Toward a "Standard Model" of Early Language Learning

George Kachergis[1], Virginia A. Marchman[1], & Michael C. Frank[1]

[1] Stanford University

Author Note

Department of Psychology

Correspondence concerning this article should be addressed to Michael C. Frank, 450 Jane Stanford Way, Stanford, CA 94305. E-mail: mcfrank@stanford.edu

Abstract

A "standard model" is a theoretical framework that synthesizes observables into a quantitative consensus. Have we made progress towards this kind of synthesis for children's early language learning? Many computational models of early vocabulary learning assume that individual words are learned through an accumulation of environmental input. This assumption is also implicit in empirical work that emphasizes links between language input and learning outcomes. However, models have typically focused on average performance, while empirical work has focused on variability. To model individual variability, we relate the tradition of research on accumulator models to Item-Response Theory models from psychometrics. This formal connection reveals that currently available datasets do not allow us to test these models fully, illustrating a critical need for theory in shaping new data collection and in creating and testing an eventual "standard model."

Keywords: early language learning; language acquisition; vocabulary development

Word count: 3352

Toward a “Standard Model” of Early Language Learning

Introduction

Early language learning is a key challenge for cognitive science: how do we go from speechless, wordless infants to children who can use language expressively and creatively? The field of early language is often portrayed as mired in controversies around issues of innateness. We see a new synthesis emerging, however, based on theoretical and empirical work on the growth of vocabulary. Our goal is to present this synthesis as the beginnings of a “standard model”[1]: a baseline theory that is accepted widely in its outlines and that should guide future work, even though it is incomplete and its assumptions still require rigorous evaluation.

In physics, the Standard Model is a widely-accepted theory from the 1970s that describes all known elementary particles along with three of the four known fundamental forces in the universe. Although incomplete and even incorrect in places, physics’ Standard Model nonetheless explains a wide variety of empirical phenomena, allowing scientists to model physical interactions with great precision and accuracy and to design tests that inform theory revision. Psychology, in general, has been criticized for lacking such formal theories that inform and drive empirical research (Muthukrishna & Henrich, 2019).

Is it possible to build a standard model for the field of language acquisition? Computational models are one potential source of unifying theory, and perhaps the only way to fully specify and quantitatively compare theories. Although models of language learning vary widely, many presuppose a common framework (Figure 1A). Its core principle is that language input accumulates via repeated exposure, resulting in learning. Although there is evidence for other factors relating input *quality* to uptake (e.g., lexical diversity:

[1] “Standard” here is not meant to be normative or prescriptive, but rather that the model is an accepted starting point for a description of early language learning. We see great value in there being a plurality of approaches to the study of early language learning, and seek only to begin gathering some related threads.

Jones & Rowland, 2017; referential transparency: Cartmill et al., 2013), a framework focusing on input *quantity* underpins much of the broader policy discussion of links between environmental stimulation and individual variability in children's outcomes (Hart & Risley, 1995 et seq.). Moreover, this framework corresponds nicely with a group of computational models that we refer to jointly as "accumulator models", which make distributional assumptions about the difficulty of words and are sometimes fit to words' average age-of-acquisition, but not to data from individual children (Hidaka, 2013; McMurray, 2007; Mollica & Piantadosi, 2017).

Our aim here is to make an explicit connection between accumulator models and the broader, but less formal, discussion of the role of language input in learning. We use Item-Response Theory (IRT) from psychometric testing as a framework for connecting empirical data about language input with accumulator models. Sitting at a level of analysis between generic regressions and cognitive process models, data-analytic cognitive models of this type offer a way to implement our verbal theories and evaluate them quantitatively, with parameters mapping to both real-world, measurable units (e.g., words heard per hour) and to psychological constructs (e.g., a child's ability to process language).

Discussions of language input typically focus on variation between children. In contrast, models of acquisition (ours included) typically focus on average patterns of acquisition (but see Jones & Rowland, 2017). By connecting such models with psychometric models that are intended to capture variation, we hope to bring individual variability back to the center of language learning theory.

Importantly, the goal is not to produce the "correct" model, but rather to explore the ways that model assumptions lead to predictions about specific patterns of data. In fact, the greatest value of such models is the ability to identify the areas of greatest mismatch between data and model, highlighting areas requiring further investigation (Tauber, Navarro, Perfors, & Steyvers, 2017). Perhaps such iterative work will lead to a true standard model for language learning, which captures differences not only in quantity, but also in the quality of various activities and sources of input.

