Predicting Age of Acquisition for Early Words Across Languages

Mika Braginsky
mikabr@stanford.edu
Department of Psychology
Stanford University

Daniel Yurovsky
yurovsky@stanford.edu
Department of Psychology
Stanford University

Virginia A. Marchman
marchman@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

woof

Keywords: language acquisition; word learning; development

Introduction

Why do children understand some words before others? Words that are more frequent are likely to be learned earlier (Goodman, Dale, & 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, 2002). No single study has assessed the relative contributions of various factors at scale, across languages, and over development.

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.

Methods

We use Wordbank (wordbank.stanford.edu), an open database of developmental vocabulary data, to estimate the age of acquisition (AoA) for words across 7 languages (English, Italian, Norwegian, Russian, Spanish, Swedish, Turkish). We then ask what factors are most important for predicting this age of acquisition.

Estimating AoA

To estimate the age at which words are acquired, we took vocabulary data collected using the MacArthur-Bates Communicative Development Inventory (CDI), a family of parent-report checklists (Fenson, 2007), specifically the Words & Gestures (infant) form for 8- to 18-month-olds. When filling out a CDI a form, parents are asked to indicate whether their child understands and/or says each of around 400 words. From these data, for each word on the CDI, we computed the proportion of children at each age that are reported to understand the word. We then fit a logistic curve to these proportions and determine when the curve crosses 0.5, i.e. at what age at least 50% of children are reported to understand the word. This point is taken to be the words' age of acquisition.

Predictors

Each of our predictors are derived from independent resources. For each word that appears on the CDI Word & Gestures form in each of our 7 languages, we obtained an estimate of its frequency in child-directed speech, its length in characters, and ratings of its concreteness, valence, arousal, and relevance to babies. While frequency was measured relative to the word's language, since the conceptual ratings were collected only in English, we mapped all the words onto translation equivalents across CDI forms, allowing us to use the ratings for English words cross-linguistically. While imperfect, this method allows us to examine languages for which limited resources exist.

Items such as *child's own name* were excluded in all languages. Each predictor was also centered and scaled so that they would all have comparable units.

Frequency: For each language, we estimated word frequency from unigram count in all corpora in the CHILDES database (MacWhinney, 2000) in that language, normalized to the length of the corpus. Each word's count includes the counts of words that share the same stem (so that *dogs* counts as *dog*) or are synonymous (so that *father* counts as *daddy*). For polysemous word pairs, such as *orange* as in color and *orange* as in fruit, each occurence of *orange* in the corpus counts for both. Finally, each word's frequency estimate is taken as the log of its count.

Length: We computed the number of characters in each word in each language, which is known to be highly correlated with number of phonemes and syllables.

Concreteness: We used previously collected norms for concreteness (Brysbaert, Warriner, & Kuperman, 2014), which were gathered by asking adult participants to rate how concrete the meaning of each word is by using a 5-point scale from abstract to concrete. For the 10 CDI words that weren't part of the collected norms (mostly animal sounds such as *baa baa*), we imputed a conreteness rating from the mean of all CDI words' concreteness rating.

Valence and arousal: We also used previously collected norms for valence and arousal (Warriner, Kuperman, & Brysbaert, 2013), for which adult participants are asked to rate words on a 1-9 happy-unhappy scale (valence) and 1-9 excited-calm scale (arousal). For the 43 CDI words that weren't part of the collected norms (mostly function words such as *her*), we imputed ratings from the mean of all CDI words' ratings.

Babiness: Lastly, we used previously collected norms of "babiness", a measure of association with infancy (Perry, Perlman, & Lupyan, 2015) for which adult participants are asked to judge how relevant to babies a word is.

Results

We fit a linear regression for each language's data, as well as a linear mixed-effects model with language as a random effect for all the data pooled. Figure 2 shows the magnitude of the coefficient for each predictor in each language and cross-linguistically. We find that frequency, concreteness, and babiness are relatively stronger predictors of age of acquisition, 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).

Discussion

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.

Conclusion

Acknowledgements

Thanks to the MacArthur-Bates CDI Advisory Board.

References

Bloom, P. (2002). *How children learn the meanings of words*. MIT press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

