

Local Predictability in the Lexicon

Max Bane & Ed King
University of Chicago

We investigate the extent to which the contents of a language's lexicon can be described as optimized for local, statistical predictability. We consider the unigram and bigram statistics of a lexicon--the frequencies of segment-sequences of length $n \leq 2$ --and quantify each lexical item's predictability according to these frequencies through the pointwise mutual information (PMI) of its constituent pairs of adjacent segments. Examining the lexica of seven languages, we find that each is a near optimum for maximizing PMI within some space of "possible" alternative lexica, where a "possible" lexicon is one derived by permuting or editing words.