Predicting Age of Acquisition in Early Word Learning Using Recurrent Neural Networks

Eva Portelance (portelan@stanford.edu)
Department of Linguistics, Stanford University

Judith Degen (jdegen@stanford.edu)
Department of Linguistics, Stanford University

Michael C. Frank (mcfrank@stanford.edu)
Department of Psychology, Stanford University

Abstract

Vocabulary growth and syntactic development are known to be highly correlated in early child language. What determines when words are acquired and how can this help us understand what drives early language development? We train an LSTM language model, known to detect syntactic regularities that are relevant for predicting the difficulty of words, on child-directed speech. We use the average surprisal of words for the model, which encodes sequential predictability, as a predictor for the age of acquisition of words in early child language. We compare this predictor to word frequency and others and find that average surprisal is a good predictor for the age of acquisition of function words and predicates beyond frequency, but not for nouns. Our approach provides insight into what makes a good model of early word learning, especially for words whose meanings rely heavily on linguistic context.

Keywords: Language model; recurrent neural network; LSTM; language acquisition; age of acquisition; child directed speech; word learning.

Introduction

Children's lexicon and syntactic abilities grow in tandem, resulting in a tight correlation between vocabulary size and grammatical complexity (Bates et al., 1994; Brinchmann, Braeken, & Lyster, 2019; Frank, Braginsky, Marchman, & Yurovsky, 2019). Additionally, children are remarkably consistent in the order in which the acquire their first words (Tardif et al., 2008; Goodman, Dale, & Li, 2008). This ordering consistency presents an opportunity: modeling of when words are acquired can help us understand what drives language learning more generally.

This general approach relies on creating quantitative models of children's average age of acquisition (AoA) for words. These analyses typically combine large-scale survey data about word acquisition with corpus estimates of language input from different children to make aggregate-level predictions. Such studies have investigated the effects of word properties such as frequency, number of phonemes, and concreteness, across a range of languages (Goodman et al., 2008; Kuperman, Stadthagen-Gonzalez, & Brysbaert, 2012; Braginsky, Yurovsky, Marchman, & Frank, 2019), with simple frequency typically being the most important factor: the more frequent a word is in the child's linguistic input, the earlier it is acquired.

These analyses have generally not considered the linguistic contexts in which words appear, however. Some analyses have considered the length of utterances as a weak proxy for sentence complexity (Braginsky et al., 2019). Some have also considered contextual diversity – a measure of semantic co-occurrence – as a possible predictor of AoA beyond frequency (Hills, Maouene, Riordan, & Smith, 2010). This could be considered a proxy for some semantic factors but does not directly measure syntactic complexity. Given the strong connection between vocabulary growth and syntactic ability in children's output, we hypothesize that the syntactic contexts in which words appear in children's language input should be an important factor as well. To test this hypothesis, we use a computational language model in concert with a broad array of child-language data sets to predict when words are acquired, exploring contextual linguistic information beyond first-order word frequency.

We use a long-short term memory (LSTM) neural model (Hochreiter & Schmidhuber, 1997), trained on child-directed speech, as our model of contextual information in sequences and as a proxy for syntactic complexity. LSTMs are a form of recurrent neural network (RNN) that can be used as language models (Sundermeyer, Schlüter, & Ney, 2012). Language models are trained on corpora to predict the next word in a sequence. Though more recent language models outperform LSTMs (Vaswani et al., 2017), they are a strong standardized baseline and have many useful analytic properties. They process utterances incrementally and make use of nested layers of hidden units to learn abstract representations that can predict sequential dependencies between words across a range of dependency lengths (Linzen, Dupoux, & Goldberg, 2016). Unlike regular RNNs, LSTM units use a gating system that allows them to 'forget' some of the previous states while 'remembering' others, thus learning to prioritize some dependencies in a sequence over others at each state. Further, regular RNNs have previously been proposed as cognitive models for language learning (Elman, 1990, 1993; Christiansen, Allen, & Seidenberg, 1998), however, these earlier models were computationally limited and could be used only with small, schematic datasets; in contrast, LSTMs can be applied to larger datasets. Thus, LSTM language models lend themselves well to our project.