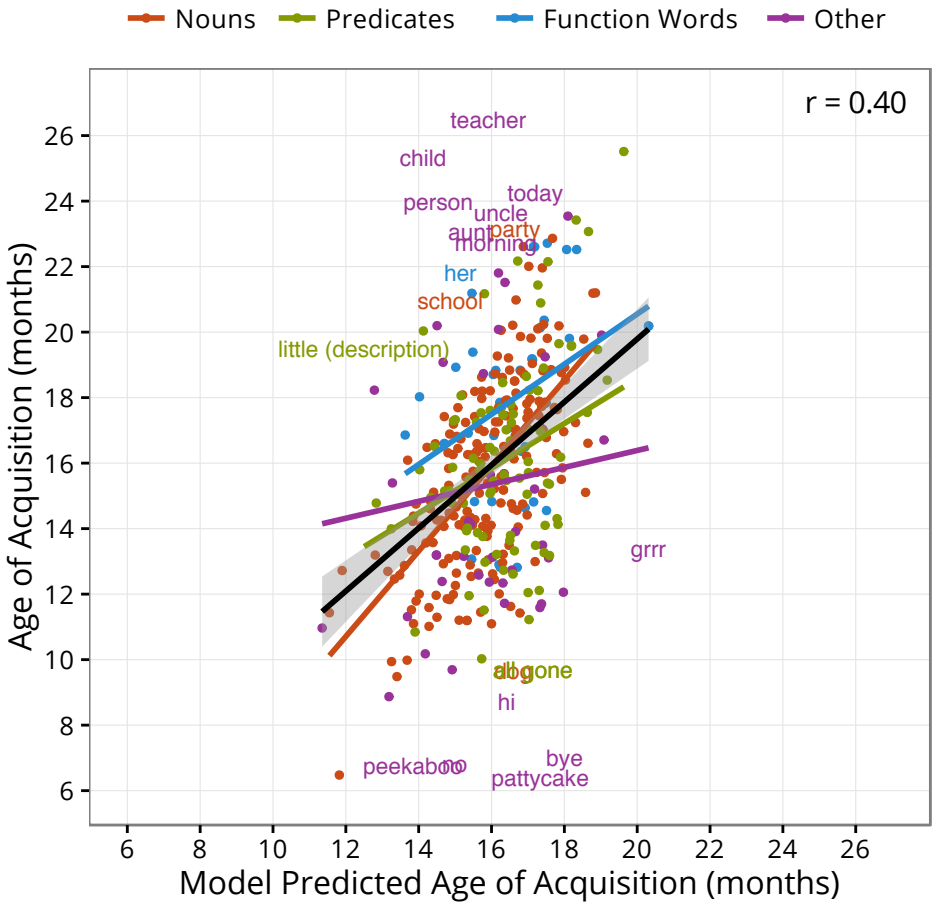

Figure 1: English model fit.

Fenson, L. (2007). *MacArthur-bates communicative development inventories: User's guide and technical manual*. Paul H. Brookes Publishing Company.

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515.

MacWhinney, B. (2000). *The cHILDES project: The database* (Vol. 2). Psychology Press.

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in english and spanish and its relation to lexical category and age of acquisition. *PloS One*, *10*(9), e0137147.

Warriner, A. B., Kuperman, V., & Brysbaert, M. (2013). Norms of valence, arousal, and dominance for 13,915 english lemmas. *Behavior Research Methods*, *45*(4), 1191–1207.

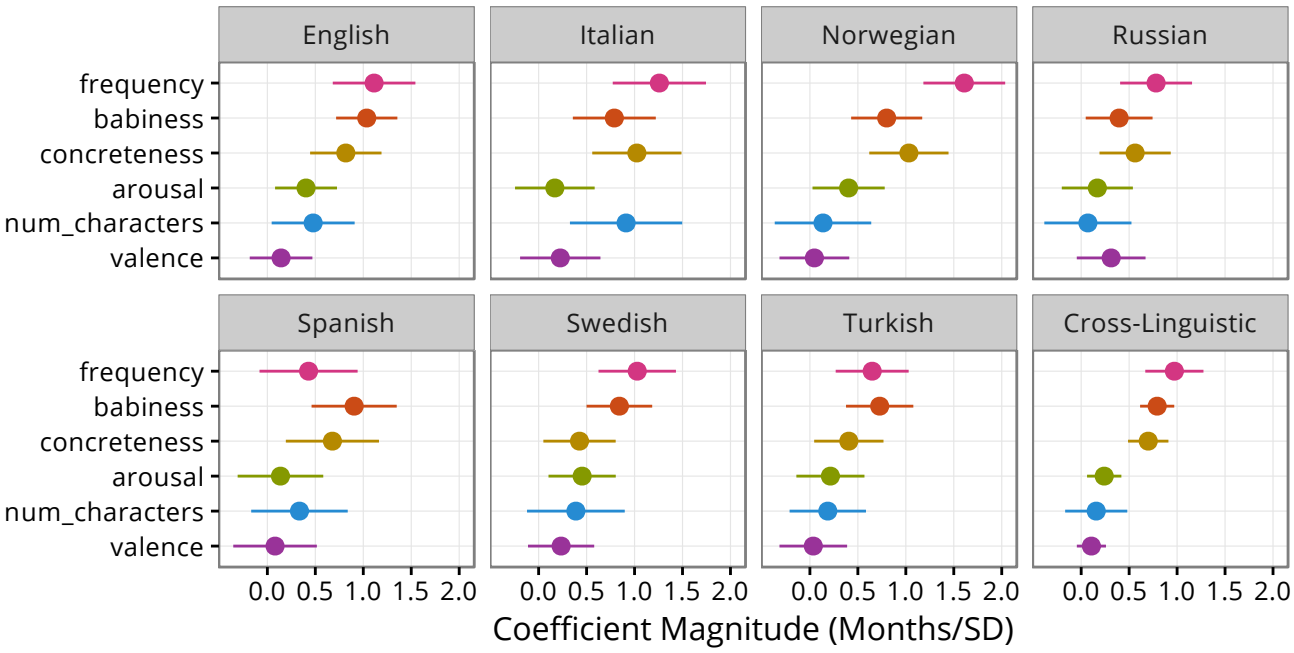

Figure 2: Magnitudes of predictor coefficients.