Consistency and variability in word learning across languages

Mika Braginsky[1], Daniel Yurovsky[2], Virginia A. Marchman[3], & Michael C. Frank[3]

[1] Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology

[2] Department of Psychology, University of Chicago

[3] Department of Psychology, Stanford University

Author Note

Correspondence concerning this article should be addressed to Mika Braginsky. E-mail: mikabr@mit.edu

Abstract

Why do children learn some words earlier than others? The order in which words are acquired can provide clues about the mechanisms of word learning. In a large-scale corpus analysis, we use data from over 38,000 children to estimate the acquisition trajectories of around 400 words in ten languages, predicting them on the basis of independently-derived environmental and conceptual factors. We examine the consistency and variability of these predictors across languages, by lexical category, and over development. The ordering of predictors across languages is quite similar, suggesting similar processes in operation. In contrast, the ordering of predictors across different lexical categories is distinct, in line with theories that posit distinct factors at play in the acquisition of content words and function words. By leveraging data at a significantly larger scale than previous work, our analyses identify candidate generalizations about the processes underlying word learning across languages.

Keywords: word learning, language acquisition, corpus analysis

Consistency and variability in word learning across languages

In spite of tremendous individual variation in rate of development (1), the first words that children utter are strikingly consistent (2, 3): they tend to talk about important people in their life ("mom", "dad"), social routines ("hi", "uh oh"), animals ("dog", "duck"), and foods ("milk", "banana"). As children learn from their experiences and according to their own interests (**???**, 4), their vocabulary grows rapidly, typically adding more nouns, but also verbs ("go") and other predicates ("hot") to their production repertoires. Over just their first three years, children learn hundreds or even thousands of words (5, 6). Why are some words learned so early and some much later?

This simple question about the order of the acquisition of first words can provide a window into the nature of children's language learning. Posed as a statistical problem, the challenge is to find what variables best predict the age at which different words are acquired. Previous work using this approach has revealed that in English, within a lexical category (e.g., nouns, verbs), words that are more frequent in speech to children are likely to be learned earlier (7). And further studies have found evidence that a variety of other semantic and linguistic factors are related to word acquisition (8–12).

But these exciting findings are limited in their generality because they use different datasets, focus on different predictors, and almost exclusively analyze English data. They do not allow for cross-linguistic comparison of the relative importance of the many relevant factors under consideration in children from different language communities. Such cross-linguistic comparisons are critical, as identifying commonalities (and differences) across languages is our best strategy for uncovering the universal mechanisms that are in play for all children and differentiating them from patterns of acquisition that emerge due to the particulars of a given language or culture (13, 14). Our goal is to extend these classic approaches by assessing the degree to which the predictors of word learning are consistent across different languages and cultures, as well as whether there are similar patterns across different word types (e.g., nouns vs. verbs).

To conduct cross-linguistic comparisons, we rely on data from Wordbank (wordbank.stanford.edu) (15). Wordbank is an open repository of cross-linguistic language development data that aggregates administrations of the MacArthur-Bates Communicative Development Inventory (1), a family of parent-report vocabulary checklists, and currently includes data from more than 60,000 children. By aggregating across large samples, idiosyncratic patterns of individual variation are less likely to overpower general trends in which words are relatively easy or hard to learn, allowing us to investigate what features affect their acquisition across languages.

In our analyses, we integrate estimates of words' acquisition trajectories from Wordbank with characterizations of the word learning environment from the CHILDES database (16) and elsewhere, a data-integration methodology originated by (7). Using this approach, we examine the impact of multiple environmental and conceptual factors. To measure environmental input, we estimated each word's frequency, mean length of utterance (MLU), frequency as sole utterance constituent, and frequency in utterance-final position. To measure conceptual factors, we obtained ratings of each word's concreteness, valence, arousal, and association with babies from previously collected norms. We used these measures to predict words' acquisition trajectories, assessing the relative contributions of each, as well as how they change over development and interact with lexical category. These analyses address two important questions.