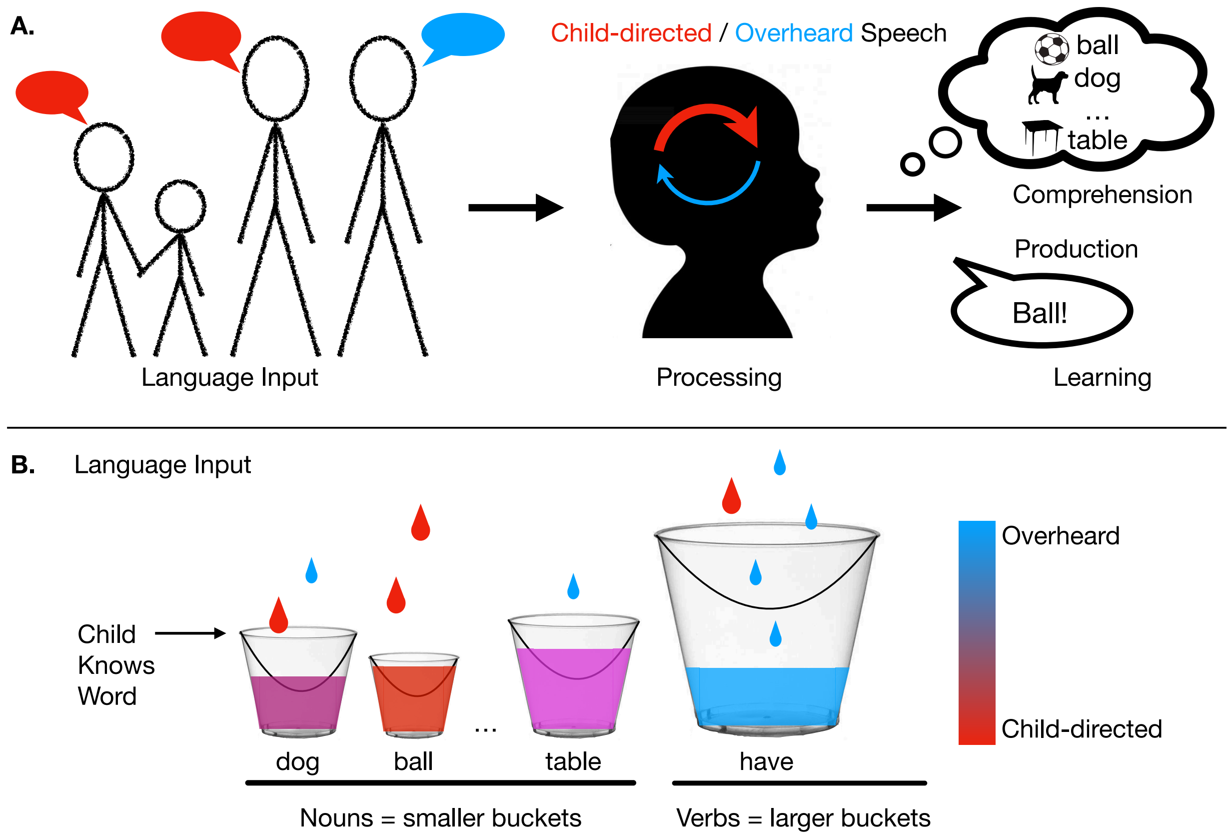

Figure 1. A) A schematic of the standard relationships between variables assumed in the literature on early language learning. It is generally assumed that child-directed speech (red) is more valuable than overheard speech (blue): it may be more attended to as well as more easily processed by the child. B) An illustration of an accumulator model: each bucket represents knowledge about a particular word, and each token is a drop in the corresponding bucket. Some words are more difficult than others (i.e., have larger buckets). When a bucket is full, the corresponding word is learned. Child-directed tokens (red) may be more valuable than overheard tokens (blue) for a variety of potential reasons. For example, overheard tokens are more likely to be ignored, child-directed tokens are more likely to be of greater interest, or are heard in more informative contexts.

Accumulator Models Are An Important Formalism For Describing Word Learning

We focus here on word learning as a central component of early language, and return to connections with morphosyntax below.[2] Accumulator models assume that experiences with words accumulate in separate registers (depicted as "buckets" in Figure 1A). When a register exceeds a particular threshold (a full bucket), the word is learned.

Accumulator models have already contributed significantly to theoretical debates. For example, McMurray (2007) elegantly demonstrated that children's "vocabulary explosion" – an acceleration in vocabulary growth in the second year – can result from the steady accumulation of word tokens without changes in the environment or learning mechanism. Other work has examined similar models using more realistic distributions of words, developmental change in learning mechanisms, and comparison to children's aggregate vocabulary growth (Hidaka, 2013; Mitchell & McMurray, 2009; Mollica & Piantadosi, 2017). Although this work is exciting, these models have not yet made direct predictions about learning in individual children (Bergelson, 2020).

One difficulty in connecting such models to data is that the relevant variables are often measured in relative, rather than absolute, units.[3] This practice is common in psychology. For example, measures of intelligence are given on a standardized scale defined

[2] What makes a word is its own separate and difficult question, especially across languages. Are "dog" and "dogs" two separate words? If not, what about "mouse" and "mice"? These questions are relatively inconsequential in morphologically-simple languages like English and Mandarin but are more difficult in morphologically-complex languages like Turkish or Inuktitut. Yet the Turkish or Inuktitut learner is accumulating *something*. Whether or not we call it a word *per se*, our hypothesis is that this unit is being accumulated and its accumulation will be subject to many of the same dynamics as words.

[3] For example, regression and correlation coefficients allow researchers to evaluate the relative strength of predictors, but often cannot be extrapolated to measurable quantities outside the experiment. In contrast, the measurement of absolute units (e.g., children's waking hours per day; words per hour of child-/other-directed speech; tokens per hour of a given word) can make real-world predictions (e.g., the

by population variability. Yet, in early language – unusually in a psychological domain – we have access to absolute units. We can count how many words a child hears and express this number as a rate that is comparable across studies (words per hour; Bergelson et al., 2019). We can similarly estimate how many words a child knows (e.g., by parents' reports of vocabulary size; Frank, Braginsky, Yurovsky, & Marchman, 2021, Ch. 5). These absolute units mean that models can make powerfully general quantitative predictions, which can be tested across different situations and populations. Few datasets have measurements of the relevant variables, however. Thus, future data collection efforts should attempt whenever possible to report measurements in absolute units.

Accumulator Models are Presupposed in the Empirical Literature

Since seminal work by Hart and Risley (1995), the connection between children's language input and the growth of vocabulary has been a topic of intense interest and debate (e.g., Sperry, Sperry, & Miller, 2019). Numerous studies have reported positive associations between the number of word tokens and types produced by caregivers (e.g., Hoff, 2003) and children's vocabulary learning – as predicted by accumulator models – though the magnitude of these relations varies across studies (Wang, Williams, Dilley, & Houston, 2020). These correlations are also partially moderated by other factors, including socioeconomic (Hoff, 2003) and genetic (Hayiou-Thomas, Dale, & Plomin, 2014) variables. Although they are costly and difficult to conduct, randomized interventions are the gold standard for estimating causal effects. When they have been done, they show modest but reliable effects (e.g., Suskind et al., 2016), providing support for a causal connection between input and outcomes.

There are many dimensions of input quality that modulate uptake including lexical diversity (Jones & Rowland, 2017), referential transparency (Cartmill et al., 2013), and

number of times a child hears "dog" each month), and allows direct comparison to other datasets.

activity context (Roy, Frank, DeCamp, Miller, & Roy, 2015). The theoretical link between input quantity and uptake is a fundamental assumption of all theories of learning, though. Reasoning from first principles, you cannot learn the word "table" if you do not hear it: input must predict learning of individual words.

Words vary in how difficult it is to learn them. For exmaple, the referent of the word "table" is often more apparent than that of the word "tomorrow," leading to easier learning. Yet traditional correlational studies typically only capture differences in input using aggregate measures of what a child may experience, asking whether variation in overall amount/quality of input relates to variation in overall vocabulary size.

An alternative approach focuses on differences in input-learning relations at the level of words, not children (Roy et al., 2015). Using regression to predict which words are easier or more difficult to learn, averaging across children (e.g., Goodman, Dale, & Li, 2008; Braginsky, Yurovsky, Marchman, & Frank, 2019), these models show a strong association between word frequency and the average age at which children acquire particular words, especially for object labels. Swingley and Humphrey (2018) extended this regression approach to predict the acquisition of individual words in individual children on the basis of the words' prevalence in their mother's speech. These findings provide convergent support for accumulator models and thereby, provide the conceptual underpinning of the relations between input (frequency for words, quantity for children) and learning.

Connecting Accumulator Models with Psychometric Models

We believe accumulator models can be the basis for an eventual "standard model" of early language. The core of our view is that individual experiences with words lead to their eventual acquisition via accumulation. Although the specific situations a word is experienced in likely vary along many dimensions that influence learning, in general, the more of these experiences a child receives, the faster their vocabulary grows. However, both children and words vary: children may learn slower or faster (perhaps also with age), and words are more or less difficult to learn. Combining these ideas, the basic hypothesis is that a child's vocabulary at a particular time should be predicted by their cumulative language exposure and learning rate, combined with the breadth of the sample of words to which they have been exposed and those words' individual difficulties.

This hypothesis describes an approach similar to Item-Response Theory (IRT; Embretson & Reise, 2013). IRT is commonly used for estimating the ability of "test takers" (here, language learners) as assessed with a particular set of "test items" (particular words). IRT models provide a convenient and broadly-used psychometric framework within which to describe and compare different model variants, which in turn represent different sets of theoretical assumptions.

