From *uh-oh* to *tomorrow*
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

[TODO: abstract]

Keywords: language acquisition; word learning; development

Introduction

One of the central problems facing a child acquiring their first language is to learn word meanings. Learners must integrate multiple information sources to figure out how to map the wordforms they hear onto representations of their meanings. Across many labratory experiments and small-scale models, a number of strategies have emerged as plausible components of word learning, including tracking co-occurence statistics between words and referents to deduce word meaning across situations; attending to social cues like pointing and eye gaze to direct hypothesis search; relying on certain biases, such as priviledging basic level category labels, to constrain inference; drawing on knowledge of relations between words to use known meanings to learn new ones; and so on.

Each of these abilities has been reliably demonstrated in the constrained learning contexts of the laboratory, indicating that they could be used for word learning. But it is less clear to what extent children actually employ them in the natural word learning environment and how they interact to create the longer-term dynamics of vocabulary acquisition. How do various word learning mechanisms differ in their relative contributions, and how does that change over the course of development?

One way to address these questions is to use naturalistic large-scale vocabulary development data to examine the contribution of various theoretically-relevant factors to vocabulary growth. We can look across children to determine how easy or hard various words are to learn, and then examine the relationship between words' difficulties and various word properties that relate to proposed word learning mechanisms. Foundational work using such an approach has revealed that in English, within lexical category, words that are more frequent in speech to children are likely to be learned earlier (Goodman, Dale, & Li, 2008; B. C. Roy, Frank, & Roy, 2009). Further studies have delved into the relevance of semantic networks (Hills, Maouene, Maouene, Sheya, & Smith, 2009), neighborhood density (Stokes, 2010), iconocity (Perry, Perlman, & Lupyan, 2015), and linguistic distinctiveness (B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015) to vocabulary construction.

However, the previous studies use different datasets, focus on different predictors, and for the most part only analyze English data. It is thus impossible to compare the relative importance of the many relevant factors and to draw robust conclusions. To remedy this issue, we present analyses based on data from Wordbank (wordbank.stanford.edu), an open repository of language development data. By aggregating administrations of the MacArthur-Bates Communicative Development Inventory (CDI; Fenson, 2007), a family of parent-report vocabulary checklists, Wordbank provides large-scale vocabulary data that we use to conduct analyses over development and across languages.

We integrate Wordbank data with characterizations of the word learning environment from the CHILDES database (MacWhinney, 2000) and elsewhere, a multiple data source approach pioneered by Goodman et al. (2008). Building on their work, we want to move beyond frequency to examine a variety of information sources. We specifically follow B. C. Roy et al. (2015) in predicting age of acquisition (AoA) as a function of several different environment predictors. In analyzing a high-density longitudinal corpus for a single English-acquiring child, Roy et al found that frequency, number of characters, and mean length of utterances were predictive of the age of a word's first production. Due to the nature of the data, this analysis was limited to one language (in fact to one subject) and could only test production, distancing the connection between properties of the input and the child's emerging understanding of words.

Our approach provides a complimentary analysis by using CDI comprehension data to look at earliest words, across languages. We estimate the age of acquisition for around 400 words in each of 7 languages. We also estimate each words' frequency and mean length of utterances (MLU) in which it appears, as well as obtaining ratings of each words' concreteness, valence, arousal, and relevance to babies from previously collected norms. We then predict words' AoA from all of these properties, and assess the relative contributions of each factor, along with the interaction of predictors with the lexical category and their changes over development.

By incorporating a variety of theoretically-important factors, as well as basing our analysis on a large samples of words and children and building towards more cross-linguistic coverage, our study presents a more thorough investigatation of the question of what properties determine words' learnability.

[TODO: need to relate predictors to theories, at least vaguely, not sure how to do this without setting up strawmen]

[TODO: something about how this whole thing will help

understand mechanisms of word learning]
[TODO: make some predictions?]

