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

Why do children learn some words so much faster than they learn other words? Regularities in the difficutly of learning different words can tell us about the mechanisms responsible for their learning. We leverage this relationship in a large-scale corpus analysis, predicting the age at which X children learn Y words in Z languages from a set of linguistic, environmental, and conceptual factors. We find broad consistency and interesting subtle variation, both across languages and across development. By leveraging data at a significantly larger scale than previous work, our analyses highlight both the power that emerges from unifying previously disparate theories, and the amount of reliable variation that still remains unexplained.

Keywords: language acquisition; word learning; development

Introduction

Word learning is one of the central challenges of language acquisition. Learners must integrate multiple information sources to figure out how to map the forms they hear onto representations of their meanings. Across many laboratory experiments and small-scale models, a number of strategies have emerged as plausible components of word learning, including tracking co-occurrence statistics between words and referents to deduce word meaning across situations; attending to social cues like pointing and eye gaze; relying on biases, such as a basic level category bias; drawing on knowledge of relations between words to use known meanings to learn new ones; and so on.

Each of these strategies have been reliably demonstrated in the constrained learning context of the laboratory, indicating that they are possible parts of the word learning process. Small-scale experimental studies typically do not tell us whether these strategies operate uniformly across children, ages, and languages, however. It is also difficult to explore how strategies interact to create the longer-term dynamics of vocabulary acquisition. How do the various strategies differ in their relative contributions? And how does their influence change over the course of development?

Our approach to addressing these questions is to use large-scale vocabulary development data to examine these interactions. We can look across children to determine how easy or hard various words are to learn, and then examine the relation between word difficulty and features that differentiate words according to different strategies. For example, distributional learning strategies rely critically on frequency. Thus, to make a first assessment of the importance of distributional learning, we can examine the relationship between the age at which words are typically acquired and word frequency in child-directed speech.

Indeed, 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 (J. C. Goodman, Dale, & Li, 2008; B. C. Roy, Frank, & Roy, 2009). Further studies have found evidence for semantic networks (Hills, Maouene, Maouene, Sheya, & Smith, 2009), neighborhood density (Stokes, 2010), iconicity (Perry, Perlman, & Lupyan, 2015), and linguistic distinctiveness (B. C. Roy, Frank, DeCamp, Miller, & Roy, 2015) as predictors of age of acquisition (AoA), suggesting that they are likely contributors to vocabulary development at scale. But these exciting findings are nevertheless limited in their generality because they used different datasets, focused on different predictors, and for the most part, only analyzed English data. It is thus impossible to compare the relative importance of the many relevant factors under consideration and to draw robust conclusions.

To remedy this issue, we present analyses based on data from Wordbank (wordbank.stanford.edu), an open repository of cross-linguistic language development data (Frank, Braginsky, Yurovsky, & Marchman, in press). 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 based on analogous instruments from more than 40,000 children in 14 different language communities. As such, Wordbank offers a novel resource for richer and more powerful analyses of vocabulary learning over development and across languages.

We integrate AoA estimates from Wordbank with characterizations of the word learning environment from the CHILDES database (MacWhinney, 2000) and elsewhere, a multiple data source approach originated by J. C. Goodman et al. (2008). Building on this work, we examine interactions between a variety of both environmental and conceptual factors. Using this same approach on a high-density longitudinal corpus for a single English-acquiring child, Roy et al. found that the length, usage frequency, and mean length of the utterances in which it occurred were all predictors of a word's AoA. But due to the nature of the dataset, this analysis used production-based AoA estimates and was further limited by relying on data from only one child (and hence one language).

Our approach provides a complimentary analysis by using CDI comprehension data available in Wordbank to look at a common set of the earliest words that children learn across several different languages. We estimate AoA for around 400 words from CDIs in each of 7 languages. We also estimate each word's frequency and mean length of utterance (MLU) based on the sentences in which it appears in CHILDES. We also obtain ratings of each word's concreteness, valence, arousal, and relevance to babies from previously collected norms. We use these measures to predict each word's AoA, assessing the relative contributions of each, as well as how predictors change over development and their interactions with lexical category. Each of these analyses has the potential to provide leverage on long-standing theoretical questions.

