

Measuring Mental Entrenchment of Phrases with Ratings, Reaction Time, and Google Frequency Statistics

This paper explores the view that lexical structures form a continuum, from word combinations which have literally fossilized into single units (*nightclub*) to those that both exist as independent units and yet have bonds, varying in tightness, with the words with which they frequently co-occur (Wray, 2002). The mental representation of these units is likely to be influenced by usage frequency. It is posited that mental entrenchment is not equivalent to corpus frequencies, but will depend on semantic cohesiveness and the overlap between a given phrase and other entrenched phrases.

Two studies on word combinations are described, illustrating how to adapt paradigms from visual word recognition to the project of studying multi-word utterances. A database of common word combinations was constructed using internet corpora and native speaker ratings of familiarity. The design included high-frequency and low-frequency collocations (Jackendoff, 1995), syntactically legal word combinations (mostly adjective+noun pairs), and random word combinations, which were illegal combinations that violated semantic constraints, following Pustejovsky's (1995) description of semantic domains. Google frequencies were obtained by placing quotation marks around each word pair.

Two-word phrases were sequentially displayed for 30 ms each followed by a pattern mask. On half of the trials the word pairs occurred in canonical order, and half in non-canonical order. Ability to recognize the words was highly determined by frequency of the word pair, and not by the frequency of the individual words, a novel finding for the word recognition literature (Author & Morris, 2008). Order of the sequential display did not influence ability to identify the words (thus, *code* followed by *zip* was identified at the same high rate as *zip code*), indicating top-down processes are at work. The probability of accurately reporting the two words (ignoring order) correlated with Google frequency at $r=0.63$ and with familiarity at $r=0.61$. Compared to lower frequency collocations and non-collocations, higher frequency collocations were more likely to be perceived in familiar order when displayed in reverse order. When the display was in the non-canonical order, the percentage of trials in which words were correctly reported but order was reversed (meaning canonical order reported) correlated with Google frequency at $r=.58$ and with subjective familiarity at $r=.60$.

Collecting familiarity ratings (5 pt scale) for a corpus of 160 word pairs from as few as 22 native speakers was sufficient to yield a correlation of $r=.81$ with Google frequencies. Phrases like *thank you* and *white house* had ratings that matched the Google frequencies. However, the familiarity of phrases like *child abuse* were underestimated relative to their Google frequency. This suggests that the conceptual semantics of this phrase, and possibly that fact that the semantics goes beyond the meaning of the two component words, influenced the high familiarity rating. In the other direction, raters gave lower-than expected familiarity ratings for objectively high frequency, but semantically noncohesive, phrases like *rather than*.

Findings support the view that common multi-word combination are mentally entrenched representations, with strength of entrenchment influenced by both semantic cohesiveness and usage frequency.

References

- Author & Morris, A. L. (2008). Fast Pairs: A visual word recognition paradigm for measuring entrenchment, top-down effects, and subjective phenomenology. *Journal of Consciousness and Cognition*.
- Jackendoff, R. (1995) The boundaries of the lexicon. In M. Everaert, E. Schenk, and R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives*. Hillsdale, NJ: Lawrence Erlbaum.
- Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, MA: MIT Press.
- Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge: Cambridge University Press.