First, we ask about the degree of consistency across languages in the relative importance of predictors. Consistency in the ordering of predictors would suggest that similar information sources are important for learners, regardless of language. Such evidence would point to the importance of similar learning mechanisms across languages despite superficial dissimilarities (e.g., greater morphological complexity in Russian and Turkish, greater phonological complexity in Danish). Conversely, variability would reveal the degree to which learners face different challenges in learning different languages.

Second, we ask which lexical categories are most influenced by environmental factors,

like frequency and utterance length, compared with conceptual factors like concreteness and valence. Division of dominance theory suggests that nouns might be more sensitive to conceptual factors, while predicates and closed-class words might be more sensitive to environmental (linguistic) factors (17). And on syntactic bootstrapping theories (18), 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) (19). Thus, examining the relative contribution of different predictors across lexical categories can help test the predictions of influential theories of acquisition.

Approach

To estimate the trajectory of words' acquisition, we used vocabulary data collected using CDI instruments adapted in many different languages, including both Words & Gestures (WG) forms for younger children and Words & Sentences (WS) forms for older children. When filling out a CDI form, parents are either asked to indicate whether their child "understands" (comprehension) or "understands and says" (production) each of around 400-700 words. Typically, both comprehension and production are queried for younger children and only production is queried for older children, but details vary from adaptation to adaptation. We use data from the items on the WG form for our comprehension measure, and data from the items in common between the WG and WS forms for our production measure. Table 1 gives an overview of our acquisition data (20–31).

For each word, CDI data yield a trajectory reflecting the number of children that are reported to produce or understand the word at each age covered by the instrument (see Figure 1 for some examples). We then use a mixed-effects logistic regression model to predict whether each child knows the word on the basis of the child's age, properties of the word, and interactions between age and each property of the word. We also fit such models separately to the words in each lexical category. The magnitude of the standardized coefficient on each feature gives an estimate of its importance in predicting whether words

are learned earlier or later. Interactions between features and age give estimates of how this effect is modulated for earlier and later-learned words. For example, a positive effect of association with babies ("babiness") means that words associated with babies are learned earlier; a negative interaction with age means that high babiness primarily leads to higher rates of production and comprehension for younger children.

Each of our predictors is derived from independent sources. For each word that appears on the CDI forms in each of our 10 languages, we used corpora of child-directed speech in that language to obtain an estimate of its frequency, the mean length of utterances in which it appears, its frequency as the sole consituent of utterance, and its frequency in utterance final position (with frequency residualized out of solo and final frequencies). Additionally, we computed each word's length in characters and included ratings of its concreteness, valence, arousal, and relatedness to babies. Since existing datasets for conceptual ratings are primarily available for English, we mapped all words onto translation equivalents across CDI forms, verified by native speaker judgements, 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. Example words for these predictors in English are shown in Table 2.

A potential concern for comparing coefficient estimates is predictor collinearity. Fortunately, in every language, the highest correlations were between MLU and solo frequency (mean over languages and measures $r =$ -0.50), as expected given the similarity of these factors; and frequency and number of characters (mean over languages and measures $r =$ -0.40), a reflection of Zipf's Law (32).

Results

Figure 2 shows the coefficient estimate for each predictor in each language. We find that babiness, frequency, MLU, and concreteness are relatively stronger predictors of age of acquisition across languages. Given the emphasis on frequency effects in the language

acquisition literature (33), one might have expected frequency to dominate, but several other predictors are just as strong in this analysis. Some factors previously argued to be important for word learning, namely valence and arousal (34), appear to have limited relevance when compared to other factors.

Overall, there is considerable consistency in the magnitudes of predictors across languages. In almost all, babiness and frequency were highest, while valence and arousal were smaller. A priori it could have been the case that different languages have wildly different effects of experiential vs. structural factors, but this pattern is not what we observe. Instead, Figure 3 shows the mean pairwise correlation of predictor coefficients across languages (i.e., the correlation of coefficients for English with coefficients for Russian, for Spanish, and so on). These means – and even the individual datapoints – are far outside of bootstrapped estimates for the average pairwise correlation in a randomized baseline created by shuffling predictor coefficients within language, suggesting that coefficient estimates are far more consistent across languages than would be expected by chance.