As our linking hypothesis between the LSTM model and children's difficulty, we use the model's average surprisal: the negative log probability of a word w_i in a given context $w_{1...i-1}$, averaged across all contexts in which it appears, C.

$$\sum_{C:w_i \in C} -\log P\left(w_i \mid w_{1\ldots i-1}\right) \times \frac{1}{|C|}$$

Since the LSTM's learning objective is to minimize the surprisal of words in context, the average surprisal of a word constitutes a measure of how difficult that word is for the model to represent. In addition to being an appropriate measure of difficulty for the LSTM, surprisal has also been shown to be a strong predictor of human processing difficulty in psycholinguistic experiments (Levy, 2008; Demberg & Keller, 2008).[1]

Our approach is as follows. First we describe the data we use for training, then we describe the LSTM model architecture and training[2]. We then compare regression models of children's AoA using average surprisal as one key predictor in concert with previous predictor sets. An important advance over previous work is the use of cross-validation to estimate out-of-sample performance. In sum, our approach allows us to determine the predictive power of this new measure and to investigate how well sequential prediction difficulty in an LSTM relates to patterns of acquisition for children.

Corpus Data

To train the model, we compiled a dataset of child-directed speech from the CHILDES database (MacWhinney, 2000). We included all of the child-directed utterances from the 39 largest single-child English corpora available through the childes-db API (Sanchez et al., 2018). To ensure a reasonable sample of child-directed speech from each contributing corpus, we selected corpora that contained at least 20,000 tokens and a ratio of child to child-directed utterances of at least 1:20. The corpora come from 18 distinct studies and were all longitudinal, with recording transcripts typically coming from regular hour-long recording sessions at the child's home. The age of children in the transcripts ranged from 9 months to $\sim$ 5 years of age, with a mean of 32 months (2;9 years). Most of the data were from children between the ages of 2 to 4 years (see Figure 1). The resulting dataset of child-directed utterances from these corpora contains over 6.5 million words. Utterances were transformed to lowercase, and all punctuation except for end of sentence tokens was removed before being passed to the model.

Language Model

We use a two-layered LSTM recurrent neural network, illustrated in Figure 2. The model has randomly initialized 100-dimensional word embeddings as its input layer, which are updated during learning. Hidden states encode information about the preceding context. At each time-step, the current word embedding w_t and the hidden state from the previous time-step h^1_{t-1} are passed through a transformation function, resulting in a new hidden state h^1_t. This hidden state h^1_t and the hidden state from the previous time-step in the second layer h^2_{t-1} are then also passed through a transformation function, resulting in a new hidden state h^2_t. This final hidden state then undergoes a transformation to produce the output layer – a distribution over the whole vocabulary representing a prediction about the upcoming word. In order to limit the number of parameters in the model, we limited vocabulary size to the 10,000 most frequent words.

The model's learning objective is to minimize the surprisal (negative log probability) of words given their preceding context in an utterance. (The average surprisal of a word across all the utterances in which it appears is therefore a measure of how difficult it is for the model to converge on a representation for the word, an observation we leverage in our simulations below). We trained the model on 80% of all of the child-directed utterances (and no child-produced utterances). The remaining 20% were used as a validation set to evaluate the model's performance during training (to avoid overfitting). Utterances were shuffled at each epoch of training. The

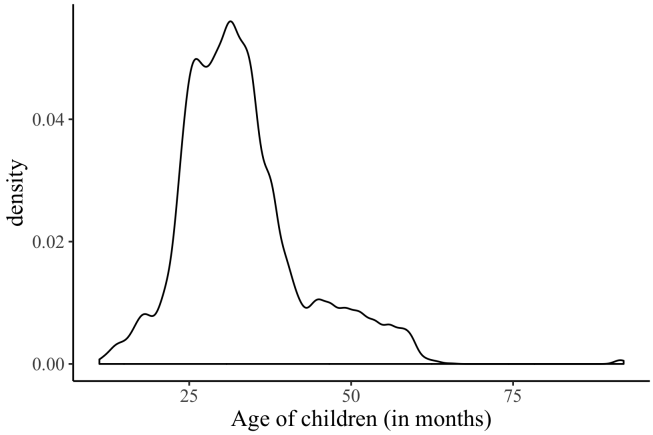

Figure 1: The overall distribution of children's age at the time of child-directed utterance production across all of the data.

