Predicting a Child’s Trajectory of Lexical Acquisition

Nicole Beckage (nicole.beckage@colorado.edu)
Department of Computer Science, University of Colorado
Boulder, CO 80309 USA

Michael Mozer (mozer@colorado.edu)
Department of Computer Science, University of Colorado
Boulder, CO 80309 USA

Eliana Colunga (eliana.colunga@colorado.edu)
Department of Psychology and Neuroscience, University of Colorado
Boulder, CO 80309 USA

Abstract

How does a child’s vocabulary production change and expand over time? Past research has often focused on characterizing population statistics of vocabulary growth. In this work, we develop models that attempt to predict when a specific word will be learned by a particular child. The models are based on two qualitatively different sources of information: a representation describing the child (age, sex, and quantifiers of vocabulary skill) and a representation describing the specific words a child knows. Using longitudinal data from children aged 15-36 months collected at the University of Colorado, we constructed logistic regression models to predict each month whether a word would be learned in the coming month. Models based on either the child representation or the word representation outperform a baseline model that utilizes population acquisition norms. Although the child- and word-representation models perform comparably, an ensemble that averages the predictions of the two separate models obtains significantly higher accuracy, indicating that the two sources of information are complementary. Through the exploration of such models, we gain an understanding of the factors that influence language learning, and this understanding should inform cognitive theories of development. On a practical level, these models support the development of interventions to boost language acquisition.

Keywords: Language acquisition; word learning; lexical acquisition

Introduction

How does a child’s current vocabulary inform and relate to their vocabulary in the future? We know that deficits in a child’s early lexicon is a predictor of future language skills (Dale et al., 2003). Potentially, if researchers can recommend words that the child is ready and able to learn, early learning deficits might be corrected. However, reliable prediction can be made only if word learning develops in a systematic way. In this paper, we explore whether there are regularities in the growth of a child’s vocabulary that allow the trajectory of an individual’s learning to be predicted.

One source of information that can be used to model vocabulary acquisition is population-level norms. The most comprehensive study (Dale & Fenson, 1996) collected productive vocabulary for over 1130 children between the ages of 16 and 30 months, based on parent reports on 649 words. Summary statistics from this *communicative development inventory* or *CDI*, describe norms of acquisition. For example, 78.7% of children produce the word dog by age 18 months. Figure 1 top frame, shows an example of the CDI norms.) These norms are typically used to assess a child’s vocabulary in relation to her peers, as quantified by a CDI percentile for a given age and vocabulary size. However, the CDI population statistics can also be used to predict an individual’s learning of a given word at a given age.

	month 16	month 17	…	month 29
airplane	38.5	39.4	…	95.0
light	35.9	30.3	…	90.0
zoo	9.0	9.1	…	66.7

	age	sex	…	voc. size	dog	house	…	zoo
child A	16.2	F	…	32	0	0	…	0
	17.1	F	…	49	1	0	…	0
	18.9	F	…	132	1	0	…	1
child B	19.3	M	…	257	1	0	…	0
	20.5	M	…	345	1	1	…	0

Figure 1: Example of normed CDI entries (top) and longitudinal CDI data for sample children (bottom).

The accuracy of these predictions for any individual depends on the nature of variability within the population. Any model based on normed data assumes that children learn in a fundamentally similar fashion to one another. For example, implicit in a model based on normed data is that late talkers (children at or below the 20th CDI percentile for vocabulary size given their age) have the same vocabulary trend as early talkers (children at or above the 80th percentile). The aggregation essentially suggests that these late talkers do not learn words in a different order, just that they learn words later. This suggestion has been directly examined and shown to be false: typical and late talkers learn not only at different rates but they also learn different lexical items (e.g., Beckage et al., 2011). More generally, limitations of the norms have been noted by many researchers. For example, the norms don’t generalize to all populations (e.g., Arriaga et al., 1998; Thal et al., 1999) and the norms mask idiosyncrasies in an individual’s learning (e.g., Mayor & Plunkett, 2011).

