From *uh-oh* to *tomorrow*: Predicting age of acquisition for early words across languages

Mika Braginsky[a,1], **Daniel Yurovsky**[b], **Stephan Meylan**[c], **Virginia Marchman**[d], **and Michael C. Frank**[d]

[a]Department of Brain and Cognitive Sciences, Massachusetts Insitute of Technology, Cambridge, MA 02139; [b]Department of Psychology, University of Chicago, Chicago, IL 60637; [c]Department of Psychology, University of California Berkeley, Berkeley, CA 94720; [d]Department of Psychology, Stanford University, Stanford, CA 94305

This manuscript was compiled on July 21, 2016

Why do children learn some words earlier than others? Regularities and differences in the age of acquisition for words across languages yield insights regarding the mechanisms guiding word learning. In a large-scale corpus analysis, we estimate the ages at which 38,212 children learn 300-400 words in ten languages, predicting them on the basis of independently-derived linguistic, environmental, and conceptual factors. Predictors were surprisingly consistent across languages, but varied across development and as a function of lexical category (e.g., concreteness predicted nouns while linguistic structure predicted function words). By leveraging data at a significantly larger scale than previous work, our analyses highlight the power that emerges from unifying previously disparate theories, but also reveal the amount of reliable variation that still remains unexplained.

TODO

Word learning is one of the central challenges of language acquisition. Learners must integrate multiple information sources to map the word 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; and drawing on knowledge of relations between words to use known meanings to learn new ones.

Each of these strategies has been reliably demonstrated in the constrained learning context of the laboratory, indicating that they are possible parts of the word learning process. However, small-scale experimental studies typically do not tell us whether these strategies operate uniformly across children, ages, and languages. 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. By aggregating across a large number of children, we can look past individual differences in acquisition to investigate not only which words are relatively easy or hard to learn, but also what features affect their acquisition. For example, distributional learning strategies rely critically on frequency. Thus, to make a first assessment of the contribution of distributional learning, we can examine the relationship between the age at which words are typically acquired and word frequency in child-directed speech.

Such an approach has revealed that in English, within a lexical category, words that are more frequent in speech to children are likely to be learned earlier (1). And further studies have found evidence for semantic networks (2), neighborhood density (3), iconicity (4), and linguistic distinctiveness (5) as additional predictors of age of acquisition (AoA), suggesting that they are likely contributors to vocabulary development. But these exciting findings are nevertheless limited in their generality because they used different datasets, focused on different predictors, and almost exclusively 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 (6). By aggregating administrations of the MacArthur-Bates Communicative Development Inventory (7), 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. Wordbank presents 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 (8) and elsewhere, a multiple data source methodology originated by (1). Building on this work, we examine interactions between a variety of linguistic, environmental, and conceptual factors. Using a similar 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 predictive 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 acquiring a single language.

Our work provides a complimentary analysis by using CDI comprehension data available in Wordbank to look at the earliest words that children learn across several different languages. We estimate AoA for approximately 400 words from CDIs in each of seven languages. We also estimate each word's frequency and mean length of utterance (MLU) based on the set of utterances in CHILDES containing the word. Additionally, we obtain ratings of each word's concreteness, valence, arousal,

Significance Statement

TODO

Author contributions: TODO

The authors declare no conflict of interest.

and relevance to babies from previously collected norms. We use these measures to predict words' AoA, assessing the relative contributions of each, as well as how they change over development and interact with lexical category. Each of these analyses has the potential to advance our understanding of the theoretical underpinnings of word learning.

A first theoretically-motivated question is which lexical categories are most influenced by input-related factors, like frequency and utterance length, compared with conceptual factors like concreteness and 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 (9). On the other hand, on syntactic bootstrapping theories (10), 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 (11).

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 grammar-related 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. Alternatively, 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.

By incorporating a variety of theoretically-important factors, basing our analysis on a large sample 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 CDI data from Wordbank to estimate the age of acquisition for words across 10 languages: Croatian, Danish, English, French (Quebec), Italian, Norwegian, Russian, Spanish, Swedish, Turkish. We then ask what factors are most important for predicting this age of acquisition. Table **??** gives an overview of our data sources.

Estimating Age of Acquisition. To estimate the age at which words are acquired, we used vocabulary data collected using the MacArthur-Bates Communicative Development Inventory, specifically the Words & Gestures (infant) form for 8- to 18-month-olds. When filling out a CDI 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 who were 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 determined when the curve crosses 0.5, i.e. at what age at least 50% of children are reported to understand the word. Following (1), we take this point to be each word's age of acquisition.

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 seven languages, we obtained an estimate of its frequency in child-directed speech, the mean length of utterances in which it appears in child-directed speech, its length in characters, and ratings of its concreteness, valence, arousal, and relevance to babies. Items such as *child's own name* were excluded. Example words for these predictors in English are shown in Table 1.

Frequency and MLU are measured relative to the word's language. But since existing 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 in `R` (6).