In their formal structure, IRT models describe individual item responses as a function of both the difficulty of specific words and the language abilities of individual children. These latent parameters can be inferred from an observed dataset. In the basic Rasch (or 1-parameter logistic) IRT model, a person i responds correctly to item j with probability determined by their ability (θ_i) and the difficulty of item j (d_j):

$$P(y_{i,j} = 1|\theta_i, d_j) = \frac{1}{1 + e^{\theta_i + d_j}}$$

These parameters can easily be mapped onto an accumulator model: items are words (e.g., d_j is the difficulty of word j), and θ_i is child i's estimated latent language ability.

In typical IRT models, both item difficulties and person abilities are standardized (normally-distributed and 0-centered) and unit-free. In principle, however, these scores can be mapped to measurements of word frequencies and rates of children's experienced input with clear units (e.g., words heard per hour hour). This mapping can then provide a quantitative linking hypothesis between measurements of input and learning.

The equivalence of accumulator models to IRT carries a variety of benefits. First, IRT extensions can be used to explore different theoretical assumptions. For example, item-level covariates – e.g., estimates of word frequency (Braginsky et al., 2019) – and person-level covariates – e.g., age, sex, socioeconomic status – can be added with the flexibility of standard regression modeling.[4] Further, tools for multivariate IRT allow consideration of whether variation in early language is uni-dimensional or multi-factorial (Frank et al., 2021). Finally, model comparisons can be made using standardized tools, a major benefit over more ad-hoc frameworks.

Comparing Model Variants as a Method for Evaluating Theories

IRT accumulator models can be fit to data, allowing different theoretical assumptions to be compared empirically. To illustrate, we can ask: is a 2-year-old different than a 1-year-old, aside from having twice as much experience with each word? Is the amount of language experience the only developmental change we predict (as assumed by the simplest version of an accumulator model)? Or are there other age-related changes that distinguish younger and older children's learning?

We fit multi-level logistic regression IRT models (De Boeck et al., 2011) to parent report data about the words that children produce. Data were from 5,429 monolingual, English-learning children (16-30 months) from Wordbank (Frank et al., 2021), a repository of data from the MacArthur-Bates Communicative Development Inventory (CDI), a reliable and valid parent report instrument containing a vocabulary checklist of 680 words.[5]

We estimated parameters in each model for each child's linguistic ability and each word's difficulty. Each word's frequency was the key parameter controlling its

[4] If X_j is a vector of person j's covariate data, then their ability can be modeled as $\theta_j = X_j\beta + \epsilon$, where β is a vector of regression parameters (covariates), and ϵ captures residual individual variation.

[5] Because they are a rich source of data, CDIs offer a good starting point for modeling, but our method is not limited to CDI data and could be applied to other vocabulary measures, including samples of language produced by children.

accumulation, hence, an item-level covariate of difficulty. To characterize individual children's environments, we estimated average word frequencies using the American English corpora in CHILDES (MacWhinney, 2000) retrieved from childes-db (Sanchez et al., 2019), following previous work (Braginsky et al., 2019; Goodman et al., 2008). Our average child receives 1,200 words/hour, 12 hours/day, for a total of 438,300 tokens/month, of which 285,200 tokens (65.1%) are words on the CDI (see Appendix).

In our simple accumulator model, the expected total tokens per word given a child's age were included as a covariate. We also included an interaction between frequency and lexical category (Braginsky et al., 2019; Goodman et al., 2008). A second model included age as a person-level covariate, separate item-level covariates of word frequency and lexical class, and interactions of the latter two.

This second model was preferred by multiple model selection metrics, suggesting support for age-related changes in the accumulation mechanism. Predicted acquisition curves by age, lexical class, and prevalence (Figure 2, left) show the expected results that nouns are learned more rapidly than verbs, and that frequent nouns are learned earlier than less common ones, with little effect of frequency for verbs and function words. Moreover, the model predicts item-level acquisition curves (Figure 2, right): e.g., "ball" is learned earlier than "dog" and "go" is easier than "have." The model also generates per-child CDI learning curves (see Appendix), which could for example be used to generate predicted acquisition curves for children receiving much more (or less) monthly input. By combining the CHILDES frequencies of non-CDI words with the lexical class parameters, one could also predict the total size of individual children's vocabulary. More generally, this simulation showcases the use of large-scale datasets to test hypotheses about the nature of learning mechanisms, and provides evidence of age-linked changes in word learning process.

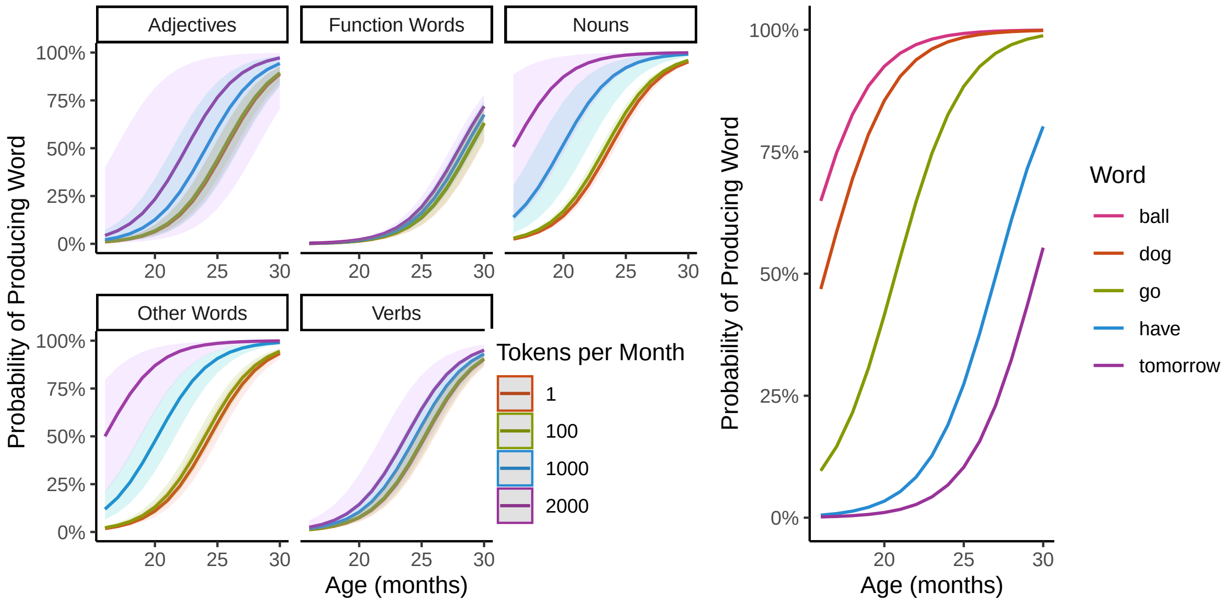

Figure 2. Predicted acquisition curves derived from a fitted age-dependent accumulator model (see text) as a function of lexical class and number of expected tokens per month (left), and for a sample of specific words (right). Nouns and Other Words show strong benefits of higher frequency, while Verbs and Function Words show less effect of frequency.

Onward Towards a Standard Model

The goal of any computational theory is to derive the predictions arising from a specific set of assumptions. Often, however, it is the *failure* of such a model to predict observed patterns of data that is most useful, as these failures point the way forward towards future refinements. We discuss three potential future directions for simple accumulator models.