Data

We use Wordbank (wordbank.stanford.edu), an open database of developmental vocabulary data, to estimate the age of acquisition 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 Age of Acquisition

To estimate the age at which words are acquired, we took vocabulary data collected using the MacArthur-Bates Communicative Development Inventory, a family of parent-report checklists, 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 using robust fitting of a generalized linear model 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, the mean utterance length (MLU) of sentences in which if appears in child-directed speech, its length in characters, and ratings of its concreteness, valence, arousal, and relevance to babies. While frequency and MLU are 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. Lexical category was determined on the basis of the conceptual categories presented on the CDI form, such that Nouns is common nouns, Predicates is verbs and adjectives, Function Words is closed-class words, and Other is the remaining items.

Frequency: For each language, we estimated word frequency from unigram count in all corpora in the CHILDES database for 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.

Language	CDI Items	CDI Admins	CHILDES Words
English	386	2,452	7,858,051
Italian	351	648	328,168
Norwegian	338	3,021	204,406
Russian	337	768	32,398
Spanish	333	778	1,458,327
Swedish	311	467	698,515
Turkish	327	1,115	44,347

Table 1: Dataset statistics

MLU: For each language, we estimated each word's MLU by calculating the mean number of words in the sentences in which that word appears in all corpora in the CHILDES database for that language. Words that only occur in one sentence were excluded.

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 120 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 119 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 et al., 2015) for which adult participants are asked to judge how relevant to babies a word is.