Word length is the one predictor of acquisition that varied substantially between measures, in that it is far more predictive for production than comprehension. Thus as measured here, length seems to be playing more the role of production constraints (i.e., how difficult a word is to say) than comprehension constraints (i.e., how difficult it is to store or access).

Next, we wanted to examine how the relative contributions of the predictors changes over development. Across languages, positive age interactions can be seen for concreteness and frequency (i.e., their effects increase with age). Conversely, there are negative age interactions for babiness and valence in comprehension and for solo frequency in production, suggesting stronger effects in words learned earlier in development.

Previous work gives reason to believe that predictors' relationship with age of acquisition differs among various lexical categories (7). To investigate these effects, we separated our data by lexical category and fit separate models for each category. Figure 4

shows the resulting coefficient estimates. Across languages, frequency had the highest magnitude for nouns and a lower magnitude for function words. In contrast, MLU was almost irrelevant for both nouns and predicates, but highly predictive for function words. These patterns are supportive of the hypothesis that different word classes are learned in different ways, or at least that the bottleneck on learning tends to be different, leading to different information sources being more or less important across categories.

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 to be measured across ages and lexical categories (7, 8, 12). Such work has identified a number of candidate predictors of word learning. By including 10 languages and 9 predictors, our work expands the scope of these studies dramatically, leading to several new findings.

First, we found consistency in the ordering of predictors across languages at a level substantially greater than the predictions of a chance model. 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. Instead, they are likely produced by processes that are similar across populations and languages. Such processes could include learning mechanisms or biases internal to children, or interactional dynamics between children or caregivers. We believe these consistencies should be an important topic for future investigation.

Second, predictors varied substantially in their weights across lexical categories. Frequent, concrete nouns were learned earlier, consistent with theories that emphasize the importance of early referential speech (35). But for predicates, concreteness was somewhat

less important. And for function words, MLU was most predictive, perhaps because it is easiest to decode the meanings of function words that are used in short sentences (or because such words have meanings that are easiest to decode). Overall these findings are consistent with some predictions of both division of dominance theory, which highlights the role of conceptual structure in noun acquisition (17), and syntactic bootstrapping theory, which emphasizes linguistic structure over conceptual complexity in the acquisition of lexical categories other than nouns (19). More generally, our methods here provide a way forward for testing the predictions of these theories across languages and at the level of the entire lexicon rather than individual words.

In addition to these new insights, several findings emerge that confirm and expand previous reports. Environmental frequency was an important predictor of learning, with more frequently heard words learned earlier (7, 12). Predictors also changed in relative importance across development. For example, certain words whose meanings were more strongly associated with babies appeared to be learned early for children across the languages in our sample (2). Finally, word length showed a disassociation between comprehension and production, suggesting that challenges in production do not carry over to comprehension (at least in parent-report data).

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 (11), 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 data are observations gleaned from parent report and are subject to both causal confounding and confounding via biases in parent observation. Thus, our conclusions will require further

testing through converging evidence from both laboratory experiments and direct observation.

In sum, by examining predictors of early word learning across languages, we identified substantial cross-linguistic consistency in the factors contributing to the ease or difficulty of learning individual words. These findings testify to the importance of building open, shared resources in the study of language learning – without the efforts of many research groups across many language communities, such studies would be impossible. In addition, we hope that our work here provides a baseline for the building of future predictive models that allow theories of language learning to be tested at scale.

Materials and Methods

All code and data to reproduce our analyses are available at `https://github.com/mikabr/aoa-prediction`.

Predictor variables

Each numeric predictor was centered and scaled so that all predictors would have comparable units. For each predictor, missing values (CDI items that were not in the relevant corpus or norms) were imputed from the mean for their respective language and measure. Placeholder items, such as "child's own name", were excluded.

Translation equivalents are available in the Wordbank database using the `wordbankr` package in `R` (15). Translation equivalents were constructed by the authors and independently hand-checked by native speakers.

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, Laplace smoothed, and then log transformed.