A first theoretical question of interest is which lexical categories are most influenced by input-related factors like frequency and utterance length, compared with conceptual factors like concreteness or valence. For example, the "division of dominance" theory suggests that nouns might be more sensitive to cognitive factors while predicates and closed-class words might be more sensitive to linguistic factors (Gentner & Boroditsky, 2001). On the other hand, on syntactic bootstrapping theories (Gleitman, 1990), nouns are argued to be learned via frequent co-occurrence (operationalized by frequency) while verbs might be more sensitive to syntactic factors (operationalized here by utterance length), and neither would be particularly sensitive to conceptual complexity (Snedeker, Geren, & Shafto, 2007).

A second question of interest is the extent to which there is variability across languages in the relative importance of predictors. For example, are there differences in the importance of syntactic factors in morphologically more complex languages like Russian and Turkish, compared with simpler ones like English? Differences of this type might be revealing of the degree to which learners face different challenges in different language environments. Or consistency may suggest the operation of similar learning mechanisms and strategies that are not as dependent on the complexities of phonology, morphology, and syntax in a particular language.

Overall, 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 investigation of the question of what properties determine words' learnability.

Data

We use Wordbank, 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. Table 1 gives an overview of our data sources.

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 a robust generalized linear model (using the `robustbase` package in `R`) and determine when the curve crosses 0.5, i.e. at what age at least 50% of children are reported to understand the word. Following J. C. Goodman et al. (2008), we take this point to be each word's age of acquisition.

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

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

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

Predictors

Each of our predictors is derived from independent sources. 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 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. All items such as *child's own name* were excluded. Examples of each of these predictors for English are shown in Table 2.