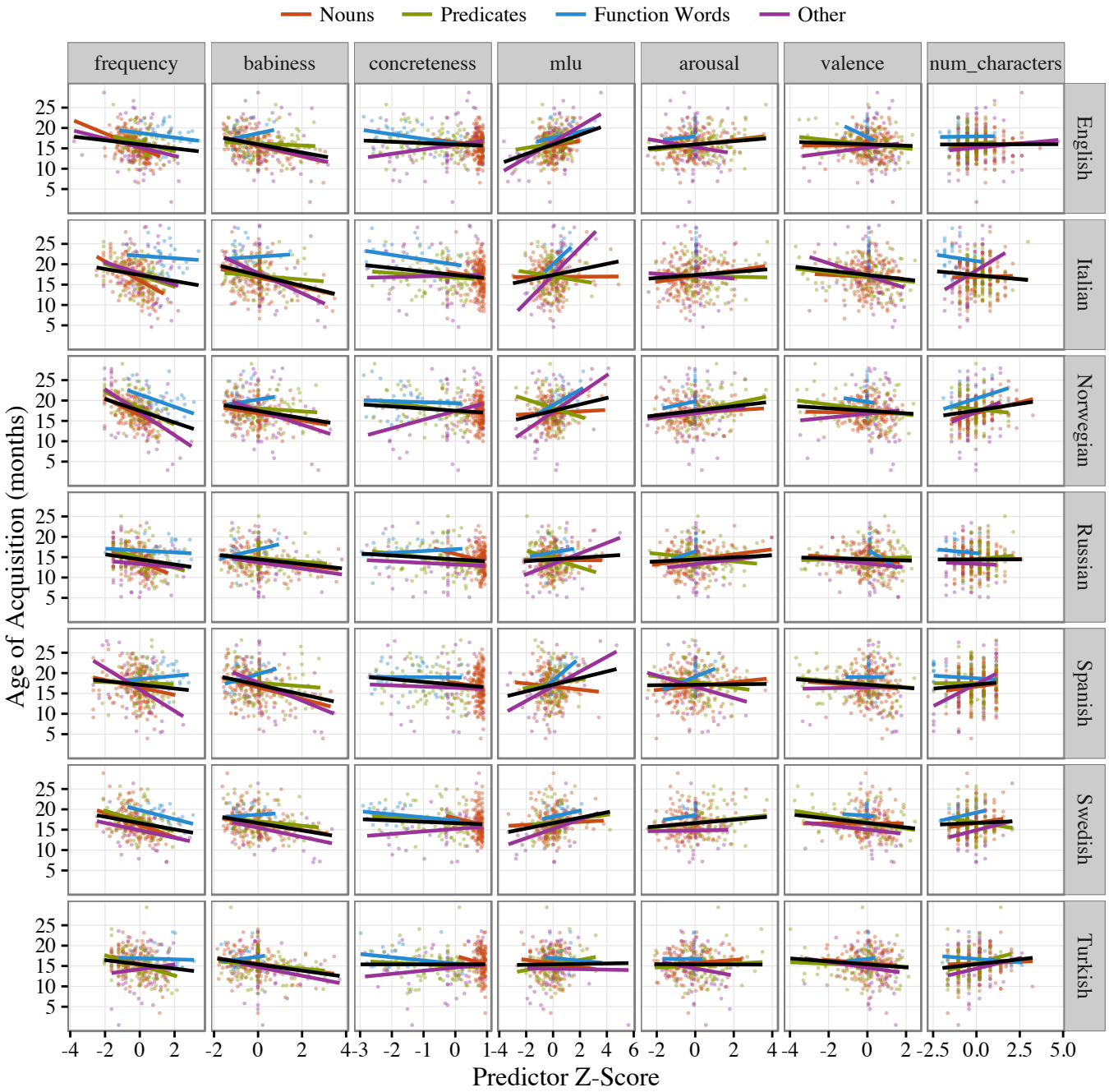

Figure 1: Relationship between predictors and AoA for each lexical category in each language.

Analysis

An overview of our entire dataset can be seen in Figure 1, which shows each word's estimated age of acquisition against its predictor values, separated by language and lexical category. We present three analyses of these data: 1) how the word properties change over development, 2) their relative contributions to predicting AoA, and 3) their interaction with lexical category.

Developmental Trajectory

To assess developmental trends, we examine how the values of each predictor change as a function of estimated AoA. Figure 2 shows these trajectories, with a cubic curve smoothing over all words. Words that are learned earlier are more frequent, higher in babiness, and appear in shorter sentences. Concreteness exhibits a U-shaped trajectory, with the earliest learned words actually being relatively abstract, such as many social routines and animal sounds.

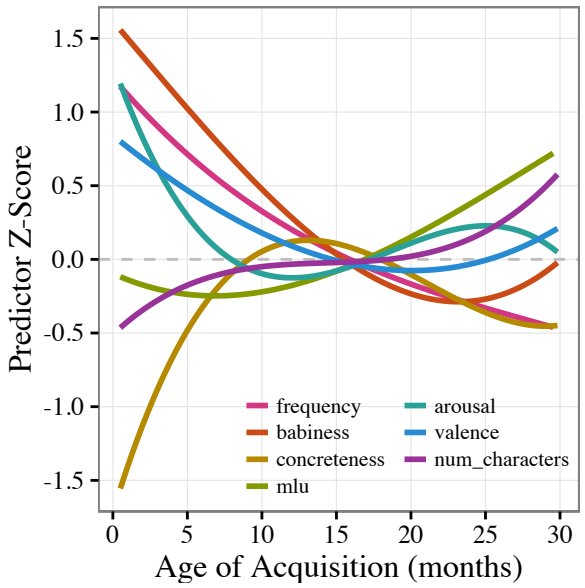

Figure 2: Predictor values over development.

Predicting AoA

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. For illustrative purposes, Figure 3 compares the predictions of the model to AoA estimates, for only English data, with outliers labelled.

Figure 4 shows the magnitude and direction of the coefficient for each predictor in each language and cross-linguistically. We find that frequency, babiness, concreteness, and MLU are relatively stronger predictors of age of acquisition, across languages and in the cross-linguistic model. Overall there's considerable consistency in how the predictors pattern in various languages, although with interesting differences: for example, MLU in English appears to be unusually strong, while frequency in Spanish look unusually weak. There is also variability in the overall fit of the models to the data, with some languages, such as Norwegian, having relatively more of the variance explained than others, such as Turkish.

Lexical Category

Previous work gives reason to believe that predictors' relationship with age of acquisition differs among various lexical categories (Goodman et al., 2008). To investigate these effects, we break down our data by lexical category and fit separate cross-linguistic linear mixed-effects models for each one. Figure 5 shows the magnitudes and directions of the resulting coefficients, leaving off the less strong predictors. We find that frequency matters most for Nouns and comparatively little for Function Words, while MLU is irrelevant for both Nouns and Predicates, but highly informative for Function Words and Other items.