Leveraged learning and the role of processing. As our results above show, older children accumulate language from their input faster than younger children. Why? Are older children simply better at remembering words, leading to “bigger drops” in their buckets? Or do they *leverage* their knowledge of language to learn faster than younger children (Mitchell & McMurray, 2009)? They could do this by reasoning about new words by exclusion (Markman & Wachtel, 1988). Or could their increasing fluency with the words they know help them learn new words? A surprising proportion of variance in the rate of vocabulary growth is accounted for by the speed with which children process familiar words (e.g., Fernald & Marchman, 2012). These theoretical proposals yield predictions that could be tested using the models we describe.

A theory for understanding acquisition in diverse contexts. The framework above formalizes an implicit assumption: namely, children learn the same way in all circumstances. Yet this assumption could very well be false. For example, some children might learn more from overheard speech and others might learn more from child-directed speech (Sperry et al., 2019), or children may learn more from words spoken in some types of activity contexts than others (e.g., book reading vs. play). The kind of data necessary to test this assumption directly are only now being collected, for example, in studies that rigorously track learning outcomes and amount of language input in diverse populations (e.g., comparisons between children in low-income, rural, indigenous communities and those in higher-income Western contexts; Casillas, Brown, & Levinson, 2020). To accurately estimate the relative value of a token of overheard vs. child-directed speech, or of book reading vs. free play, it will be necessary to have accurate measurement (i.e., long-term, given the variability in children's experience) of 1) how much children experience each context (say, monthly hours), and 2) the amount of speech of each type is heard in each context (we elaborate on this extension in the Appendix).

Beyond vocabulary. We have described a model of the accumulation of words. Yet language is a rich, complex system in which words are inflected morphologically and composed syntactically to express compositional meanings. In early generative views, syntax and morphology were conceptualized as distinct and unconnected from the lexicon, but this conception has not been borne out empirically. Instead, evidence shows again and again that the language system is "tightly woven," with extremely tight correlations between the acquisition of words, morphology, and syntax (Frank et al., 2021).

Theoretically, accumulator models are generic models of skill acquisition, and could thus model more than word learning. If language learning is a form of skill acquisition (Chater & Christiansen, 2018), such connections could lead the way towards extensions of the standard model to the accumulation of broader units of language-like constructions.

Conclusions

An implicit theory drives research and policy-making on early language acquisition: early language accumulates through discrete experiences with individual words. The more experiences, the faster the words are learned. This implicit theory can be expressed within a family of computational models that make quantitative predictions for individual children based on environmental sources of variation, advancing theory (see Haines et al., 2020). By situating these models in a common psychometric framework, we show how they can be used to connect to large-scale datasets. This modeling framework synthesizes measures of language input and vocabulary growth, allowing us to formalize, test, and iteratively improve our understanding. Moreover, this formalization identified a specific gap in current empirical approaches: Studies must report absolute units (e.g., words per hour rather than standardized scores) to allow integrated modeling across studies. Perhaps one day soon, these developments together will lead to a true "standard model" of language learning.

Recommended Readings

- Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). (See References). A large-scale analysis of multiple language input corpora of the type that will be needed to help constrain the standard model, along with discussion of demographic variation in input.
- Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). (See References). A book examining the influence of many child- and word-level factors on

early language learning, along with an introduction to the use of psychometric models on CDI data.

- McMurray, B. (2007). (See References). A short introduction to the idea of accumulator models, and a demonstration of how they can inform theories of language learning.
- Swingley, D. and Humphrey, C. (2018). (See References). An example of how statistical models can be used to infer the relative importance of different language input factors in predicting children's learning of individual words.

References

Bergelson, E. (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives*, *14*(3), 142–149.

Bergelson, E., Casillas, M., Soderstrom, M., Seidl, A., Warlaumont, A. S., & Amatuni, A. (2019). What do north american babies hear? A large-scale cross-corpus analysis. *Developmental Science*, *22*(1), e12724.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, (3), 52–67. https://doi.org/10.1162/opmi_a_00026

Cartmill, E. A., Armstrong, B. F., Gleitman, L. R., Goldin-Meadow, S., Medina, T. N., & Trueswell, J. C. (2013). Quality of early parent input predicts child vocabulary 3 years later. *Proceedings of the National Academy of Sciences*, *110*(28), 11278–11283.

Casillas, M., Brown, P., & Levinson, S. C. (2020). Early language experience in a tseltal mayan village. *Child Development*, *91*(5), 1819–1835.

Chater, N., & Christiansen, M. H. (2018). Language acquisition as skill learning. *Current Opinion in Behavioral Sciences*, *21*, 205–208.

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., … others. (2011). The estimation of item response models with the lmer function from the lme4 package in r. *Journal of Statistical Software*, *39*(12), 1–28.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory.* Psychology Press.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project.* MIT Press.

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.

Haines, N., Kvam, P. D., Irving, L., Smith, C., Beauchaine, T. P., Pitt, M. A., & Turner, B. (2020). Theoretically informed generative models can advance the psychological and brain sciences: Lessons from the reliability paradox. *https://psyarxiv.com/xr7y3/*.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children*. Baltimore, MD: Paul H Brookes Publishing.

Hayiou-Thomas, M. E., Dale, P. S., & Plomin, R. (2014). Language impairment from 4 to 12 years: Prediction and etiology. *Journal of Speech, Language, and Hearing Research*, *57*(3), 850–864.

Hidaka, S. (2013). A computational model associating learning process, word attributes, and age of acquisition. *PLoS ONE*, *8*(11), e76242. https://doi.org/10.1371/journal.pone.0076242

Hoff, E. (2003). The specificity of environmental influence: Socioeconomic status affects early vocabulary development via maternal speech. *Child Development*, *74*(5), 1368–1378.

Jones, G., & Rowland, C. F. (2017). Diversity not quantity in caregiver speech: Using computational modeling to isolate the effects of the quantity and the diversity of the input on vocabulary growth. *Cognitive Psychology*, *98*, 1–21.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. Transcription format and programs* (Vol. 1). Psychology Press.

Marchman, V. A., Adams, K. A., Loi, E. C., Fernald, A., & Feldman, H. M. (2016). Early language processing efficiency predicts later receptive vocabulary outcomes in children born preterm. *Child Neuropsychology*, *22*(6), 649–665.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–157.

McMurray, B. (2007). Defusing the childhood vocabulary explosion. *Science*, *317*(5838), 631.

Mitchell, C., & McMurray, B. (2009). On leveraged learning in lexical acquisition and its relationship to acceleration. *Cognitive Science*, *33*(8), 1503–1523.

Mollica, F., & Piantadosi, S. T. (2017). How data drives early word learning: A cross-linguistic waiting time analysis. *Open Mind*. https://doi.org/10.1162/opmi_a_00006

Muthukrishna, M., & Henrich, J. (2019). A problem in theory. *Nature Human Behaviour*, *3*(3), 221–229.

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.

Sanchez, A., Meylan, S. C., Braginsky, M., MacDonald, K. E., Yurovsky, D., & Frank, M. C. (2019). Childes-db: A flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, *51*(4), 1928–1941.