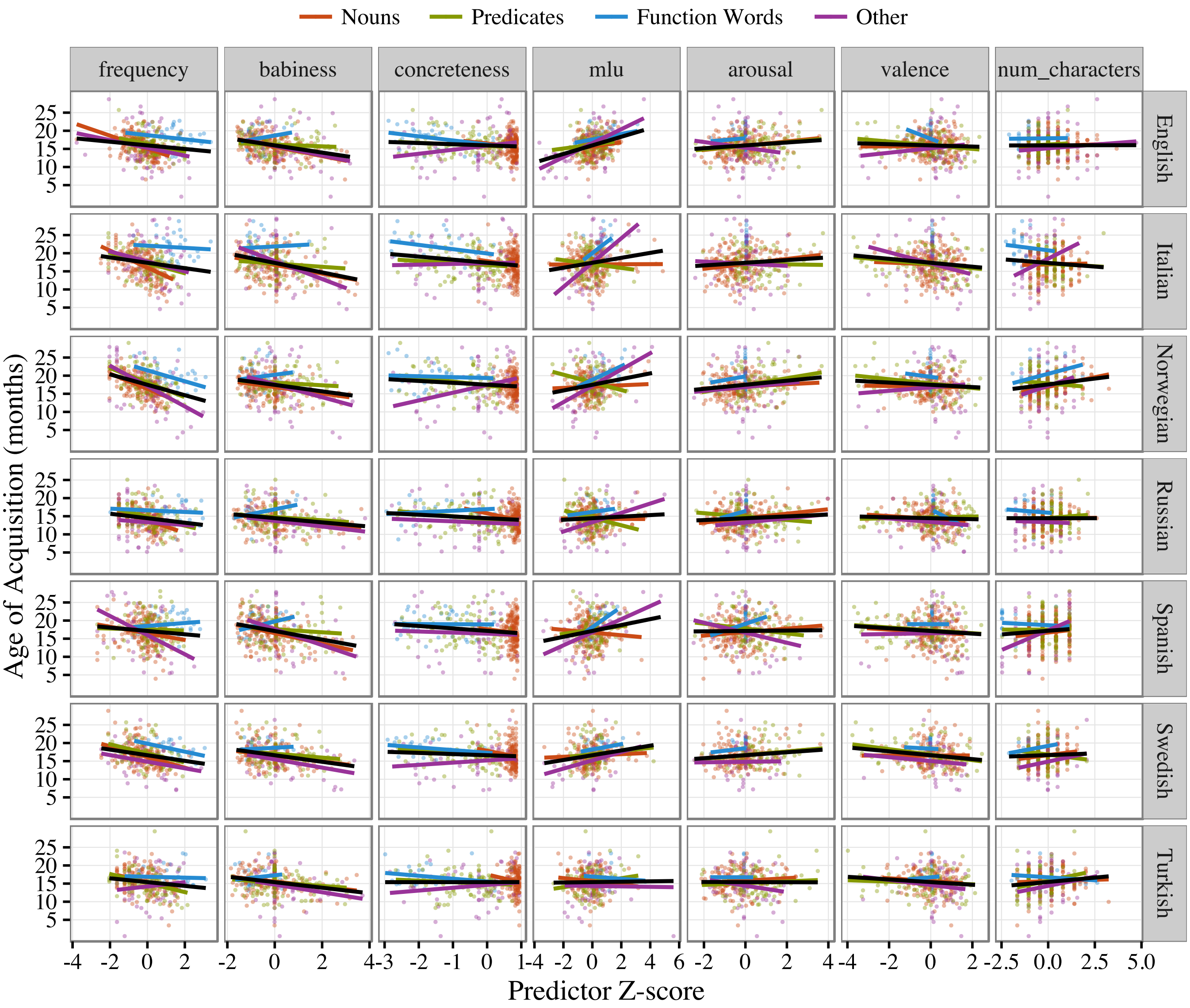

Figure 1: Relationship between predictors and AoA for each lexical category in each language. Each point represents a word, with lines indicates linear model fits for each lexical category (in colors) and overall (in black).

Frequency and MLU are measured relative to the word's language. But since extant datasets for conceptual ratings are primarily available for English, we mapped all words onto translation equivalents across CDI forms, allowing us to use the ratings for English words across languages. While necessarily imperfect, this method allows us to examine languages for which limited resources exist. Translation equivalents are available in the wordbank database using the `wordbankr` package (Frank et al., in press).

Each numeric predictor was centered and scaled so that all predictors would have comparable units. Finally, lexical category was determined on the basis of the conceptual categories presented on the CDI form (e.g., "Animals"), such that the noun category contained common nouns, predicates contained verbs and adjectives, function words contained closed-class words, and other contained the remaining items (following Bates et al., 1994).

Frequency For each language, we estimated word frequency from unigram counts based on all corpora in the CHILDES database for that language. 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, the occurrences of *orange* in the corpus were split uniformly between the two senses.[1] Finally, counts were normalized to the length of each corpus and then log transformed.

[1] Polysemy is puzzle for our analysis because word sense disambiguation is not possible across languages. Our approach here is a pragmatic one, and should not affect our overall results given the relatively small number of polysemous words in each language.

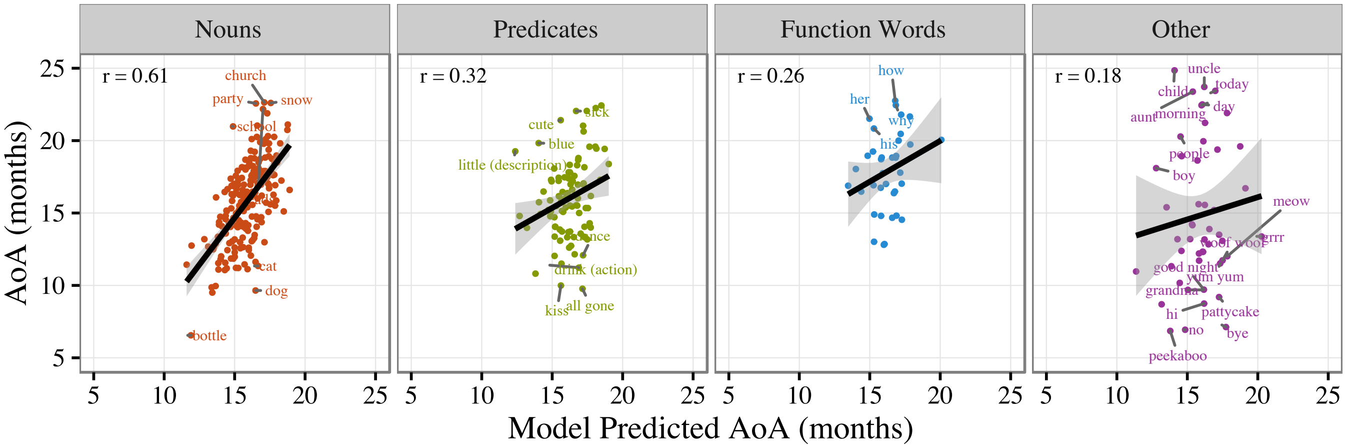

Figure 2: Comparison between the model predicted and actual ages of acquisition for words in English. Words with an absolute error above 5 months are labelled for reference.

