Grounding corpus-linguistic measures of multiword association in human judgments of idiomaticity

Stefanie Wulff & Debra Titone
University of Florida & McGill University

Usage-based theories of language representation, processing, and acquisition are gaining ground in cognitive linguistics, corpus linguistics, and psycholinguistics. One key assumption in usage-based approaches is that speakers display a sensitivity to the distributional properties of their linguistic environment, including, among others, the frequency with which a linguistic item occurs; the cue validity of a linguistic item to predict the presence of another item; whether or not an item has been primed in previous discourse; or whether an item has been preempted by another linguistic item. Speakers' implicit analysis of these distributional properties is argued to shape comprehension and production biases, the speed and accuracy with which linguistic items are acquired, the way an item is represented mentally, and, in turn, likely also shapes speakers' intuitions regarding the use of a given linguistic item (Ellis 2006, Bybee 2010, Boyd and Goldberg 2011). However, the extent to which these distributional properties indeed constitute distinctive processes, whether they are best seen as language-specific cognitive abilities or rather stem from general cognitive skills, how they interact, and which algorithmic implementations may serve to model them most adequately are some of the most hotly debated questions across disciplines.

This paper attempts to contribute to this ongoing debate by examining the relationship between the cue validity of the component words of idiomatic phrases and human idiomaticity judgments on various properties of these idioms. To this end, we extracted 14, 439 tokens of 54 V NP-idiom types such as *bite the bullet* or *make a bundle* from the *Corpus of Contemporary American English*. We calculated the respective cue validity of the verb and noun phrase for the idiomatic phrase via an array of association measures, including raw frequency, faith, DeltaP, Fisher Yates exact tests, Mutual Information, log-likelihood, and chi-squared tests. The different association measures were then correlated with human judgment data of the idiom's perceived meaningfulness, decomposability, and final word predictability (Libben & Titone, 2008).

Significant correlations involved both raw verb frequencies, which were found to correlate highly with familiarity and meaningfulness ratings, as well as specific measures of contingency – for example, verb-idiom cue validity as expressed in a DeltaP score was highly correlated with final word predictability ratings. We will discuss the implications of these and other findings for (i) the adequacy of corpus-linguistic measures of association strength to represent cue validity, (ii) the relationship between input data and speaker intuitions, and (iii) the appeal of converging corpus-linguistic and behavioral evidence in usage-based approaches to language.

References

Ellis, Nick C. 2006. Language acquisition as rational contingency learning. *Applied Linguistics* 27.1:1-24.
Bybee, Joan. 2010. *Language, use and cognition*. Cambridge: Cambridge University Press.
Boyd, Jeremy K. and Adele E. Goldberg. Learning what not to say: categorization and statistical preemption in 'a-adjective' production. *Language* 87.1:1-29.
Libben, M., & Titone, D. 2008. The multidetermined nature of idiomatic expressions. *Memory & Cognition* 36:1103-1131.