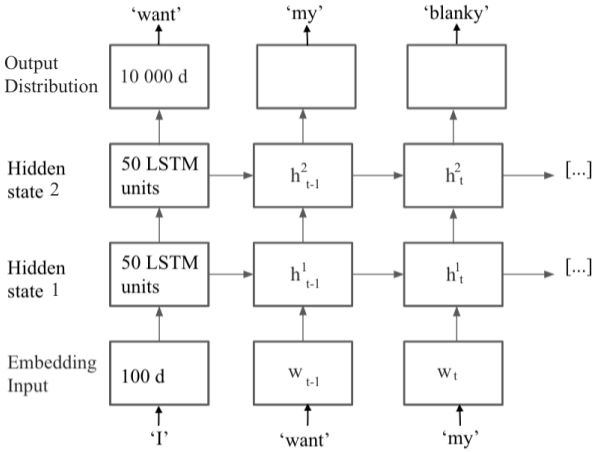

Figure 2: The LSTM model architecture incrementally processing the utterance 'I want my Blanky'.

[1]For an overview of empirical evidence supporting the validity of surprisal as a predictor of processing difficulty, see Hale, 2016.

[2]All code and data for this paper are available at www.github.com/eporte2/aoa_lstm_prediction.

model saw 32 sentences per epoch before performing backpropagation, after which we calculated the categorical cross entropy loss of the model on 32 unseen utterances from the validation set. If the loss on the validation set did not improve for more than 10 epochs in a row, we terminated model training.

We performed cross-validation tests to find the parameter settings for the LSTM language models that best minimized surprisal. The parameters we tested were the number of LSTM layers (1 or 2); the size of the hidden dimension for the hidden state (5, 10, 25, 50); the size of the output word vectors (10, 50, 100); the regularization method to be used (L1, L2, and elastic net). We found that the optimal parameter combination was to have 2 hidden LSTM layers with a hidden dimension of 50 units, 100-dimensional word vectors, and elastic net regularization.

Prediction Task

Our goal was to understand whether sequential predictability – surprisal in the trained LSTM model – is related to difficulty of acquisition for children. In other words, are words that are difficult for the model to predict also difficult for children to learn? In this section, we describe the data and experimental procedures used to explore this question.

Data

Following Braginsky et al., 2019, we estimated AoA using the MacArthur-Bates Communicative Development Inventories (CDI) (Fenson et al., 1993)[3]. The CDI is an instrument for parents to report their child's vocabulary – essentially, a checklist of words. We accessed American English CDI data through Wordbank, a cross-linguistic repository for CDI data (Frank, Braginsky, Yurovsky, & Marchman, 2016). We used the "Words & Sentences" version of the CDI, which asks parents to report on which words their child produces and is recommended for use with children ages 16 – 30 months.

To select the list of words for AoA prediction, we began with the 375 word list used by Braginsky et al., 2019 so that we could make use of that study's other predictors. We then discarded any item which contained multiple words (e.g. *peanut butter*, *choo choo*), since the LSTM would have considered these as separate items. We also discarded any item for which the lexical category was either unavailable or marked as "other" since we were interested in the interaction between the lexical category of a word and its average surprisal or frequency. This left a total of 314 words for which we could predict AoA (189 nouns, 91 predicates, and 34 function words).

[3]A reviewer asked why we chose to use these AoA estimates over those of Kuperman et al., 2012. Though Kuperman et al., 2012 have estimates for a much larger vocabulary, they are based on adult estimates of their own AoA, rather than timely reports of children's AoA. Thus, we favored using AoA estimates collected from CDI instruments.

Table 1: Variance inflation factor (VIF) for all predictors.

Predictor	VIF
`average surprisal`	1.40
`frequency`	1.42
`number of phonemes`	1.19
`concreteness`	3.54
`valence`	1.05
`arousal`	1.17
`babiness`	1.12
`lexical category`	5.63

Age of acquisition estimates

Our predictive target is the age at which a word is acquired. We assume that AoA correlates with ease of acquisition. Since not all children learn a given word at the same time, we instead quantified AoA as the age at which 50% of children are reported to produce a word (Goodman et al., 2008).

There are a number of methods to estimate the 50% point. The simplest method is to determine the youngest age group at which the empirical proportion of children producing the word is $> 50\%$, but this approach has several shortcomings. If words are very hard or very easy to learn, then it is possible that for the covered age range some words never reach the 50% point (e.g., *beside*), or have already surpassed the 50% point (e.g., *Mommy*). Such words would have to be discarded if we were to use this method. Another issue is that this approach is susceptible to bias AoA estimates towards ages for which more CDI instruments were available since the number of observations at each age is not equal (i.e., there may be more CDI instruments filled with 24-month-olds than with 20-month-olds in the dataset).