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

Length We computed the number of characters in each word in each language. While imperfect, this metric of length is highly correlated with number of phonemes and syllables in each word (Lewis & Frank, under review).

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 on 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 concreteness ratings as the mean of all other words' ratings.

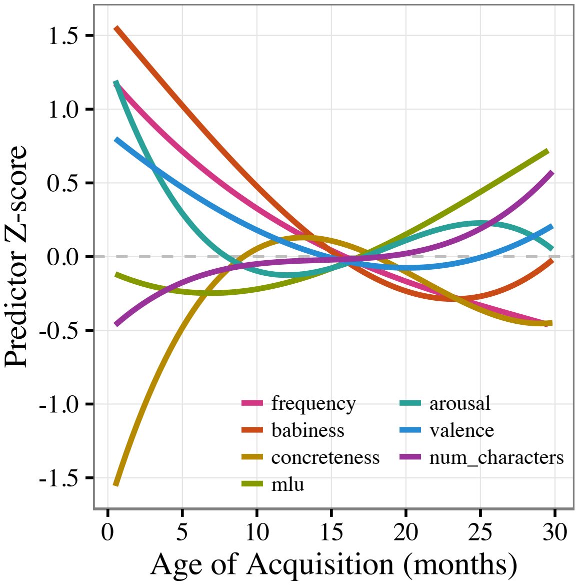

Figure 3: Predictor values over development, with a cubic curve smoothing over all items in all languages.

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) in which adult participants are asked to judge how relevant to babies a word is.

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 predictor values change for words learned earlier and later, 2) their relative contributions to predicting AoA, and 3) their interaction with lexical category.

Developmental Trajectories

To assess developmental trends, we examine how the values of each predictor change as a function of estimated AoA. Figure 3 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 (e.g., social routines and animal sounds).

