**Predicting Age of Acquisition for Early Words Across Languages**

Why do children understand some words before others? Words that are more frequent are likely to be learned earlier (Dale, Goodman, & Li, 2008), but many other factors have also been argued to predict the order of acquisition of words, including acoustic features, distributional properties, lexical category, and conceptual complexity (Bloom, 2000; Clark, 2009). No single study has assessed the relative contributions of various factors at scale, across languages, and over development.

We use Wordbank (wordbank.stanford.edu), an open database of developmental vocabulary data, to estimate the age of acquisition for words across nine languages (Croatian, English, Hebrew, Italian, Norwegian, Russian, Spanish, Swedish, Turkish). In each language, for each word on the MacArthur-Bates Communicative Development Inventory (CDI), we compute the age at which it is first reported as understood for at least 50% of children. We then ask what factors are most important for predicting this age of acquisition.

Each of our predictors are derived from independent resources. We estimate word frequency from unigram count in CHILDES corpora, normalized to the length of the corpus. We use previously collected norms for concreteness (Brysbaert, Warriner, & Kuperman, 2013), arousal and valence (Warriner, Kuperman, & Brysbaert, 2013), and "babiness" (a measure of association with infancy; Perry & Lupyan, 2015). To apply these norms cross-linguistically, we map each word onto translation equivalents and apply English-based values to other languages, under the assumption that cognitive and affective measures are relatively language-independent. While imperfect, this method allows us to examine languages for which limited resources exist. We also compute the number of characters in each word in each language, which is known to be highly correlated with number of phonemes and syllables.

We then fit a linear regression for the English data and a linear mixed-effects model with language as a random effect for the cross-linguistic data. Each predictor is scaled and centered so that coefficients are comparable and provide standardized estimates of the strength of the effect for each predictor. We find that concreteness, frequency, and babiness are relatively stronger predictors of age of acquisition, both in English and across languages (Figure 1). Follow-up analyses showed that lexical class was also a strong predictor but was highly correlated with concreteness both in English and across languages.

To assess developmental trends, we compare predictor values in two groups of words to determine which features best distinguish words with age of acquisition estimates earlier and later than 14 months. Within English and across all languages, early learned words are distinguished from later learned words on all of our measured factors (i.e. they are shorter, more concrete, etc). However, the magnitude of this difference is especially large for frequency and babiness (Figure 2).

Overall, we find that while frequency is predictive of age of acquisition across languages, conceptual factors are at least as important, especially for the earliest learned words. Our results imply that distributional theories of word learning should be constrained by such conceptual factors, which are likely to apply across languages.