For these reasons, we base our AoA estimates on the model provided by Frank et al. (2019), a Bayesian generalized linear model with hand-tuned prior parameters that were fitted to the English Words & Sentences CDI instruments from Wordbank (for more detailed description see Appendix E of Frank et al., 2019).

Predictors

We next describe the predictors of AoA that we use in our regression models.

Average surprisal We used the LSTM model described above to compute the average surprisal of each word across all of the utterances in the corpus. To set an upper bound on surprisal, we added a small normalizing constant $\varepsilon \approx 2.22 \times 10^{-16}$ to the predicted probability of each word. This step was necessary because in very unlikely contexts, given the size of the vocabulary, the probability of a word was on occasion so small that it led to infinite surprisal values. Thus, the average surprisal values are capped at an upper bound of $-\log \varepsilon$.

Frequency We calculated the raw frequency counts of words for each of the 39 CHILDES corpora used to train the model and then weighted the counts based on each corpus size. We then averaged the weighted frequencies across all the corpora, excluding any zero counts. We did not aggregate counts across inflected forms (e.g. 'give' and 'gave') since the LSTM language model considered inflected forms separately during learning.

Other predictors We also included predictors used by Braginsky et al. (2019): the number of phonemes in each word, which is a proxy for production difficulty; concreteness, capturing conceptual difficulty; valence and arousal, capturing emotional aspects of the words); and babiness, which captures the association of particular words with babies and acts as a proxy for how likely babies are to be exposed to and to attend to specific items. For a full description of these predictors see Braginsky et al., 2019.

Lexical category interactions All of the previously listed predictors were tested for interactions with lexical category. We considered three lexical categories: nouns (common nouns), predicates (verbs, adjectives, and adverbs), and function words (closed-class words) following Bates et al., 1994. Word categories were derived from the categories on the CDI forms (e.g., verbs are listed as "action words").

Collinearity analysis

We did not observe strong correlations between any of the predictors. We also did not find evidence for multicollinearity. The strongest Pearson correlation coefficient we found was between surprisal and concreteness, $r = 0.41$. Frequency and surprisal had a correlation of $r = -0.35$. All other correlation coefficients were negligible ($\leq$ +/- 0.3). Table 1 shows the variance inflation factor (VIF) for all predictors. VIFs were relatively low, with the exception of concreteness and lexical category: both these predictors are categorical variables with relatively few levels (5 and 3 levels respectively). Thus, these slightly higher VIFs can be safely ignored.

Regression models

To determine if the average surprisal from the LSTM language model increased the accuracy of AoA predictions, we compared linear regression models with different predictor sets using leave-one-out (LOO) cross-validation. We evaluated the mean absolute deviation (MAD) of these models' predictions across all words (each word represents one instance of a LOO model fit). The absolute deviation of a word is the absolute difference in months between the actual AoA estimate and the predicted AoA.

The full model we examined contained all predictors and their interactions with lexical category and was specified as: `AoA ~ lexical category * ( average surprisal + frequency + number of phonemes + concreteness + valence + arousal + babiness )` .

We were interested in the relationship of surprisal and frequency, including whether the two predictors each made unique contributions to prediction performance. Thus, we compared performance of the full model to: a null model with no predictors; models including only frequency or surprisal; and the various combinations of surprisal, frequency, and the other predictors above.

Table 2: Mean absolute deviation (MAD) in months of predicted AoA to actual AoA estimate across all words using LOO cross-validation by model, `AoA ~ lexical category * predictors`. Other predictors are `number of phonemes`, `concreteness`, `valence`, `arousal`, and `babiness`.

Predictors	MAD	95%CI
`null model`	2.35	[2.14, 2.56]
`surprisal`	2.12	[1.91, 2.32]
`frequency`	2.00	[1.82, 2.18]
`surprisal + frequency`	2.00	[1.81, 2.21]
`surprisal + others`	1.97	[1.79, 2.18]
`frequency + others`	1.92	[1.74, 2.09]
`surprisal + frequency + others`	1.88	[1.70, 2.05]

Results and Discussion

The MAD for each model after cross-validation is shown in Table 2. The `null model` represents the null hypothesis where we predict the mean AoA for all words simply from the intercept. This baseline model's MAD is 2.35 months. In other words, on average, the `null model` is off by 2.35 months across all words and LOO cross-validation folds. The best model improves on this baseline by reducing the MAD to 1.88 months.