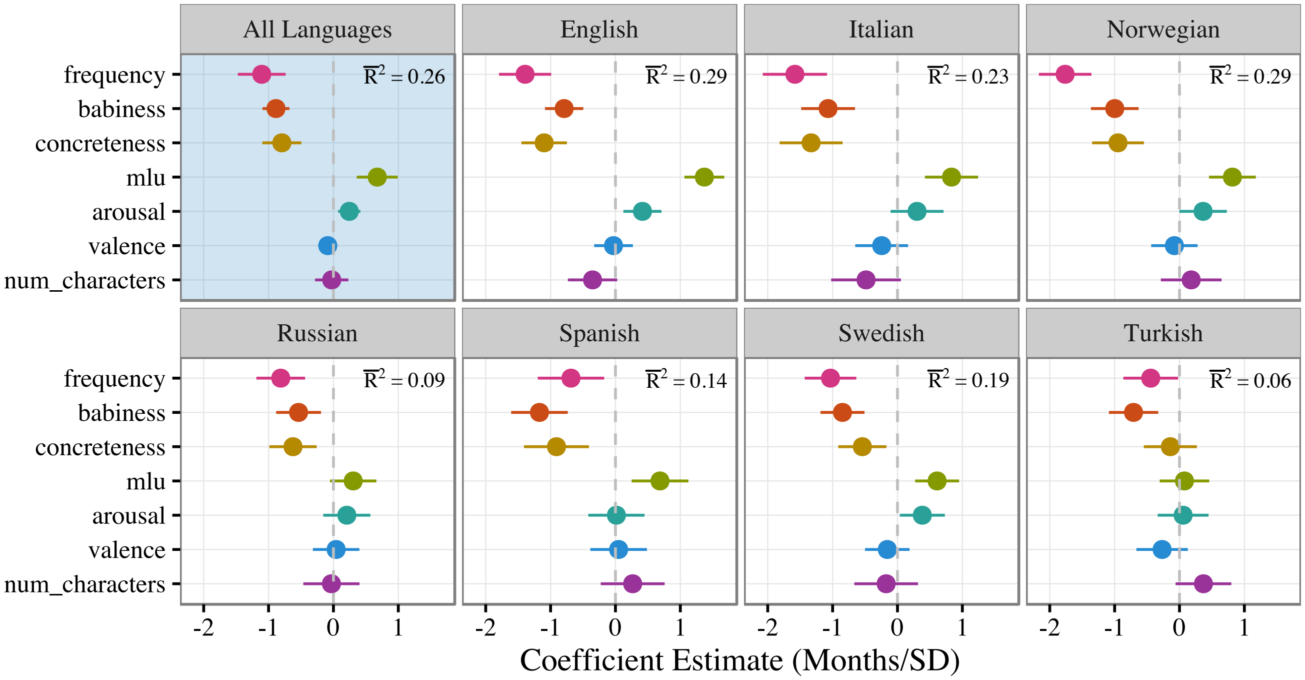

Figure 4: Estimates of predictor coefficients by language and for the all language model. Values above 0 indicate a positive relationship (i.e. words with higher MLU tend to have a higher AoA), while values below 0 indicate a negative relationship (i.e. words with higher frequency tend to have a lower AoA.

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 across all languages. For illustrative purposes, Figure 2 shows the predictions of the English model to the empirical AoA estimates.

Figure 4 shows the coefficient estimate for each predictor in each language and for all languages combined. We find that frequency, babiness, concreteness, and MLU are relatively stronger predictors of age of acquisition, across languages and in the all languages model. Overall there was considerable consistency in how the predictors pattern in various languages, although with some 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 much 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 (J. C. Goodman et al., 2008). To investigate these effects, we separated our data by lexical category and fit separate linear mixed-effects models for each. Figure 5 shows the resulting coefficient estimates (leaving off the weaker predictors for illustrative purposes). Frequency mattered most for nouns and comparatively little for function words, while MLU was irrelevant for both nouns and predicates, but highly informative for function words and other items.

Discussion

What makes words easier or harder for young children to learn? Previous experimental work has largely addressed this question using small-scale experiments. While such experiments can identify sources of variation, they typically do not allow for different sources to be compared in detail. In contrast, observational studies allow the effects of individual factors (with frequency being the most common) to be measured across ages and lexical categories (e.g., J. C. Goodman et al., 2008). Scale comes at a cost in terms of detail, however: The availability of both predictors and outcome data has been quite limited.

By including 7 languages and as many predictors, our current work expands the scope of previous observational studies of age of acquisition. Our data show a number of patterns that confirm and expand previous reports. First, predictors differed in importance across even very early development. For example, certain concepts that were more strongly associated with babies appeared to be learned early for children across languages (as in Tardif et al., 2008). Second, we found general consistency in predictor coefficients across languages (even as overall model fit varied, at least in part due to the amount and quality of data for different languages). This consistency supports the idea that differences in culture or language structure do not lead to fundamentally different acquisition strategies, at least at the level of detail we were able to examine.

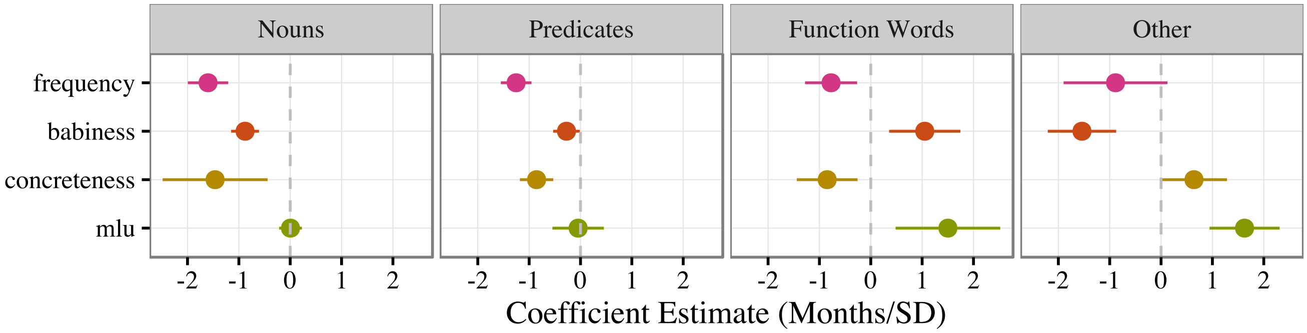

Figure 5: Estimates of predictor coefficients by lexical category, without any separation by language and omitting the three weaker predictors..