Despite their shortcomings, the CDI norms could still be useful for characterizing an individual child’s lexical growth.

In this paper, we compare predictions based on the CDI norms with predictions based on child-specific sources of information. Specifically, we have two sources of information at our disposal from a data set we'll describe shortly. First, we have *child features*: the child's age, sex, vocabulary size, and language skill as estimated by CDI percentile. Second, we have the specific productive vocabulary of a child at a particular moment of time, as assessed by parent report; we characterize the vocabulary as a binary vector of *word features* indicating whether or not a word is known. These two sources of information come from longitudinal studies. Figure 1 bottom frame shows several examples of specific children's vocabulary trajectories.

We test two hypotheses. First, are child- and word-features as useful as the population acquisition norms for predicting whether a specific word will be learned by a child in a certain window of time? Second, do the child- and word-features provide redundant information, or can the two qualitatively different sources combine to yield greater predictive power than either individually?

The child language literature suggests that information about an individual learner may be useful in predicting the learning of unknown words. For example, the sex of the child is a significant factor in language development as vocabulary size and the sex of the child are correlated: females have larger vocabularies on average than their age-matched male peers (Fenson et al., 1994). Clearly, age is a critical child feature as well: certain words are more likely to be learned earlier than others. The CDI percentile, which is formed by combining information about the child's age and vocabulary size as compared to peers, is itself useful for predicting the specific words a child knows (Beckage & Colunga, 2013). Thus, we find justification for predicting word learning using the child features of age, sex, vocabulary size, and CDI percentile.

Nonetheless, these child features don't tell the whole story. The content of the child's vocabulary may reflect the language learning environment, the child's interests and possibly learning strategies that the child has. Consequently, the words known by the child may be predictive of which words they learn next; co-occurrence of words in a child's vocabulary increases predictability of future language learning above and beyond the normed age of acquisition data (Beckage & Colunga, 2013). Work also suggest that there is a strong relationship between what words a child will learn and the language learning environment of the child (Weizman & Snow, 2001) and their specific interests (DeLoache et al., 2007). These aspects of learning may be better captured by the content of the child's vocabulary than by features related to the child's age and vocabulary size.

In this article, we compare models that utilize child features and/or word features to predict the learning of individual words over a time window of roughly a month. That is, we use information about the child and the child's vocabulary at time t to predict whether an individual word not known at time t will be learned by time $t+\Delta t$. (Ideally, observations are a month apart, but as we explain in the methodology section, Δt varies across observations.) We build logistic regression models for each word individually and include features related to the child and/or to the vocabulary of the child. We discuss the modeling assumptions in detail below but to summarize, we compare performance of logistic regression models to a model based on the age of acquisition data. The performance of the logistic regression models, with child- and/or word- features, helps us understand the features relevant to predicting the learning of individual words, informing our models of lexical acquisition in young children.

Methodology

Vocabulary Data

We use data collected as part of a 12-month longitudinal study in Dr. Colunga's Lab at the University of Colorado Boulder. The data were collected over three recruitment phases in which parents and children came to the lab for recurrent visits over 12 consecutive months. Visits were timed at nearly monthly intervals and, on average, we have 9 visits for each child in the study. Overall, we include 112 monolingual children. At each visit, parents were asked to fill out a vocabulary report. The parental vocabulary report was collected using the MacArthur-Bates Communicative Development Inventory (CDI, Dale & Fenson, 1996) and included 680 commonly used English words. Across all recruitment phases, we have a total of 996 CDI *snapshots* of children's' vocabulary knowledge.

The study represents many different types of language learners with the age of the children in the study ranging from 15.3 months to 33 months. The median age of a child across all the CDIs is 22 months. We also have a full range in language ability represented as well. To approximate language ability, we utilize the CDI percentile which is calculated based on the size of the child's vocabulary as compared to their age matched peers. The range of the CDI percentile represented in the CDI vocabulary snapshots was between 0 and 99, with a median percentile of 54. We should note that recruitment of participants in the longitudinal study was biased to over-represent late-talkers as late-talkers are a population of particular interest in language acquisition.