A model including frequency alone was more accurate on average than a model including surprisal alone. Thus, we did not find that surprisal could subsume frequency. However, the best model was the full model with all predictors, `surprisal + frequency + others`, demonstrating that surprisal added information beyond frequency when predicting held out data during cross-validation. A nested model comparison between this full model and the one without surprisal and its interactions with lexical category as predictors (`frequency + others`) fitted on all of the data revealed that surprisal and its interactions with lexical category explained variance above and beyond frequency ($F_{3,290} = 4.32$, $p < .01$).

Figure 3 illustrates the direction of the estimates for frequency and surprisal in the best model (`surprisal + frequency + others`) for each lexical category from the cross-validation. Frequency has negative estimates across all lexical categories. This means that the more frequent a word is in the input, the lower its AoA, or the easier it is to acquire. In contrast, for surprisal, coefficients were positive for function words and predicates (but not for nouns). Thus, the higher the average surprisal of a predicate or function word — i.e. the less predictable it is across contexts — the later its

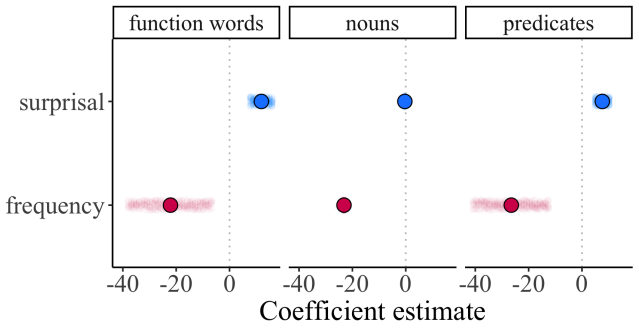

Figure 3: Estimates for frequency and average surprisal for best model (`average surprisal + frequency + others`) by lexical category (main effect of predictor + main effect of lexical category + interaction between predictor and lexical category). Each faded point represents a fold of the leave-one-out cross validation, i.e. one word. The large bordered dots represents the mean of each estimate across folds.

AoA. This relation was not present for nouns, however, suggesting that surprisal did not predict difficulty of acquisition.

When we compared the model with frequency and other predictors to the full model that included surprisal, we found that the words which benefited the most from the addition of the surprisal predictor were function words and predicates. The top 50 words with the largest reduction in absolute deviation are shown in Figure 4. Overall, adding surprisal saw a mean increase in MAD of 0.007 for nouns – that is, the full model did slightly worse at predicting the AoA of nouns – but a mean decrease of MAD of 0.218 for function words and 0.082 for predicates.

Function words and predicates are words that require other dependent words to fully express their meaning (Gleitman, 1990). Thus, it makes sense that the variability in their contexts of use – as captured by model surprisal beyond frequency – should be an important indicator of their learnability. In contrast, nouns in child-directed speech tend to be concrete: The mean concreteness score taken from Brysbaert, Warriner, & Kuperman, 2014, for all nouns in our data on a scale of 1 through 5 (1=abstract, 5=concrete) is 4.86 – almost all the nouns are concrete (for comparison, the mean score for function words is 2.71 and for predicates is 3.50). It is therefore possible that children use other cues in their environment when they hear these words which may contain much richer sources of information beyond the linguistic context in which the word was uttered, such as pointing, gaze, or joint attention to a given object.

General Discussion

A fully-explicit theory of early language should ideally provide strong predictions about the course of acquisition. Towards this general goal, previous work has explored the specific challenge of predicting which words are learned earlier or later (e.g. Goodman et al., 2008; Braginsky et al., 2019).