Finally, different predictors appeared more important across lexical categories. Frequent, concrete nouns were learned earlier, consistent with theories that emphasize the importance of early referential speech (e.g., Baldwin, 1995). But for predicates, concreteness was somewhat less important, and for function words, MLU was most predictive. Overall these findings are consistent with theories that emphasize the role of linguistic structure over conceptual complexity in the acquisition of other lexical categories beyond nouns (Gentner & Boroditsky, 2001; Snedeker et al., 2007).

Despite its larger scope, our work still shares a number of important limitations with previous studies. First and foremost, our approach is to predict one set of individuals with data about the experience of a completely different set and ratings of concepts gathered from yet others. In contrast to dense-data approaches (B. C. Roy et al., 2015), this approach fundamentally limits the amount of variability we will be able to capture. In addition, the granularity of the predictors that can be extracted from corpus data and applied to every word is necessarily quite coarse. Ideally, predictors could be targeted more specifically at particular theoretical constructs of interest (for example, the patterns of use for particular predicates).

Perhaps the most important theoretical challenge in the study of early language is linking individual observations or experiments to the broader patterns of acquisition observed in large datasets. We have strong theories of how individual learning situations proceed (M. C. Frank, Goodman, & Tenenbaum, 2009; McMurray, Horst, & Samuelson, 2012), and although we cannot yet discern unambiguous signatures of those theories in the large-scale data we have available, our analyses highlight the importance of studying aggregated data. They also demonstrate the value of analyzing a number of factors, including those predicted to be relevant by different and even competing theories.

Even for English, in which our seven predictors capture the most variance (r^2 = 0.29), much still remains unexplained. Further, this variance is highly reliable – cross-validation using half of the English-speaking children to predict ages of acquisition for the other half yields r^2 = 0.98. While adjudicating between theories is a critical component of the scientific process, it is important not to forget that our working theories are incomplete (Breiman, 2001). Unification across these theories may be the other critical step in making progress on understanding language learning (Newell, 1973).

All data and code for these analyses are available at `https://github.com/mikabr/aoa-prediction`

Acknowledgements

Thanks to the MacArthur-Bates CDI Advisory Board. This work supported by NSF BCS #1528526.

References

Baldwin, D. A. (1995). Understanding the link between joint attention and language. *Joint Attention: Its Origins and Role in Development*, 131–158.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., . . . Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(01), 85–123.

Breiman, L. (2001). Statistical modeling: The two cultures. *Statistical Science*, *3*, 199–231.

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.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (in press). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. *Psychological Science*, *20*, 578–585.

Gentner, D., & Boroditsky, L. (2001). Individuation, relativity, and early word learning. In *Language acquisition and conceptual development* (p. 215). Cambridge University Press.

Gleitman, L. (1990). The structural sources of verb meanings. *Language Acquisition*, *1*(1), 3–55.

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.

Lewis, M. L., & Frank, M. C. (under review). The length of words reflects their conceptual complexity.

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

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831.

Newell, A. (1973). You can't play 20 questions with nature and win: Projective comments on the papers of this symposium. In W. G. Chase (Ed.), *Visual information processing* (pp. 1–26). New York: Academic 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*, *112*(41), 12663–12668.

Snedeker, J., Geren, J., & Shafto, C. L. (2007). Starting over international adoption as a natural experiment in language development. *Psychological Science*, *18*(1), 79–87.

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.

Tardif, T., Fletcher, P., Liang, W., Zhang, Z., Kaciroti, N., & Marchman, V. A. (2008). Baby's first 10 words. *Developmental Psychology*, *44*(4), 929.

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.