Sperry, D. E., Sperry, L. L., & Miller, P. J. (2019). Reexamining the verbal environments of children from different socioeconomic backgrounds. *Child Development*, *90*(4), 1303–1318.

Suskind, D. L., Graf, E., Leffel, K. R., Hernandez, M. W., Suskind, E., Webber, R., … Nevins, M. E. (2016). Project aspire: Spoken language intervention curriculum for parents of low-socioeconomic status and their deaf and hard-of-hearing children. *Otology & Neurotology*, *37*(2), e110–e117.

Swingley, D., & Humphrey, C. (2018). Quantitative linguistic predictors of infants' learning of specific english words. *Child Development*, *89*(4), 1247–1267.

Tauber, S., Navarro, D. J., Perfors, A., & Steyvers, M. (2017). Bayesian models of cognition revisited: Setting optimality aside and letting data drive psychological theory. *Psychological Review*, *124*(4), 410.

Wang, Y., Williams, R., Dilley, L., & Houston, D. M. (2020). A meta-analysis of the predictability of lena automated measures for child language development. *Developmental Review*, *57*, 100921.

Appendix to 'Toward a "Standard Model" of Early Language Learning'

George Kachergis, Virginia A. Marchman, and Michael C. Frank

August 18, 2021

In the standard model framework we would like to keep real, measurable units on all parameters, so it is important to use empirical values where possible in order to ground the fitted parameters in interpretable units. In the following, we report how we estimate 1) children's average rate of experienced input (tokens of child-directed speech), and 2) the expected number of tokens per word heard by children in a typical month. We then provide details of the models reported in the paper, as well as a few simpler baseline models. All of the analysis scripts, data sources, and fitted models are available on OSF.[1]

Finally, we offer a schematic of how the standard model could be extended to model the differential contributions of child-directed speech and adult-directed speech to children's early word learning, were a suitable dataset available.

Children's Experienced Rate of Input

To estimate the average hourly input rate of child-directed speech experienced by children, we use the distribution of rates measured from 113 individual children from Hart and Risley (1995), Sperry et al. (2018), and Weisleder and Fernald (2013). The rates from the children in Hart & Risley (1995) were copied from the reported tables, and the remaining datasets were shared with us by the authors, for which we are grateful. Figure 1 plots the distribution of hourly rate of child-directed speech tokens from the primary caregiver (left, $M = 1198$ tokens/hour), as well as the higher rate overheard speech from all adults (right; $M = 1798$), which was only available from 83 children (including 12 additional children from Soderstrom and Wittebolle (2013) for whom caregiver child-directed tokens were not available). Although ultimately we would like to estimate the individual contributions of both child-directed speech and overheard speech from all adults to children's word learning in future iterations of the standard model (see below section on Standard Model 2.0), for now we use the average rate of primary caregiver's child-directed speech, $M \approx 1200$ tokens/hour, since child-directed speech is typically a better predictor of vocabulary growth. Assuming that the average child experiences 1,200 tokens per hour of child-directed speech, 12 hours per day, and 30.44 days per month, they will hear on average 438,336 tokens each month. We can use this average estimate of input in two ways: 1) in combination with each child's fitted latent ability to estimate their amount of experienced input, and 2) in combination with corpus-estimated word frequencies to estimate the number of monthly tokens per word that are experienced by the average child, as we explain in the next section.

Word Frequencies from CHILDES

To estimate how often children hear particular words, we used a selection of child-directed speech from CHILDES (MacWhinney, 2000). We retrieved transcripts of American English utterances using the childes-db R package (Sanchez et al., 2018) and tabulated 7.680345×10^6 tokens of 45598 word types. The majority of transcripts were from children between two to four years of age, and were collected during hour-long

[1] https://osf.io/c5u49/?view_only=a483598c3e7940df939b4c4ab5b4fa04

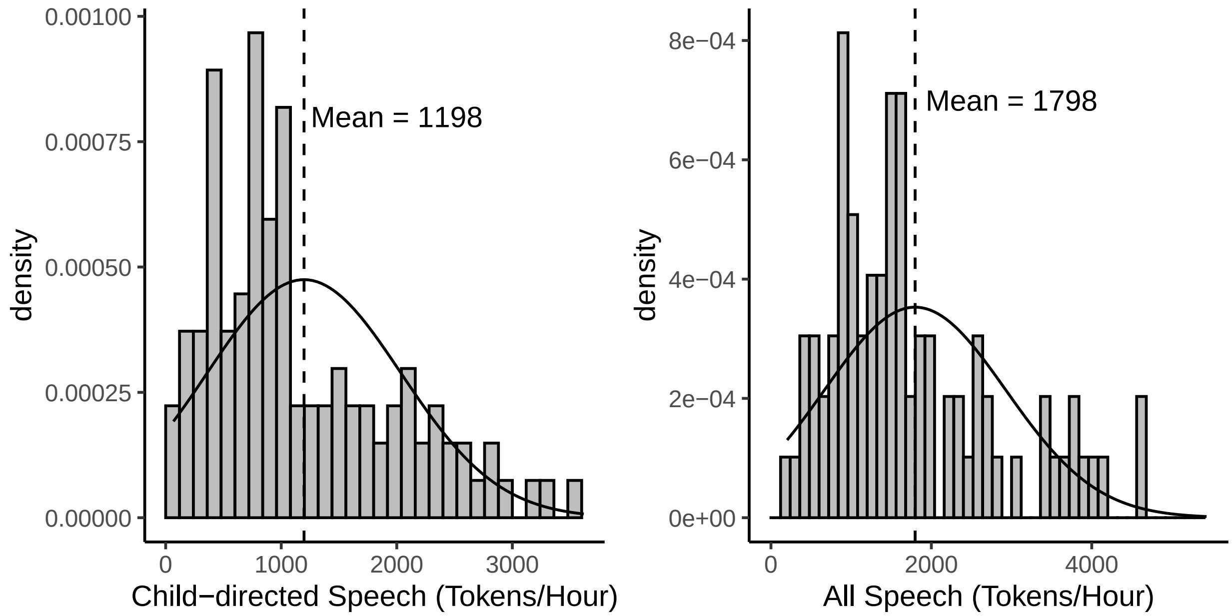

Figure 1: Distribution of hourly child-directed speech rates heard by individual children from their primary caregivers (left) and from all adults (right). On average, two out of three words children hear are child-directed – but individual children's average ratios (and overall rates) likely differ.

at-home recording sessions. We further processed the word counts, combining counts for some variants (e.g., "Mom", "mama", "momma"), removing non-words (e.g., "www") and removing any *hapax legomena* (words appearing only once; e.g. "'s", "A_a"; see file get_childes_freqs.R for details). After this processing, the lexicon contained 27673 unique words that appeared a total of 7.646081×10^6 times in the transcripts, with the median word appearing 6 times. We mapped 659 of these words to words on the CDI Words & Sentences (CDI:WS) form (see preprocess_word_freq.R for details). The 21 CDI:WS words that did not appear in the CHILDES lexicon were assigned the minimum value in the corpus (count = 2). The word counts were converted to probabilities of occurrence: e.g., the most frequent word "you" occurred 340,268 times in the 7.646081×10^6-word corpus, and thus for a randomly-selected word w from the corpus, the probability that w is "you" was $P(w) = 0.0445$. In the models we further convert the $P(w)$ for each word to an expected number of tokens per month based on the above estimated average of 1200 hourly child-directed tokens experienced by children (438,336 child-directed tokens per month). So for example, the average child can expect to hear "you" $0.0445 \times 438336 = 19507$ times in child-directed speech in a typical month, but will only hear "ball" an average of 331 times, and "dog" on average 232 times. Figure 2 shows the overall distribution of expected monthly tokens for the 680 CDI items by lexical class, estimated from child-directed speech in CHILDES. The average child can expect to hear the median words ("doll" and "lion") 58 times each month by the average child, although the frequency distribution is of course right-skewed (mean = 419).