Each numeric predictor was centered and scaled so that all predictors would have comparable units. Lexical category was determined on the basis of the conceptual categories presented on the CDI form (e.g., "Animals"), such that the Nouns category contains common nouns, Predicates contains verbs and adjectives, Function Words contains closed-class words, and Other contains the remaining items (12).

Frequency. For each language, we estimated word frequency from unigram counts based on all corpora in CHILDES 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 (e.g., *orange* as in color or fruit), occurrences of the word in the corpus were split uniformly between the senses on the CDI. Counts were normalized to the length of each corpus and then log transformed.

MLU. For each language, we estimated each word's MLU by calculating the mean length in words of the utterances in which that word appeared, for all corpora in CHILDES for

Predictor	Value	Words
frequency	highest	you, it, that
	lowest	cockadoodledoo, grrr, church
MLU	highest	when (question), day, store
	lowest	peekaboo, ouch, hello
final_frequency	highest	book, it, there
	lowest	give, when (question), put
solo_frequency	highest	no, yes, what
	lowest	tooth, feed, aunt
length	highest	cockadoodledoo, refrigerator, rocking chair
	lowest	i, go, hi
concreteness	highest	apple, baby, ball
	lowest	how, now, that
valence	highest	happy, hug, love
	lowest	sick, hurt (description), ouch
arousal	highest	naughty, money, scared
	lowest	shh, asleep, blanket
babiness	highest	baby, bib, bottle
	lowest	donkey, penny, jeans

Table 1. Examples of words with the lowest and highest values for each predictor.

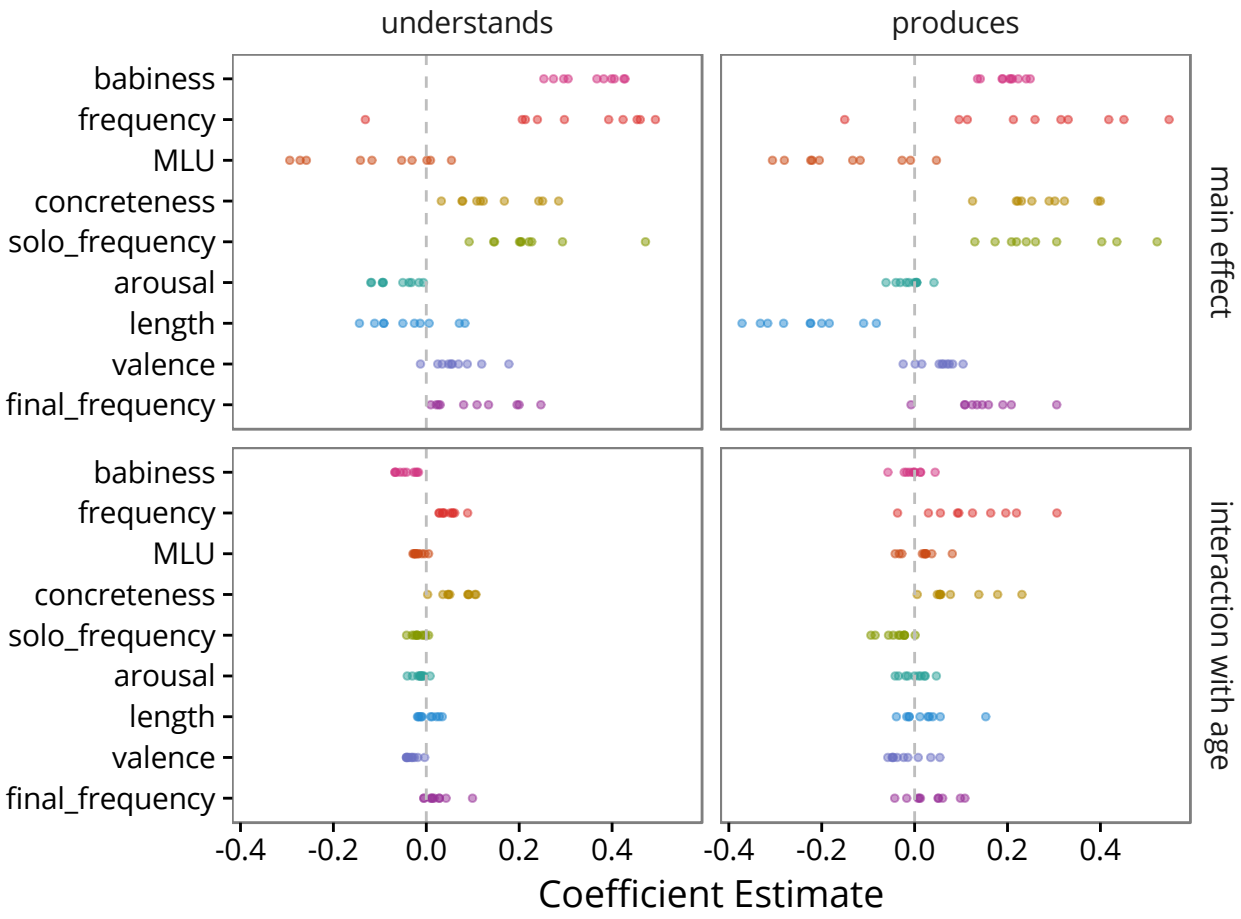
understands
produces
babiness
frequency
MLU
concreteness
solo_frequency
arousal
length
valence
final_frequency
main effect
interaction with age
-0.4
-0.2
0.0
0.2
0.4
Coefficient Estimate

Fig. 1. TODO

that language. Words that only occurred in one utterance 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 (13).