Solo and Final Frequencies. Using the same dataset as for frequency, we estimated the frequency with which each of word occurs as the sole word in an utterance, and the frequency with which it appears as the final word of an utterance (not counting single-word utterances). As with frequency, solo and final counts were normalized to the length of each corpus, Laplace smoothed, and log transformed. Since both of these estimates are by necessity highly correlated with frequency, we then residualized unigram frequency out of both of them, so that values reflect an estimate of the effect of solo/final frequency over and above frequency itself.

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 that language. For words that occurred fewer than 10 times, MLU estimates were not used (i.e. treated as missing).

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 (36).

Concreteness. We used previously collected norms for concreteness (37), 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.

Valence and Arousal. We also used previously collected norms for valence and arousal (38), for which adult participants were asked to rate words on a 1-9 happy-unhappy scale (valence) and 1-9 excited-calm scale (arousal).

Babiness. Lastly, we used previously collected norms of "babiness", a measure of association with infancy (10) for which adult participants were asked to judge a word's association with babies on a 1-10 scale.

Lexical category

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 (39).

Analysis

For all analyses, we used logistic mixed-effects regression models (fit with `lme4 1.1-12` in `R`) to predict whether each child understands/produces each word from age, the above predictors, and the interactions of age with the predictors. Each model was fit to all data

from a particular language community and included a random intercept for each word and a random slope of age for each word.

Acknowledgements

Thank you to the labs and individuals who contributed data to Wordbank and to NSF BCS #1528526 for support.

Author Contributions

M.B. and D.Y. conducted data processing and analysis, with supervision from V.A.M. and M.C.F.; all authors contributed to writing the paper.

Competing Interest Statement

The authors declare no conflict of interest.

References

1. Fenson L, et al. (2007) *MacArthur-Bates Communicative Development Inventories* (Brookes Publishing Company).
2. Tardif T, et al. (2008) Baby's first 10 words. *Developmental Psychology* 44(4):929.
3. Schneider R, Yurovsky D, Frank MC (2015) Large-scale investigations of variability in children's first words. *Proceedings of the Cognitive Science Society*.
4. Mayor J, Plunkett K (2014) Shared understanding and idiosyncratic expression in early vocabularies. *Developmental science* 17(3):412–423.
5. Fenson L, et al. (1994) Variability in early communicative development. *Monogr Soc Res Child Dev* 59(5).
6. Mayor J, Plunkett K (2011) A statistical estimate of infant and toddler vocabulary size from CDI analysis. *Dev Sci* 14(4):769–785.
7. Goodman JC, Dale PS, Li P (2008) Does frequency count? Parental input and the acquisuisition of vocabulary. *J Child Lang* 35(3):515.
8. Hills TT, Maouene M, Maouene J, Sheya A, Smith L (2009) Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisuisition? *Psychol Sci* 20(6):729–739.
9. Stokes SF (2010) Neighborhood density and word frequency predict vocabulary size in toddlers. *J Speech Lang Hear Res* 53(3):670–683.
10. Perry LK, Perlman M, Lupyan G (2015) Iconicity in English and Spanish and its relation to lexical category and age of acquisuisition. *PloS One* 10(9):e0137147.
11. Roy BC, Frank MC, DeCamp P, Miller M, Roy D (2015) Predicting the birth of a spoken word. *Proc Natl Acad Sci* 112(41):12663–12668.
12. Swingley D, Humphrey C (2017) Quantitative linguistic predictors of infants' learning of specific English words. *Chi Dev*.
13. Slobin DI (1985) *The crosslinguistic study of language acquisition: Theoretical issues*

(Psychology Press).

