

More than words: speakers are sensitive to the frequency of multi-word sequences

Are speakers sensitive to the frequencies of multi-word sequences? There is ample evidence that speakers are sensitive to the frequencies of units on many levels of linguistic analysis; they are sensitive to the frequency of the phonotactics in their language, to how often words appear, and to what a prototypical *t* sounds like (Fidelholz 1975; Rhodes 1992, 1996). But most of this research has focused on the word level and below (phonemes, morphemes, words). Much less is known about the status of multi-word sequences. In a series of studies we show that speakers are sensitive to the frequencies of four-word sequences (four-grams), and that this happens across the frequency continuum (it is not limited to highly frequent sequences). We discuss the implications for models of the lexicon.

We tested speaker's sensitivity to sequence frequency using a phrasal-decision task. Participants saw four-word sequences and had to decide whether they were possible sequences in English or not (all test items were possible, attested sequences. For fillers we used sequences with impossible word order like '*I saw man the*'). Each test item consisted of two four-gram sequences that differed only on the final word (e.g. don't have to worry vs. don't have to wait). The pair had the same bigram and unigram frequency but differed on the sequence frequency (estimated from the 20 million word Fisher corpus). If sequence frequency affects recognition, then the high frequency variant will be decided on faster than the low frequency one.

In experiments one to three we show an effect of frequency in three frequency bins. High frequency sequences (above ten per million) were decided on faster than low frequency ones ($p > .01$ using a mixed effect regression model). Mid frequency sequences (between five and ten per million) were decided on faster than low frequency ones (under five), $p < .05$ and low frequency ones were faster than very low ones (between 0.5 and one per million), $p < .01$. These effects were found controlling for unigram and bigram frequencies. Together, the experiments show an effect of sequence frequency across the frequency continuum. We also performed a meta-analysis of these three experiments to explicitly test the hypothesis that participants' reaction times are better predicted by a continuous measure of frequency than a categorical one, in which frequencies were binned at a threshold of 10/million. In model comparisons using mixed logit models, log frequency is a better predictor of behavior ($p < .001$) than a binary measure ($p < .25$)

What are the implications of these findings? They show that speakers keep track of co-occurrence patterns beyond the bigram level and call for lexical models in which larger chains of relations are stored, and complement recent results in children (Bannard and Matthews 2008). We present one possible model, consistent with usage-based views of grammar (Goldberg 1995; Bybee 2006; Hay and Bresnan 2006), in which sequences co-exist alongside their parts. The results also highlight the need to incorporate sequence frequency into parsing models.