It should be noted that these rates are sure to vary greatly both within children (month-to-month), as well as across children (e.g., perhaps especially if they have a dog). The large amount of expected per-word variation is why, if we want to explain per-child variation in word learning, it is critical to measure not only children's overall input rates, but also the rates of each individual word in the input.

Although we have focused thus far on both the number of tokens of and prevalence of words in child-directed speech, future versions of the Standard Model should take into account two additional sources of input variations. First, children also overhear speech that is not directed at them (other-directed speech; ODS): see the distribution of all daily tokens heard by children at right in Figure 1. Second, the frequency with which particular words appear in child-directed vs. other-directed may vary significantly (e.g., we presumably talk less with our children about 'taxes', and more with them about 'toys'.) See below discussion of Standard

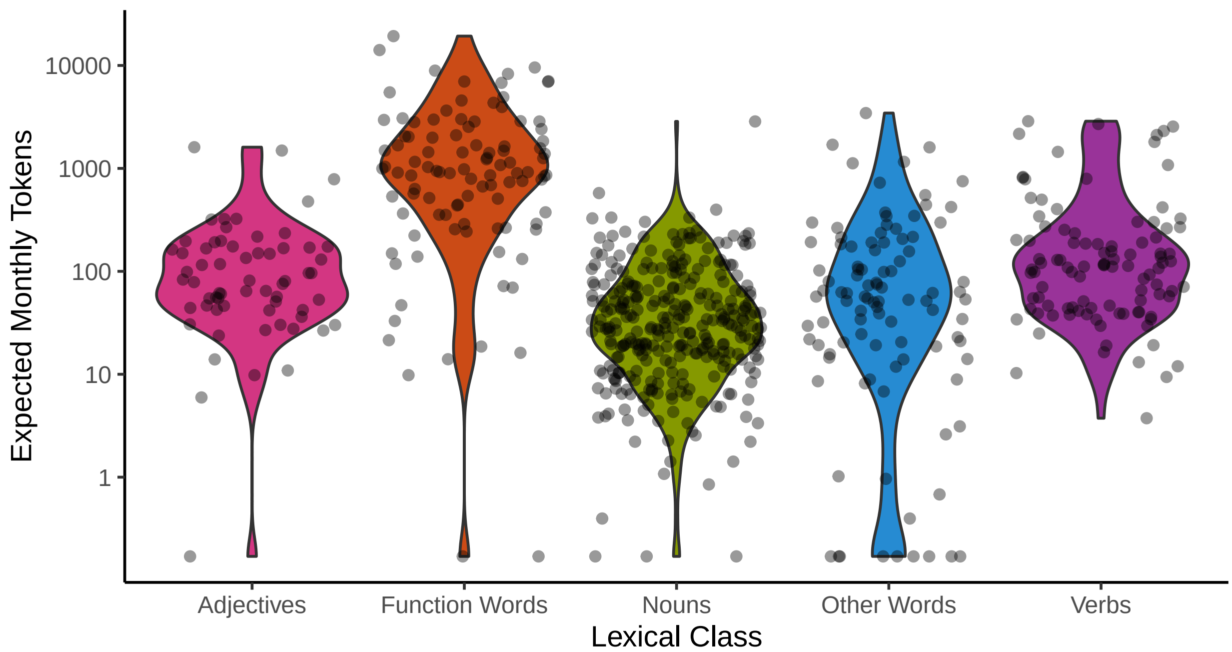

Figure 2: Distribution of expected monthly tokens for the 680 CDI items by lexical class, based on CHILDES word frequencies.

Model v2.0 for how to estimate contributions of other-directed speech to children's early word learning.

Models

Finally, we fitted models to American English Wordbank production data (Words & Sentences form) from 5,492 children 16-30 months of age. The following models were fitted in R using the lme4 package:

- **m_1pl**: produces ~ -1 + (1 | item) + (1 | person)
- **m_tokmo**: produces ~ -1 + tok_per_mo + (1 | item) + (1 | person)
- **m_lc**: produces ~ -1 + lexical_class + (1 | item) + (1 | person)
- **m2**: produces ~ -1 + lexical_class * lifetime_tokens + (1 | item) + (1 | person)
- **m3**: produces ~ -1 + lexical_class * tok_per_mo + age + (1 | item) + (1 | person)
- **m4**: produces ~ -1 + lexical_class * tok_per_mo * age + (1 | item) + (1 | person)

The first three models are baseline models: m_1pl is the 1-parameter logistic (Rasch) model, estimating a single latent ability parameter per child and a difficulty per item, both as random intercepts. m_tokmo includes the item-level covariate tok_per_mo which represents the CHILDES-estimated monthly child-directed tokens of each word received by the average child (i.e., a child hearing ~438,000 tokens/month). m2 includes lexical class and the expected number of lifetime tokens (tok_per_mo * age) per word received by the average child, and their interaction. m3 includes lexical class, tokens per month, and their interaction, and estimates a separate age effect. Finally, m4 extends m3 to include possible interactions of tok_per_mo and age, and of lexical_class, tok_per_mo, and age.

Table 1: Model comparisons.

	npar	AIC	BIC	logLik	deviance	Chisq	Df	Pr(>Chisq)
m_1pl	2.00	2,253,598.74	2,253,625.00	-1,126,797.37	2,253,594.74	NA	NA	NA
m_tokmo	3.00	2,253,584.33	2,253,623.73	-1,126,789.17	2,253,578.33	16.41	1.00	0.00
m_lc	7.00	2,253,329.63	2,253,421.56	-1,126,657.81	2,253,315.63	262.70	4.00	0.00
m2	12.00	2,252,621.52	2,252,779.12	-1,126,298.76	2,252,597.52	718.11	5.00	0.00
m3	13.00	2,248,868.06	2,249,038.79	-1,124,421.03	2,248,842.06	3,755.47	1.00	0.00
m4	22.00	2,244,676.63	2,244,965.56	-1,122,316.32	2,244,632.63	4,209.42	9.00	0.00

Table 2: Coefficients from m3.