14. Bates E, MacWhinney B (1987) Competition, variation, and language learning. *Mech of Lang Acquis*:157–193.
15. Frank MC, Braginsky M, Yurovsky D, Marchman VA (2016) Wordbank: An open repository for developmental vocabulary data. *J Child Lang*.
16. MacWhinney B (2000) *The CHILDES project: The database* (Psychology Press).
17. Gentner D, Boroditsky L (2001) Individuation, relativity, and early word learning. *Lang Acquis and Concept Dev* (Cambridge University Press).
18. Gleitman L (1990) The structural sources of verb meanings. *Lang Acquis* 1(1):3–55.
19. Snedeker J, Geren J, Shafto CL (2007) Starting over: International adoption as a natural experiment in language development. *Psychol Sci* 18(1):79–87.
20. Kovacevic M, Babic Z, Brozovic B (1996) A Croatian language parent report study: Lexical and grammatical development. *Seventh International Congress for the Study of Child Language, Istanbul, Turkey*.
21. Bleses D, et al. (2008) The Danish Communicative Developmental Inventories: Validity and main developmental trends. *J Child Lang* 35(03):651–669.
22. Boudreault M, Cabirol E, Poulin-Dubois D, Sutton A, Trudeau N (2007) MacArthur Communicative Development Inventories: Validity and preliminary normative data. *La Revue d'Orthophonie et d'Audiologie* 31(1):27–37.
23. Trudeau N, Sutton A (2011) Expressive vocabulary and early grammar of 16-to 30-month-old children acquisuiring Quebec French. *First Lang* 31(4):480–507.
24. Caselli MC, Rinaldi P, Stefanini S, Volterra V (2012) Early action and gesture "vocabulary" and its relation with word comprehension and production. *Chi Dev* 83(2):526–542.
25. Caselli MC, et al. (1995) A cross-linguistic study of early lexical development. *Cog Dev* 10(2):159–199.
26. Simonsen HG, Kristoffersen KE, Bleses D, Wehberg S, Jørgensen RN (2014) The

Norwegian communicative development inventories: Reliability, main developmental trends and gender differences. *First Lang* 34(1):3–23.

27. Vershinina E, Yeliseyeva M (2011) Some norms of speech development of children from 8 to 18 months. *Special Education*.

28. Yeliseyeva M, Vershinina E (2009) Some norms of speech development of children from 18 to 36 months (based on the materials of the MacArthur survey). *Problems of Developmental Linguistics, Saint-Petersburg*, p 22.

29. Jackson-Maldonado D, et al. (2003) *MacArthur Inventarios del Desarrollo de Habilidades Comunicativas: User's guide and technical manual* (Brookes Publishing Company).

30. Eriksson M, Berglund E (2002) *Instruments, scoring manual and percentile levels of the Swedish Early Communicative Development Inventory, SECDI* (Högskolan i Gävle).

31. Acarlar F, et al. (2008) Adapting MB-CDI to Turkish: The first phase. *Essays of Turkish Linguistics: Proceedings of the 14th International Conference on Turkish Linguistics*, pp 6–8.

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

33. Ambridge B, Kidd E, Rowland CF, Theakston AL (2015) The ubiquity of frequency effects in first language acquisuisition. *J Child Lang* 42(02):239–273.

34. Moors A, et al. (2013) Norms of valence, arousal, dominance, and age of acquisuisition for 4,300 dutch words. *Behav Res Meth* 45(1):169–177.

35. Baldwin DA (1995) Understanding the link between joint attention and language. *Joint Attention*:131–158.

36. Lewis ML, Frank MC (2016) The length of words reflects their conceptual complexity. *Cognition* 153:182–195.

37. Brysbaert M, Warriner AB, Kuperman V (2014) Concreteness ratings for 40 thousand generally known English word lemmas. *Behav Res Meth* 46(3):904–911.

38. Warriner AB, Kuperman V, Brysbaert M (2013) Norms of valence, arousal, and

358 dominance for 13,915 English lemmas. *Behav Res Meth* 45(4):1191–1207.

359 39. Bates E, et al. (1994) Developmental and stylistic variation in the composition of early

360 vocabulary. *J Child Lang* 21(01):85–123.

Language	CDI Items	N production	N comprehension
Croatian	390	627	250
Danish	383	6,112	2,398
English	393	5,967	1,821
French	396	1,364	537
Italian	396	1,400	648
Norwegian	381	7,466	2,374
Russian	410	1,805	768
Spanish	399	1,872	778
Swedish	371	1,367	467
Turkish	396	3,537	1,115

Table 1

Dataset statistics for acquisition data from Wordbank.