[TODO: give more examples everywhere of specific words that are low/high in AoA, predictor values]

[TODO: mention predictors' correlations to each other]

[TODO: mention that predictors have different variabilities, that probably matters]

[TODO: emphasize the sketchiness of using ratings cross-linguistically]

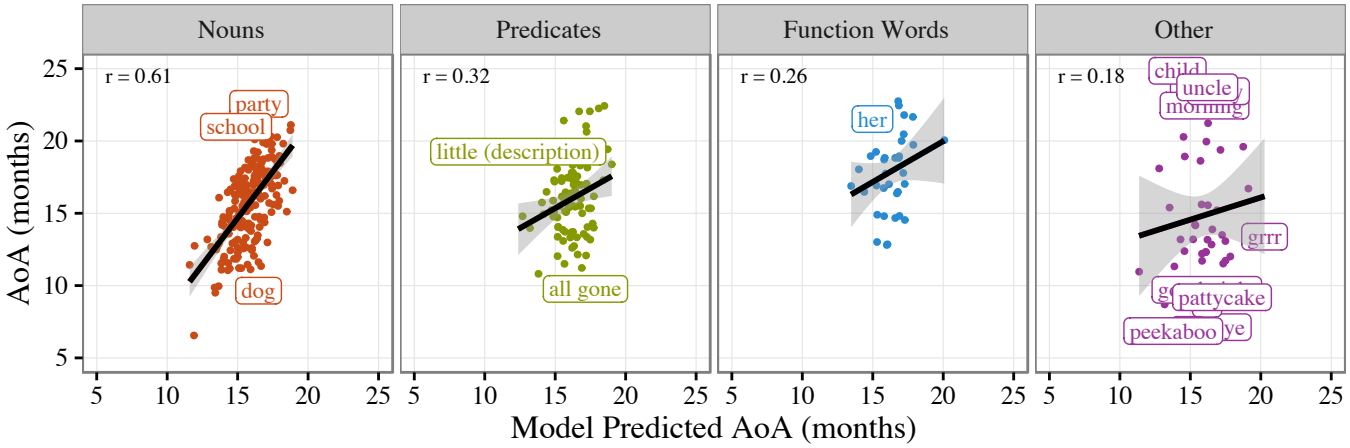

Figure 3: English model fit.

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. *[TODO: actually say something]*

Acknowledgements

Thanks to the MacArthur-Bates CDI Advisory Board.

References

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.

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.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, *20*(6), 729–739.

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.

Roy, B. C., Frank, M. C., & Roy, D. (2009). Exploring word learning in a high-density longitudinal corpus. *Proceedings of the Cognitive Science Society*.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2015). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*.

Stokes, S. F. (2010). Neighborhood density and word frequency predict vocabulary size in toddlers. *Journal of Speech, Language, and Hearing Research*, *53*(3), 670–683.

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.