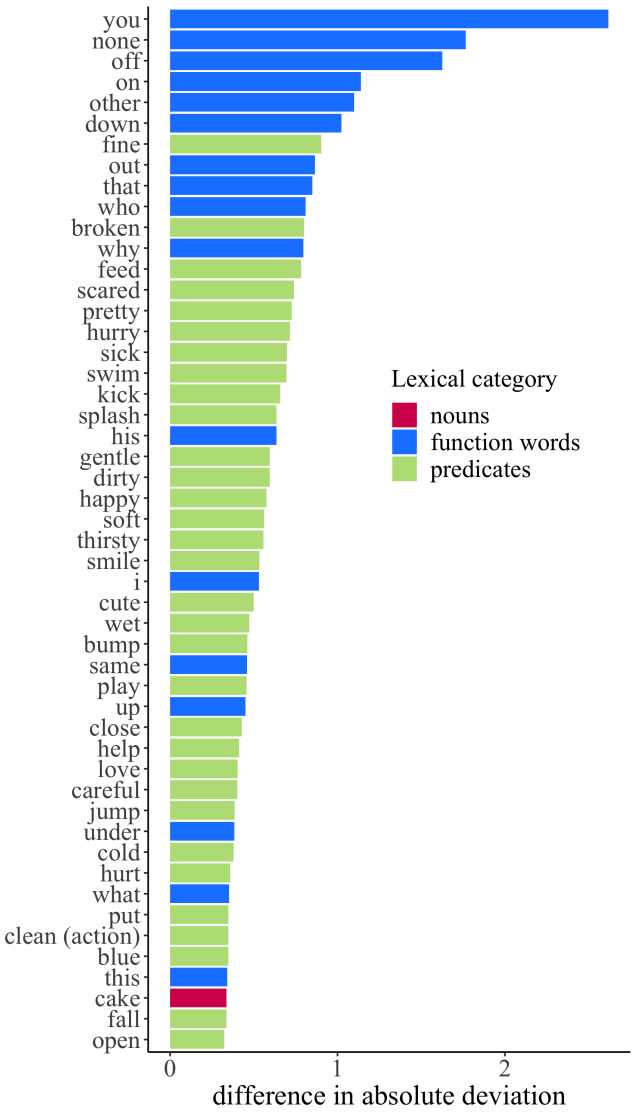

Figure 4: The top 50 words for which adding average surprisal as a predictor improves model fit. The difference in absolute deviation by word (top 50/314 words or 16% of the corpus) between a model with `average surprisal + frequency + other` and `frequency + other`. These words represent 17/34 (50%) of function words, 32/91 (35.16%) of predicates, and only 1/189 (0.53%) of nouns.

Here we focused on understanding the role of sequential predictability in what makes particular words easy or hard to acquire. We used an LSTM – a generic language model – trained on a corpus of child-directed speech, to estimate the average surprisal of specific words. When added to regression models, these surprisal estimates increased predictive accuracy over and above simple frequency and other predictors from prior research. The LSTM was especially useful in predicting children's difficulties on those words for which linguistic context is important to meaning: function words and predicates.

Though the LSTM is a useful tool for understanding sentence predictability, it is only one of many architectures we could have used and is not a proposal for a cognitive model of children's sequential processing. Further, we were limited by the amount of data available in CHILDES in English; LSTM language models are usually trained on larger data sets than the one used here (though more data does not necessarily mean that models learn better representations; van Schijndel, Mueller, & Linzen, 2019). Our corpus was – by necessity – assembled from many sub-corpora from different children, and does not represent a true estimate of the regularity, idiosyncrasy, and contextual diversity found in the language targeted to a single child. Additionally, our corpus contained utterances directed at children who were older in some cases than those surveyed for the AoA estimates. Ideally, these sentences would span the exact same developmental stages. Future work should address these issues.

Theories of language learning must explain not just words like "ball" and "dog" but also words – especially function words – whose meaning in context is almost entirely dependent on other words. Sequential models like the LSTM we used here may be a promising avenue for helping to explain the acquisition of these "hard words". A goal for future work is to understand how sequential (syntactic) information can be combined with other contextual and semantic information, and how the interaction of these factors might lead to the different acquisition trajectories for nouns and function words.

Acknowledgments

Thank you to the members of the Language and Cognition Lab, ALPS Lab, as well as our CogSci 2020 reviewers for their valuable comments! Thank you to all the contributors to both CHILDES and Wordbank for making work like this possible. The first author is supported by the Social Sciences and Humanities Research Council of Canada Doctoral Fellowship and the Goodan Family Fellowship.

References

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P., Reznick, J. S., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123.

Bloom, L., Hood, L., & Lightbown, P. (1974). Imitation in language development: If, when, and why. *Cognitive Psychology*, *6*(3), 380–420.

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and variability in children's word learning across languages. *Open Mind*, 1–16.

Brinchmann, E. I., Braeken, J., & Lyster, S.-A. H. (2019). Is there a direct relation between the development of vocabulary and grammar? *Developmental Science*, *22*(1), e12709.