Predictor	Highest	Lowest
Babiness	baby, bib, bottle	donkey, penny, jeans
MLU	when, day, store	peekaboo, ouch, hello
Frequency	you, it, that	cockadoodledoo, grrr, church
Concreteness	apple, baby, ball	how, now, that
Solo frequency	no, yes, what	tooth, feed, aunt
Arousal	naughty, money, scared	shh, asleep, blanket
Length	cockadoodledoo, refrigerator, rocking chair	i, go, hi
Valence	happy, hug, love	sick, hurt, ouch
Final frequency	book, it, there	give, when, put

Table 2

Items with the highest and lowest values for each predictor in English.

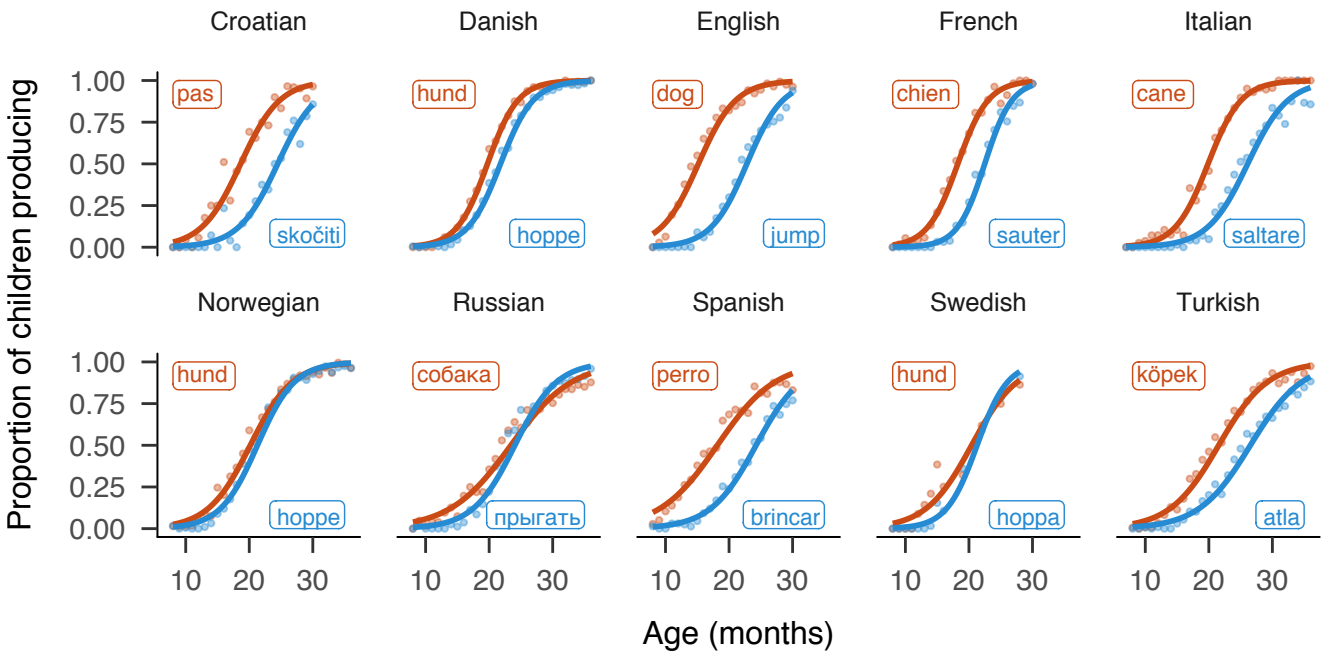

Figure 1. Example production trajectories for the words "dog" and "jump" across languages. Points show average proportion of children producing each word for each one-month age group. Lines show the best-fitting logistic curve. Labels show the forms of the word in each language.