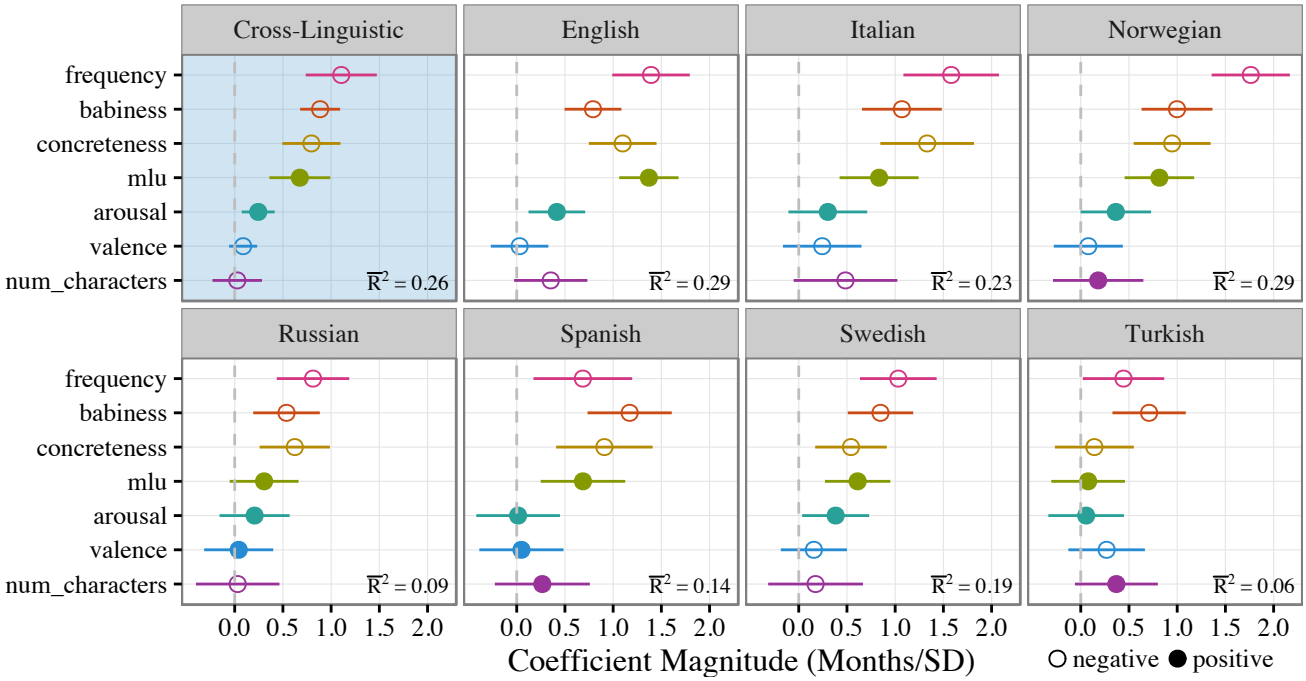

Figure 4: Magnitudes of predictor coefficients.

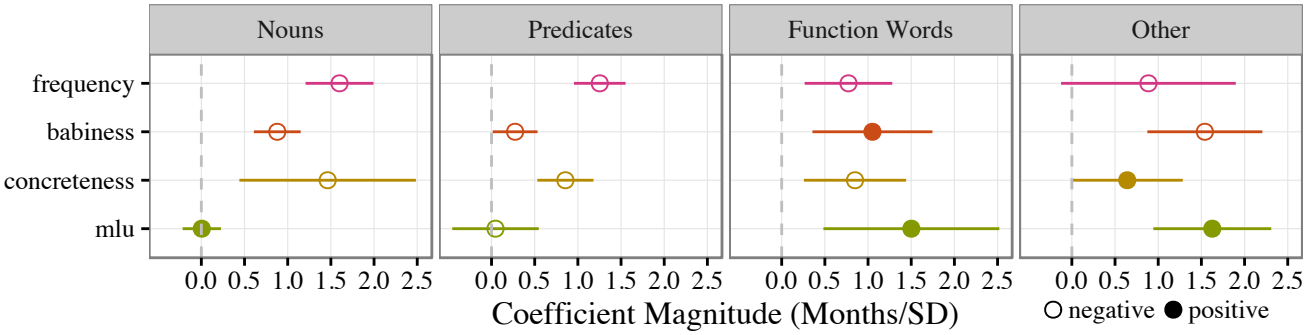

Figure 5: Magnitudes of predictor coefficients by lexical category.

Measure	Lowest	Highest
aoa	mommy, bottle, peekaboo	babysitter, teacher, naughty
frequency	living room, cockadoodledoo, grrr	you, it, that
babiness	donkey, penny, jeans	baby, bib, bottle
concreteness	how, now, that	apple, ball, banana
mlu	cockadoodledoo, peekaboo, uh oh	babysitter, when (question), day
arousal	shh, asleep, blanket	naughty, money, scared
valence	sick, owie, ouch	happy, hug, love
num_characters	i, in, it	cockadoodledoo, refrigerator, living room

Table 2: Examples of words with the lowest and highest values for age of acquisition and each predictor.