Of the 680 words on the full CDI, 649 of these words are normed. These 649 words are the words we use to represent an individual child's vocabulary in this study. As part of these 649 words, all types of word classes are represented. The most common are concrete nouns (dog, chair, etc.) followed by action verbs (drink, run, etc.) as well as connecting words, descriptive words and words about time and routine. Because of the variation in the type of words as well as the baseline knowledge of a word (both in the norms and our observed data) we construct an independent logistic regression model for each word. We utilize different information from each CDI snapshot as the input to our model and predict acquisition forward to the next CDI–capturing the probability of learning a specific word in approximately one month's time. Though we model each word individually, we are not interested in performance across different types of words so we consider the performance of a model to be based on the features included in training across all types of words–that is to say a model refers to the features included in training, not the specific word we train on.

Model Construction and Evaluation

We construct separate models for each target word in the CDI. To generate training and test sets for each target, we use the snapshots from all children up to and including the point in time at which the child transitions from not knowing to knowing the target. (We use the terms 'know' and 'learn' loosely; the CDI snapshots are in fact a parent's report of a child's productive vocabulary, however, we hope they capture something about learning and the acquisition process.) The point of transition can vary from one child to the next as well as one target to the next. For example, one child may show initial learning of the word 'dog' at month 18, and if CDIs are available for that child for the preceding months 15, 16, and 17, then that child will provide 3 separate snapshots (predicting to month 16, 17 and 18) from which model training and testing is performed, 2 of which involve a prediction of not knowing the target and one of which involves knowing the target.

We explore a set of alternative models for each target word, as we will describe. The models take as input a representation of a child's snapshot at some time t, and predict whether or not the target is known at the next snapshot, collected at $t+\Delta t$. Specifically, the model outputs the probability of target acquisition at $t+\Delta t$. In all cases of training and test, the target is not known at t. We make predictions conditional the target not being known because once a target is learned it remains known, and one can trivially use the conditional models we develop to make unconditional predictions.

For each target, the full data set is split into training and test sets, and the same split is held constant for all alternative models. The training and test sets are created by selecting all children who do not know the target at the beginning of the study. We then place 80% of the children in the training set and the remaining 20% in the test set. Because the number of children who initially know a word varies across words, the training and test set is created uniquely for each target.

We evaluate each alternative model for each target via the log-likelihood over the test set. This measure weights each prediction equally, and thus later learning children play more heavily into the measure. To obtain a single measure of performance for each alternative model, we sum log-likelihoods over all 649 target words. To determine the reliability of difference between alternative models across targets, we compute a paired t-test treating target as the random variable.

Baseline Normed Acquisition Model

We constructed a baseline *normed model* utilizing published CDI statistics that indicate the normative age of acquisition (Dale & Fenson, 1996). These norms are based on 1130 CDIs collected for children between the ages of 16 and 30 months. In the Dale study, the CDIs are binned by age (rounded to the nearest month) and then the percentage of children who were reported to produce a specific word is calculated. We use these values, for each word, for each month, to estimate the probability of learning a currently unknown word. In the normed model, only one feature is used for prediction: the age of the child. Because the norms exist only for children between 16 and 30 months and the children in our study are occasionally younger or older, we establish boundary conditions–for children over 30 months age or younger than 16 months, we use either the 30 month or 16 month norms.

To use the CDI norms for prediction, we must transform them from a probability of *knowing* a word at a given age to the probability of *learning* a currently unknown word at a given age. The difference between the CDI norms at month m and month $m-1$ might seem like a measure of learning, but the difference is occasionally negative (due to the fact that the data used to construct the norms are cross-sectional: the children in the 16 month group are not the same children as in the 17 month group). To ensure monotonicity of the normed model, we smooth out negative differences by replacing them with the rate of vocabulary change over the minimum time span that yields a positive rate of change.