	Estimate	Std. Error	z value	Pr(>|z|)
lexical_classadjectives	-12.18	0.27	-44.48	0.00
lexical_classfunction_words	-13.72	0.24	-58.01	0.00
lexical_classnouns	-11.26	0.17	-67.90	0.00
lexical_classother	-11.60	0.22	-52.07	0.00
lexical_classverbs	-12.01	0.23	-53.12	0.00
tok_per_mo	0.00	0.00	1.01	0.31
age	0.47	0.01	81.67	0.00
lexical_classfunction_words:tok_per_mo	0.00	0.00	-0.73	0.47
lexical_classnouns:tok_per_mo	0.00	0.00	1.21	0.23
lexical_classother:tok_per_mo	0.00	0.00	1.52	0.13
lexical_classverbs:tok_per_mo	0.00	0.00	-0.49	0.63

Model comparison

Model comparisons are shown in Table 1, arranged in order of increasing complexity. The successively more complex models show significant improvement in fit (smaller values of AIC and BIC are preferred). **m_tokmo** offers a small but significant improvement in fit compared to the Rasch model (**m_1pl**). Including lexical class rather than monthly tokens per word further improves fit (**m_lc**), but including both lexical class, expected lifetime tokens (a person-item covariate), and their interaction in **m2** is even better. **m3**, with lexical class, monthly tokens, their interaction, and a separate effect of age fits better than **m2**. Finally, allowing for interactions of lexical class and monthly tokens with age in **m4** achieves superior fit. However, for the sake of brevity and interpretability we focus on the comparison of m2 and m3 below and in the paper.

Effects

Table 2 shows the estimated coefficients from m3. All lexical class coefficients are significant, as well as age, but there is no significant effect of monthly tokens per word nor interactions of it with lexical class.

Table 3 shows the estimated coefficients from m4.

m4 revealed significant main effects of lexical class and age, but not of monthly tokens. Significant two-way interactions of age and lexical class and significant three-way interactions of lexical class, monthly tokens, and age show that words' quantities and class influence children's ability to learn them to a varying degree across ages. A non-significant interaction of age and monthly tokens was trending positive, leaving room for future investigations with more sensitive models and larger, perhaps cross-linguistic datasets to seek evidence for changes in overall language ability across development.

Table 3: Coefficients from m4.

	Estimate	Std. Error	z value	Pr(>|z|)
lexical_classadjectives	-12.70	0.28	-46.17	0.00
lexical_classfunction_words	-14.13	0.24	-59.13	0.00
lexical_classnouns	-11.11	0.17	-66.93	0.00
lexical_classother	-10.40	0.22	-46.63	0.00
lexical_classverbs	-13.25	0.23	-58.33	0.00
tok_per_mo	0.00	0.00	0.74	0.46
age	0.50	0.01	81.94	0.00
lexical_classfunction_words:tok_per_mo	0.00	0.00	-0.37	0.71
lexical_classnouns:tok_per_mo	0.00	0.00	0.69	0.49
lexical_classother:tok_per_mo	0.00	0.00	0.90	0.37
lexical_classverbs:tok_per_mo	0.00	0.00	-0.29	0.77
lexical_classfunction_words:age	-0.01	0.00	-2.19	0.03
lexical_classnouns:age	-0.03	0.00	-15.27	0.00
lexical_classother:age	-0.07	0.00	-34.45	0.00
lexical_classverbs:age	0.03	0.00	13.72	0.00
tok_per_mo:age	0.00	0.00	1.85	0.06
lexical_classfunction_words:tok_per_mo:age	0.00	0.00	-2.38	0.02
lexical_classnouns:tok_per_mo:age	0.00	0.00	2.79	0.01
lexical_classother:tok_per_mo:age	0.00	0.00	3.76	0.00
lexical_classverbs:tok_per_mo:age	0.00	0.00	-1.15	0.25

Plotting Standard Model Predictions

Below we plot selected predictions from the fitted models, with a focus on m3. Figure 3 shows marginal predicted effects of lexical class and word frequency for a typical ability 23-month-old child. The overall distributions of CDI words' CHILDES frequencies are shown at left in Figure 2.

Plot Sample of Children and Words

Figure 4 shows predicted acquisition curves from m3 for a small subset of the 680 words (left) and for a random sample of 10 children from the Wordbank sample (right). The learning curves of both individual words and individual children show much variation – for words, based on their estimated difficulty, and for children, based on their estimated language ability. Note that although each of these children were only tested once, the model predicts an entire trajectory per child.

Standard Model 2.0: Child-directed vs. Overheard Speech

The version of the Standard Model (v1.0) presented above and in the paper uses word frequencies from child-directed speech (from the CHILDES database) and expected hourly rates of child-directed speech (CDS) to estimate the number of monthly tokens heard per word, on average. But this ignores the fact that children also overhear speech between adults, and that although this overheard speech may be less attended or otherwise valuable for learning, it may nonetheless explain a significant portion of variation (see Sperry et al., 2018). Figure 5 sketches how a Standard Model v2.0 could use per-child concurrent measures of child-directed and other-directed speech (ODS) to estimate the relative (per-token) value of each type of speech, and to predict learning of individual words (e.g., from the CDI).[2]

[2]We *would* fit this model, but we do not currently have such a dataset.

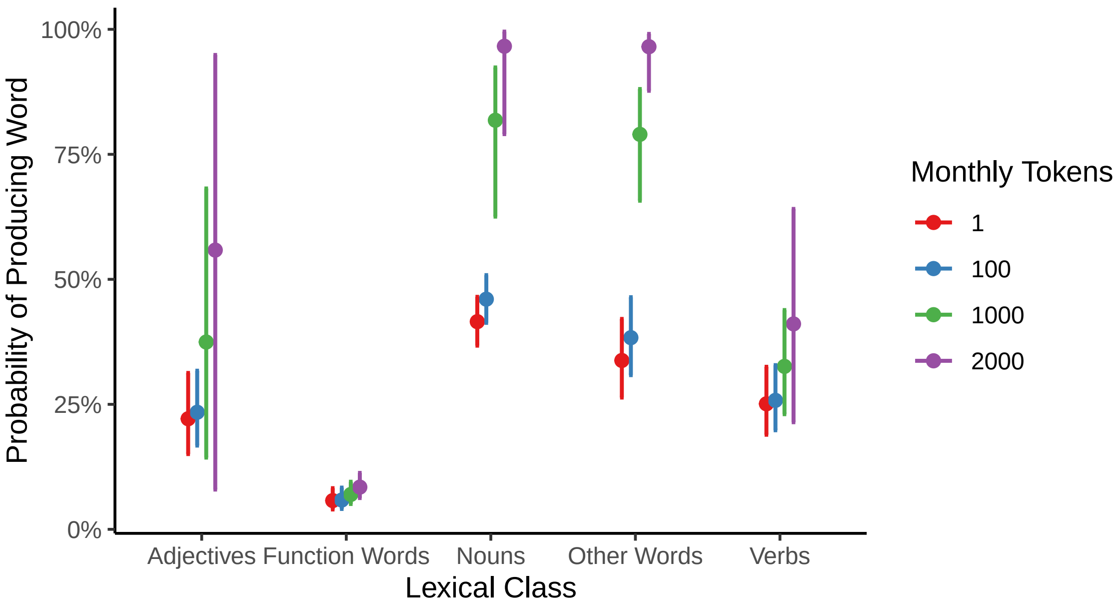

Figure 3: Model m3's predicted probability of the average 23-month-old child producing CDI words of varying frequency and lexical class.

