# New short words and compounds in English: probabilities matter

Søren Wichmann
*Leiden University & Kazan Federal University*
wichmannsoeren@gmail.com

Little is known about how new lexical roots come about even if, across the world's languages, this is something that has happened millions of times. It is hypothesized here that such a process must involve speakers' perception of frequencies of form and meaning and their pairing. This talk looks at new lexemes in English, in addition to the graphic behavior of compounds.

Raw data come from the Corpus of Historical American English (COHA) supplemented by the Google N-Grams corpus. Derived data include word frequencies, distances in form measured as edit (Levenshtein) distances, and semantic distances measured through frequencies of words occurring in the context of the target word, specifically by a metric similar to Jensen-Shannon.

It is found that (1) there is a tendency for new short words of a given decade to be more similar to new short words of recent decades than to words from decades that are temporally more remote, suggesting that the formation of lexemes obey principles of fashion; (2) successful new words tend to have a better linear correlation between semantic and phonological distance to other words in the lexicon than unsuccessful words, suggesting that that the relation between sound and meaning is subtly non-arbitrary (cf. also Blasi et al. 2016 and Dautriche et al. 2016); (3) the process of compounds moving from being written as separate word to being hyphenated to being written as unhyphenated single words shows regularities relating to the timing of the three stages and to frequencies of the forms, indicating the importance of speakers' memories of frequencies in language change.

**References:** • Blasi, D.E. et al. (2016): Sound-meaning association biases evidenced across thousands of languages. *Proc Natl Acad Sci* USA 113, 10818-10823. • Dautriche et al. (2016): Wordform similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Sci* 1-21.