Because the CDI norms are binned by months and we may be required to make a prediction for a child at age $t+\Delta t$ which may lie between two months, linear interpolation on the smoothed differences of the CDI norms is performed.

Logistic Regression Models

We use lasso regression, a penalized (L1-regularized) logistic regression model that performs feature selection to exclude (set coefficients to zero) features that do not meaningfully contribute to the prediction. In principle, lasso regression serves to regularize a model; that is, it attempts to prevent overfitting by reducing the number of nonzero coefficients in the model. We perform lasso regression in R using the library glmnet (Friedman et al., 2010), which internally performs cross-validation using the training data to select the regularization parameter, and the remainder of the data are used to determine model coefficients.

We develop a set of alternative logistic regression models that differ in the features provided as input. The two sets of features we consider are *child features* and *word features*. The child features are: the sex of the child, the number of words spoken by the child, the CDI percentile, and the age of the child at snapshot t. We include two additional features pertaining to the child: Δt and the session (visit) number. The reason for including Δt is that the time between snapshots is designed to be one month, but this desideratum is not always satisfied and the variation of Δt may be useful for prediction. We include the session number to capture how long into the 12 month longitudinal study the child is. We have found that the child's participation in this longitudinal study positively affects their vocabulary growth and influences their vocabulary size and percentile, thus this child-level feature may affect our ability to predict the acquisition of words.

Turning to the word-level features, we construct an indicator vector with one element per word. The *i*th element of the vector is set to 0 or 1 depending on whether the parent reports that the child can produce word *i* at snapshot *t*.

Results

We first compare performance of the normed model, the child-feature model, and word-feature model (Table 1). The performance is assessed via log-likelihood; values closer to zero are better. As expected, the fit to the training set (column 2) is related to the complexity of the model. The model with the most free parameters, the word-feature model, best fit the training data. However, on the test set (column 3), this model did not perform as well as the child-feature model, due to overfitting of the training set. Nonetheless, both the child- and word-feature models outperform the normed model (using a paired two-tailed t-test, child $t(649)$=44.71, $p <.001$; word $t(648)$=30.62, $p <.001$).

Table 1: Performance of the normed, child-feature and word-feature models.

	llk train	llk test	% best fit	% vs norms
norms	-123881	-31092	3.05	—
child	-84698	-22774	67.64	96.92
word	-65703	-24059	30.51	91.06

The log-likelihood score combines performance across individual words. We can also examine which of the models—normed, child-, and word-feature—performs the best for each word. Column 4 of Table 1 indicates the percentage of words for which a given model outperforms the other two. Column 5 indicates the same for the child- and word-feature models compared separately against the normed model. Consistent with the log-likelihood results, both child- and word-feature models outperform the norms, and the child-feature model outperforms the word-feature model.

Because lasso regression discards input features it deems to be irrelevant, we use the presence or absence of a feature as a proxy for importance. Since a model is trained independently for each predicted word, we determine the percentage of models that include a particular child-feature to measure the importance of a feature. The child-feature models have an average of 4.7 parameters, and range from having 1 parameter (the intercept) to 7 (all child features plus intercept). Across child-feature models, 64.1% included gender, 64.1% included either age and age-at-prediction, but only 22.8% included both, suggesting that the time between visits is less important than the general age of the child. The session visit appears significant in nearly 60% of models. Most important to the child-feature model are percentile and total vocabulary size. Percentile is present in in 87.1% and total vocabulary size at time *t* appears in 73.1% of the models respectively.

For the word-feature model, the number of parameters could range from 1 (intercept) to 650 (each of the 649 words plus the intercept). The actual range based on the model fit for each word was between 1 and 83 features with an average of 31 features. Since only a subset of the 649 words ended up in the logistic regression models we can conclude that a localized representation was at least as useful in predicting acquisition than the full vocabulary. Of the features included in the model, 83% had a positive weight indicating an increased probability of learning the target word if the word was known. We hope to investigate the relationships between individual words, as well as why some of the coefficients in the model were negative in future work.