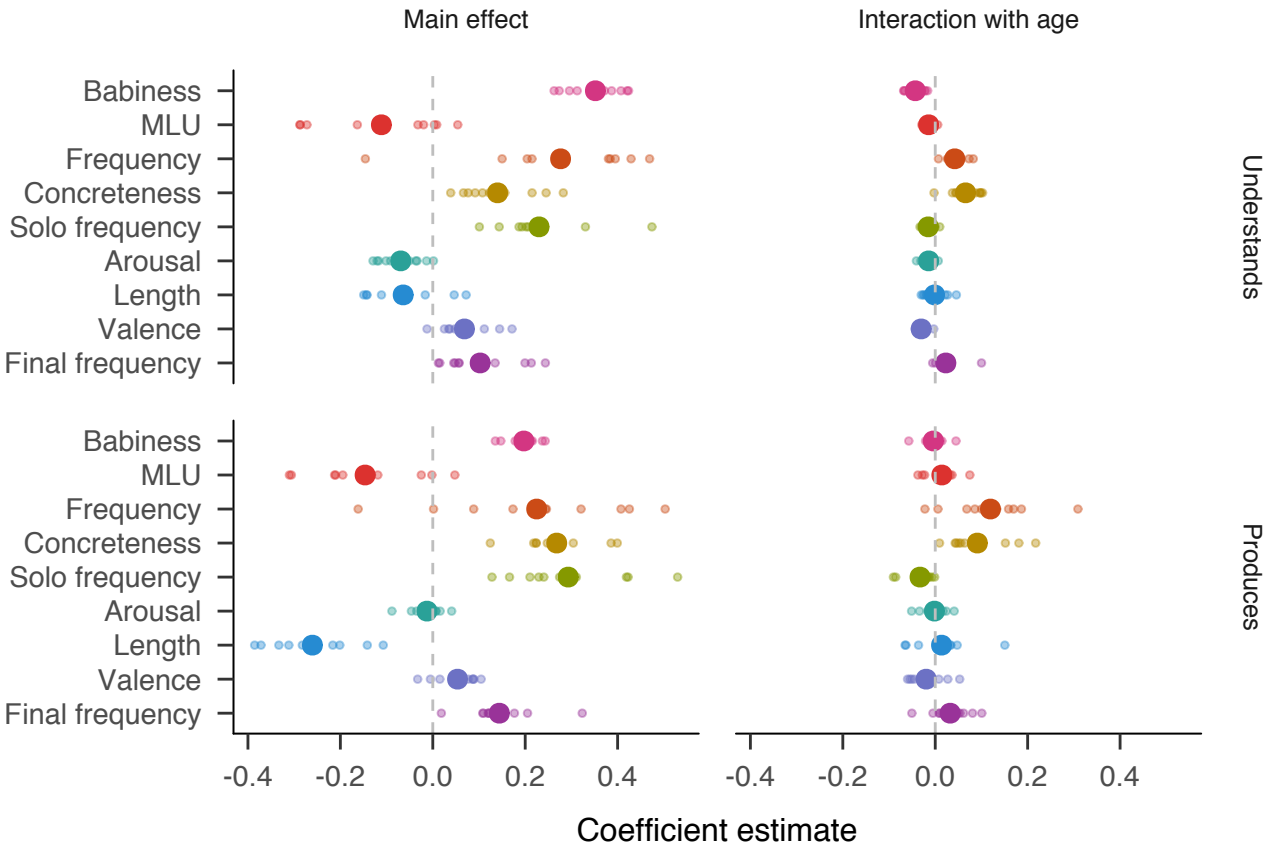

Figure 2. Estimates of coefficients in predicting words' developmental trajectories. Each point represents a predictor's coefficient in one language, with the large point showing the mean across languages. Larger coefficient values indicate a greater effect of the predictor on acquisition: positive main effects indicate that words with higher values of the predictor tend to be understood/produced by more children, while negative main effects indicate that words with lower values of the predictor tend to be understood/produced by more children; positive age interactions indicate that the predictor's effect increases with age, while negative age interactions indicate the predictor's effect decreases with age.