Brown, R. (1973). *A first language: The early stages*. Harvard University Press.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, *46*(3), 904–911.

Christiansen, M. H., Allen, J., & Seidenberg, M. S. (1998). Learning to segment speech using multiple cues: A connectionist model. *Language and cognitive processes*, *13*(2-3), 221–268.

Clark, E. V. (1978). Awareness of language: Some evidence from what children say and do. In R. J. A. Sinclair & W. Levelt (Eds.), *The child's conception of language* (pp. 17–43). Springer.

Demberg, V., & Keller, F. (2008). Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, *109*(2), 193–210.

Demetras, M. (1989). Changes in parents' conversational responses: A function of grammatical development. *ASHA, St. Louis, MO*.

Demetras, M. J., Post, K. N., & Snow, C. E. (1986). Feedback to first language learners: The role of repetitions and clarification questions. *Journal of child language*, *13*(2), 275–292.

Demetras, M. J.-A. (1989). Working parents' conversational responses to their two-year-old sons.

Demuth, K., & McCullough, E. (2009). The prosodic (re) organization of children's early english articles. *Journal of Child Language*, *36*(1), 173–200.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*(1), 71–99.

Fenson, L., Dale, P., Reznick, J. S., Thal, D., Bates, E., Hartung, J., ... Reilly, J. (1993). MacArthur Communicative Inventories: User's guide and technical manual. *San Diego*.

Frank, M. C., Braginsky, M., Marchman, V., & Yurovsky, D. (2019). *Variability and Consistency in Early Language Learning: The Wordbank Project.* (https://langcog.github.io/wordbank-book/index.html)

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694.

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

Goodman, J. C., Dale, P. S., & Li, P. (2008). Does frequency count? Parental input and the acquisition of vocabulary. *Journal of Child Language*, *35*(3), 515–531.

Hale, J. (2016). Information-theoretical complexity metrics. *Language and Linguistics Compass*, *10*(9), 397–412.

Henry, A. (1995). *Belfast English and Standard English: Dialect variation and parameter setting*. Oxford University Press.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. *Journal of Memory and Language*, *63*(3), 259–273.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Kuczaj II, S. A. (1977). The acquisition of regular and irregular past tense forms. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 589–600.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990.

Levy, R. (2008). Expectation-based syntactic comprehension. *Cognition*, *106*(3), 1126–1177.

Lieven, E., Salomo, D., & Tomasello, M. (2009). Two-year-old children's production of multiword utterances: A usage-based analysis. *Cognitive Linguistics*, *20*(3), 481–507.

Linzen, T., Dupoux, E., & Goldberg, Y. (2016). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *TACL*, *4*, 521–535.

MacWhinney, B. (2000). *The CHILDES Project: Tools for analyzing talk. Third Edition.* Erlbaum.

MacWhinney, B., & Snow, C. (1990). The child language data exchange system: An update. *Journal of Child Language*, *17*(2), 457–472.

Peters, A. M. (1987). The role of imitation in the developing syntax of a blind child. *Text-Interdisciplinary Journal for the Study of Discourse*, *7*(3), 289–311.

Rowland, C. F., & Fletcher, S. L. (2006). The effect of sampling on estimates of lexical specificity and error rates. *Journal of Child Language*, *33*(4), 859–877.

Sachs, J. (1983). Talking about the there and then: The emergence of displaced reference in parent-child discourse. In K. E. Nelson (Ed.), *Children's language* (Vol. 4, pp. 1–28). Erlbaum.

Sanchez, A., Meylan, S., Braginsky, M., MacDonald, K., Yurovsky, D., & Frank, M. C. (2018). childes-db: a flexible and reproducible interface to the Child Language Data Exchange System.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *13th annual conference of the international speech communication association.*

Suppes, P. (1974). The semantics of children's language. *American Psychologist*, *29*(2), 103–114.

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

Theakston, A. L., Lieven, E. V., Pine, J. M., & Rowland, C. F. (2001). The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, *28*(1), 127–152.

van Schijndel, M., Mueller, A., & Linzen, T. (2019). Quantity doesn't buy quality syntax with neural language models. *arXiv:1909.00111*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998–6008).

Weist, R. M., & Zevenbergen, A. A. (2008). Autobiographical memory and past time reference. *Language Learning and Development*, *4*(4), 291–308.