Concreteness. We used previously collected norms for concreteness (14), 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 TODO CDI words that were not part of the collected norms, we imputed ratings from the mean of all CDI words' ratings.

Valence and Arousal. We also used previously collected norms for valence and arousal (15), for which adult participants were asked to rate words on a 1-9 happy-unhappy scale (valence) and 1-9 excited-calm scale (arousal). For the TODO CDI words that were not part of the collected norms (mostly function words), 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 (4) for which adult participants were asked to judge a word's relevance to babies.

Analysis

An overview of our entire dataset can be seen in Figure **??**, 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 over development, 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 **??** 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 utterances. Concreteness exhibits a U-shaped trajectory, with the earliest learned words actually being relatively abstract (e.g., social routines and animal sounds).

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 languages. For illustrative purposes, Figure **??** shows the predictions of the English model plotted against the empirical AoA estimates.

Figure 1 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 full, cross-linguistic model. Overall there is 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 looks unusually weak. There is also variability in the overall fit of the models to the data, with some languages (e.g., Norwegian), having much more of the variance explained than others (e.g., Turkish).

A potential concern for comparing these coefficient estimates is predictor collinearity. Fortunately, in every language, the only high correlations were between frequency and number of characters, a reflection of Zipf's Law (16), and between frequency and concreteness, probably as a consequence of the complexity bias (13).

Lexical Category. Previous work gives reason to believe that predictors' relationship with age of acquisition differs among various lexical categories (1). To investigate these effects, we separated our data by lexical category and fit separate linear mixed-effects models for each, limiting the predictors to the four that were significantly predictive overall. Figure **??** shows the resulting coefficient estimates. 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.

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 (1). Scale comes at a cost in terms of detail, however, since the availability of both predictors and outcome data has been quite limited.

By including seven 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 changed in relative importance across development. For example, certain concepts that were more strongly associated with babies appeared to be learned early for children across languages (17).

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.

Lastly, the predictors varied in strength across lexical categories. Frequent, concrete nouns were learned earlier, consistent with theories that emphasize the importance of early referential speech (18). 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 (9, 11).

Despite its larger scope, our work 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 analyses (5), 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 specific predicates).

Finally, our work underscores the incompleteness of the current understanding of vocabulary development. Even for English, the language in which our model captures the most variance ($r^2 = 0.29$), much still remains unexplained. Furthermore, this variance is highly reliable—cross-validation using half of the English-speaking children to predict ages of acquisition for the other half yields $r^2 = 0.98$. This gap highlights an important theoretical challenge in the study of early language: linking individual datapoints to the broader patterns of acquisition. We have strong theories of how individual learning situations proceed, but must unify these theories to make progress on understanding language learning at scale.

Supporting Information (SI). TODO

ACKNOWLEDGMENTS. TODO

1. Goodman JC, Dale PS, Li P (2008) Does frequency count? Parental input and the acquisition of vocabulary. *Journal of child language* 35(3):515.

2. Hills TT, 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.

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

4. Perry LK, 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.

5. Roy BC, Frank MC, 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.

6. Frank MC, Braginsky M, Yurovsky D, Marchman VA (in press) Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*.

7. Fenson L (2007) *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual* (Paul H. Brookes Publishing Company).

8. MacWhinney B (2000) *The CHILDES project: The database* (Psychology Press).

9. Gentner D, Boroditsky L (2001) Individuation, relativity, and early word learning. *Language Acquisition and Conceptual Development* (Cambridge University Press).

10. Gleitman L (1990) The structural sources of verb meanings. *Language acquisition* 1(1):3–55.

11. Snedeker J, Geren J, Shafto CL (2007) Starting over: International adoption as a natural experiment in language development. *Psychological science* 18(1):79–87.

12. Bates E, et al. (1994) Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language* 21(01):85–123.

13. Lewis ML, Frank MC (under review) The length of words reflects their conceptual complexity.

14. Brysbaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior research methods* 46(3):904–911.

15. Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods* 45(4):1191–1207.

16. Zipf GK (1935) The psycho-biology of language.

17. Tardif T, et al. (2008) Baby's first 10 words. *Developmental Psychology* 44(4):929.

18. Baldwin DA (1995) Understanding the link between joint attention and language. *Joint attention: Its origins and role in development*:131–158.