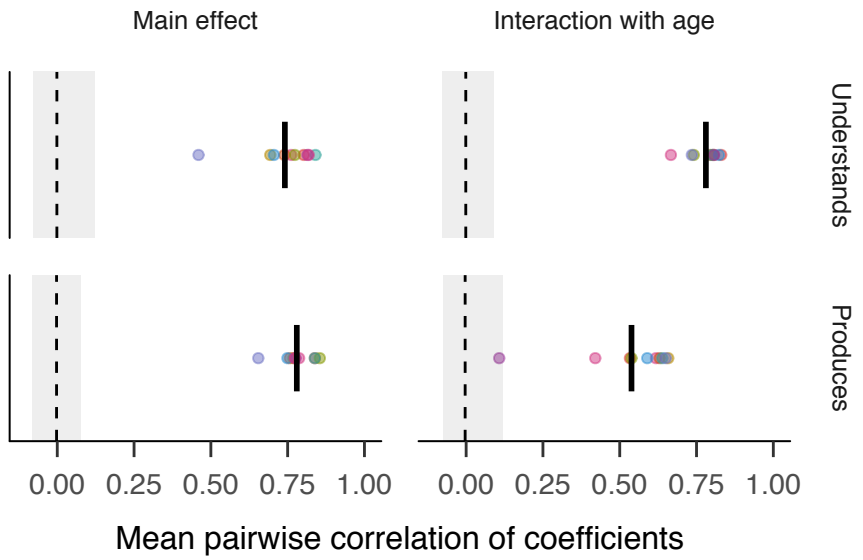

Figure 3. Correlations of coefficients estimates between languages. Each point represents the mean of one language's coefficients' correlation with each other language's coefficients, with the black line indicating the overall mean across languages. The grey region and dashed line show a bootstrapped 95% confidence interval of a randomized baseline where predictor coefficients are shuffled within language.

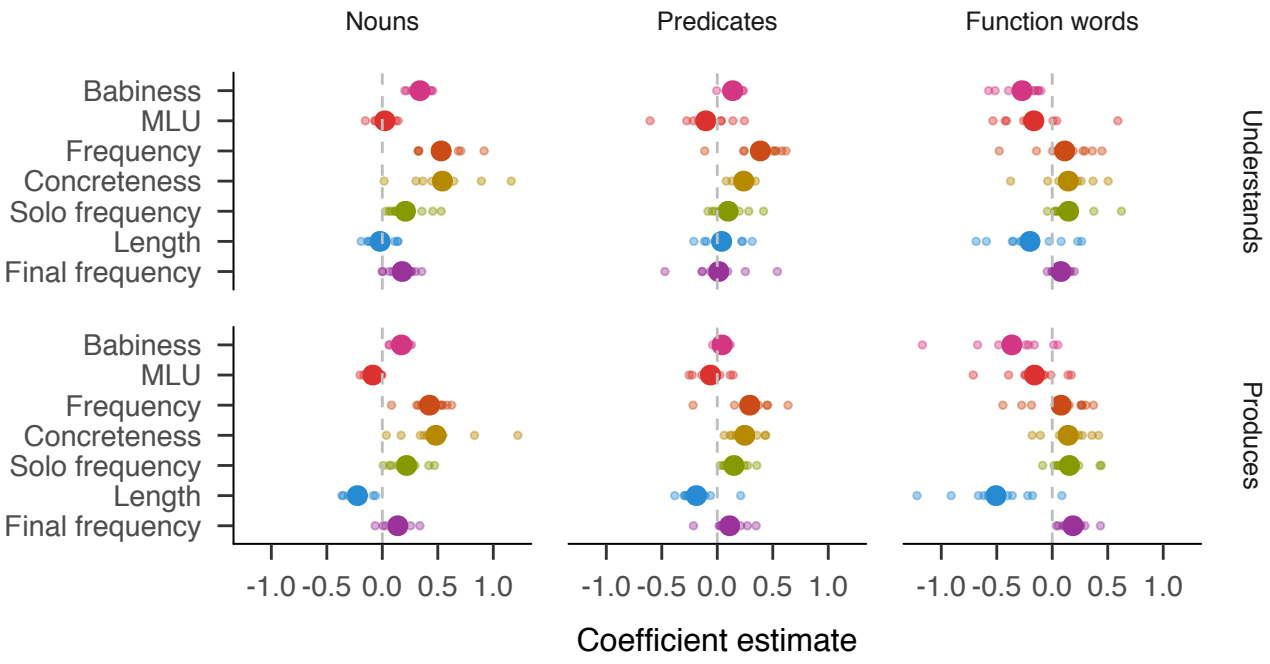

Figure 4. Estimates of coefficients in predicting words' developmental trajectories (as described in Figure 2), with separate models for each lexical category.