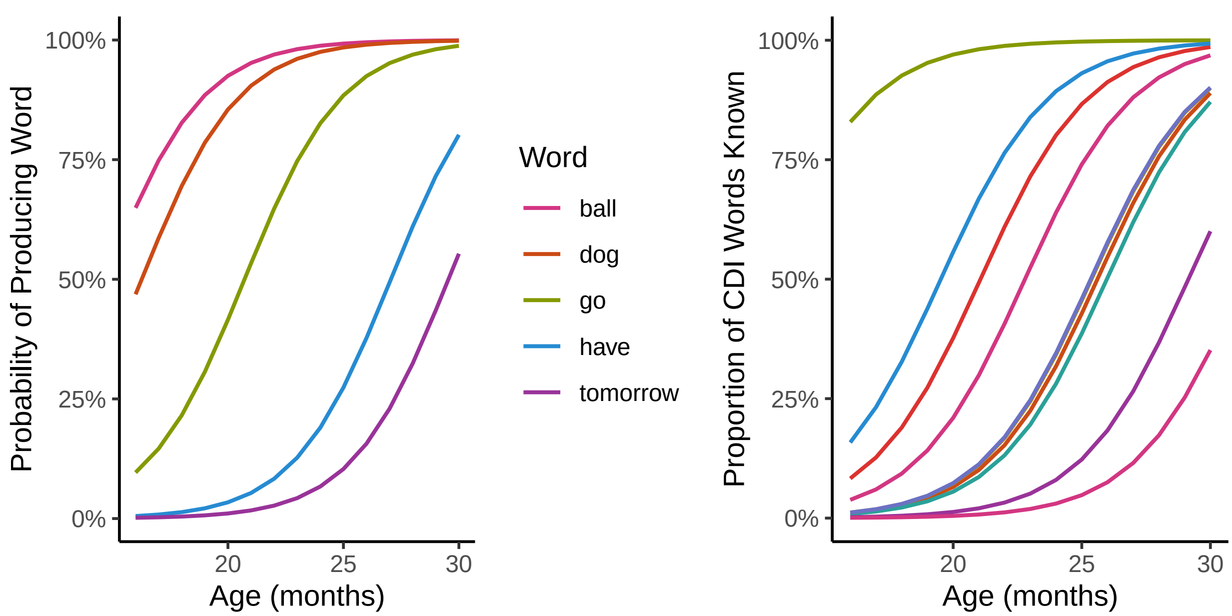

Figure 4: Predicted acquisition curves for a sample of words (left): concrete nouns are learned early while verbs and abstract words (e.g. 'tomorrow') show slower learning. At right are shown predicted learning curves for a sample of 10 children of varying ability.

We show estimated normal distributions for children's hourly rates of input for both CDS ($M = 1200$ words/hour) and ODS ($M = 600$ words/hour). Thus, an average child would hourly receive a total of 1800 hourly tokens, with two thirds of it being child-directed speech. In the figure, we show a hypothetical Child 1 who consistently hears 1.5SD more ODS than average (1830 words/hour), and 1SD less CDS than expected (360 words/hour), for a total of (2190 words/hour). An open question is how much more valuable is 1 token of CDS compared to 1 token of ODS, but this multiplier can be estimated if by leveraging input (CDS and ODS) and acquisition data from the same children.

However, it is also important to note that word frequencies sometimes greatly vary between child-directed speech and adult speech, the latter of which we measure using the frequency distribution of 51 million words in subtitles from American English movies (SUBTLEX; Brysbaert & New, 2009). For example, while "you" appears in CDS and ODS at roughly the same rate (44502 tokens/million words in CHILDES; 41857 tokens/million words in SUBTLEX), "ball" is far more frequent in CDS (6631 vs. 105 tokens/million words), and "tomorrow" is more frequent in ODS than CDS (336 vs. 240). Experienced CDS and ODS tokens both contribute to learning each given word (see pie charts at right), but may contribute differentially – not only due to varying word frequency and a child's varying mean rates of CDS and ODS, but due to the relative value of 1 token of CDS vs. 1 token of ODS, which is an estimated parameter of the model.

Extending the Standard Model framework to model bilinguals would schematically look similar (substitute "L1" for CDS, and "L2" for ODS) – although researchers may then want to measure both CDS and ODS in both languages.

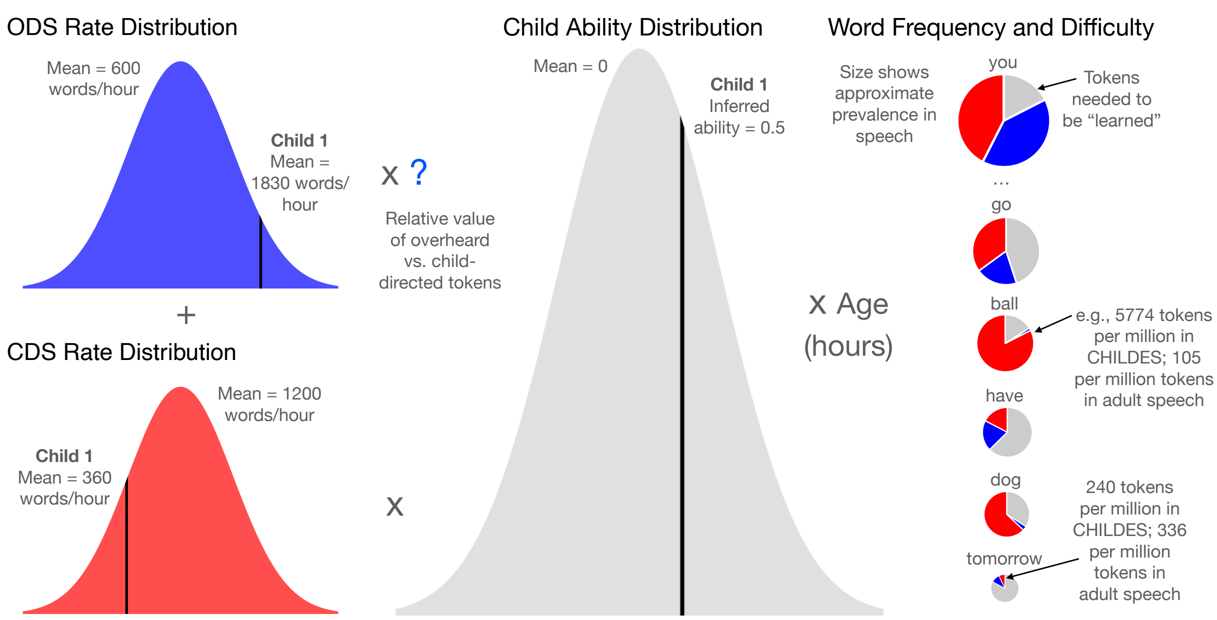

Figure 5: A schematic of how Standard Model v2.0 could incorporate measures of child-directed speech (CDS; red) and other-directed speech (ODS; blue) to also estimate the relative value of the two types of speech. Such a model would also incorporate separate word frequency covariates for CDS and ODS, as word frequencies can vary greatly in CDS vs. ODS (see pie charts at right).