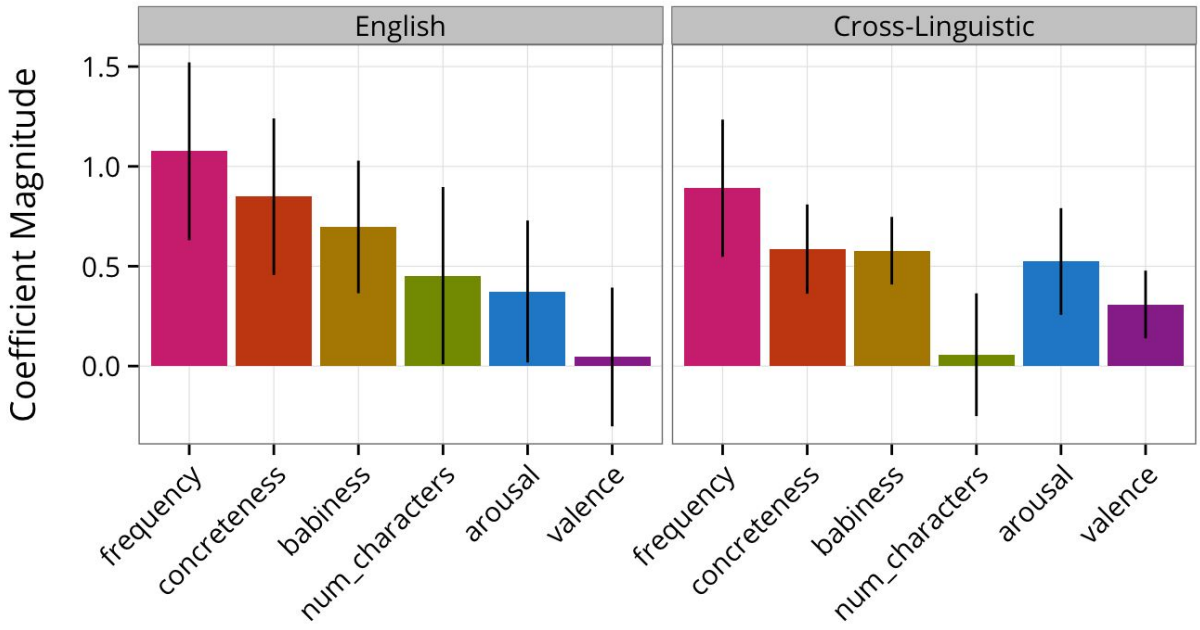


Figure 1: Estimates of each predictor's standardized coefficient in predicting age of acquisition. Ranges show 95% confidence intervals.

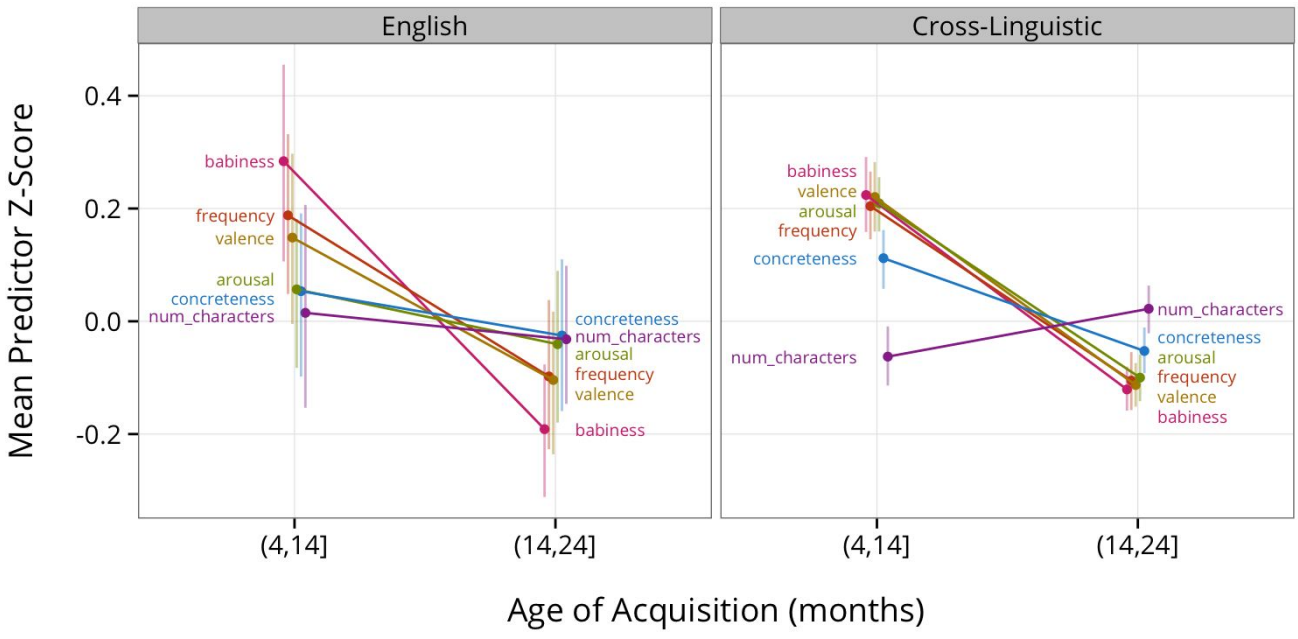


Figure 2: Mean standardized value of each predictor for words whose age of acquisition estimate lies in each age bin. Ranges show bootstrapped 95% confidence intervals.