We can conclusively say that both the child and word features outperform the model based on the acquisition norms. We can also conclude that the set of child features outperform the word-vector features. To see how much of the increase in performance of the child features over the word features was due to overfitting of the word-feature model, we perform a dimensionality reduction on the word features using principal component analysis on all 996 vocabularies, regardless of whether a specific vocabulary is in the training or test set. We use the first 18 components of the PCA reduction based on a Scree plot of the components. We then take the full binary vocabulary of the child and multiply it by the reduced PCA representation for each word, for each snapshot resulting in a vector of 18 features, representing each child's vocabulary snapshot.

Utilizing this representation of the vocabulary data (redux-word), we now find that the total likelihood of this model is less than the child-feature model. We compare this reduced word model to the child-feature model and a model that contains both the child and reduced word features. Column 2 of Table 2 shows the total log-likelihood (llk) for the models based on child features, word features, reduced-word features (redux-word), and both reduced word and child features. Referring back to Table 1, we confirm that all of these models outperform the model based on the CDI production norms. The number of free parameters (which is correlated with performance on the training set) is included in column 3, as are the average number of features seen in each model across all words (column 4). On the test data, model fit is best for the child- and the reduced word feature models. These two models, are not significantly different in a paired t-test ($t(649)$=1.44, $p = 0.148$) but the reduced-word-feature model is significantly better than all other models, we report the results of a t-test between the two redux models ($t(649)$=3.07, $p = 0.002$). We find that the child-feature model is not significantly different than the reduced word model with word+child features.

A priori, it seems likely that adding additional child-features should only improve the performance of the reduced-word (redux-word) model. We instead find that the model with the extra child-features performs statistically worse than only the reduced word-feature model. We believe this is be-

Table 2: Performance of the logistic regression models with different features.

	total llk	poss. params	# params
child	-22774	7	4.70
word	-24059	650	31.20
redux-word	-22654	19	8.39
redux-word+child	-22848	25	9.72

cause lasso regression is a greedy algorithm and therefore, if a set of variables capture more information than a single variable the model will choose the single variable and discard the others. In the case of variables such as vocabulary size or percentile, there may be an ensemble of other variables that represent the feature better–something the greedy algorithm cannot easily capture. To account for this shortcoming of lasso-regression, we construct a final ensemble model in hopes of understanding the role of child-level and word-level features in modeling acquisition.

Our final ensemble model simply averages the prediction from the child- and the reduced word-feature model. This ensemble model significantly outperforms any other model, with a lower total likelihood value of -22263. We confirm this improvement in a paired t-test to both the child- and the reduced word-feature models (child: $t(649)=9.96$ $p<.001$; word: $t(648)=9.52$ $p<.001$). In Figure 2, we visualize the types of information the two models might be using that result in a better ensemble model. The first frame captures the relationship between percentile and vocabulary size, colored based on age of the child, highlighting the information in the child-level features. Frame 2 of Figure 2 captures the information relevant in the dimensionality reduction. We color the data points based on the number of CDIs in which the word was reported as known. The first two components, visualized in the graph, seem to encode information about the population rate of knowing a specific word across all of the 996 CDIs.

Conclusions

Our results show that models can predict the acquisition of a particular target word by a specific child. In contrast, past research has primarily focused on characterizing general population trends in vocabulary growth. We find that two qualitatively different sources of information are useful for prediction: features that describe the child (such as age and sex and total vocabulary size) and features that specify the vocabulary content. Models based on either child- or word-features outperform the traditional age-of-acquisition norms in predicting whether a specific word will be learned by a specific child.

We investigated which of the child features were most useful for prediction via lasso regression, and found that CDI percentile (chosen for 87% of models), vocabulary size (chosen for 75% of models), and age (chosen for 64% of models) were common features. Although CDI percentile is a function of both vocabulary size and age, the three features were often (38% of models) included together in the model for a specific word, consistent with previous work suggesting that CDI percentile contains useful information about the interaction between age and vocabulary size as compared to the peer group (Thal et al., 1999; Beckage & Colunga, 2013).

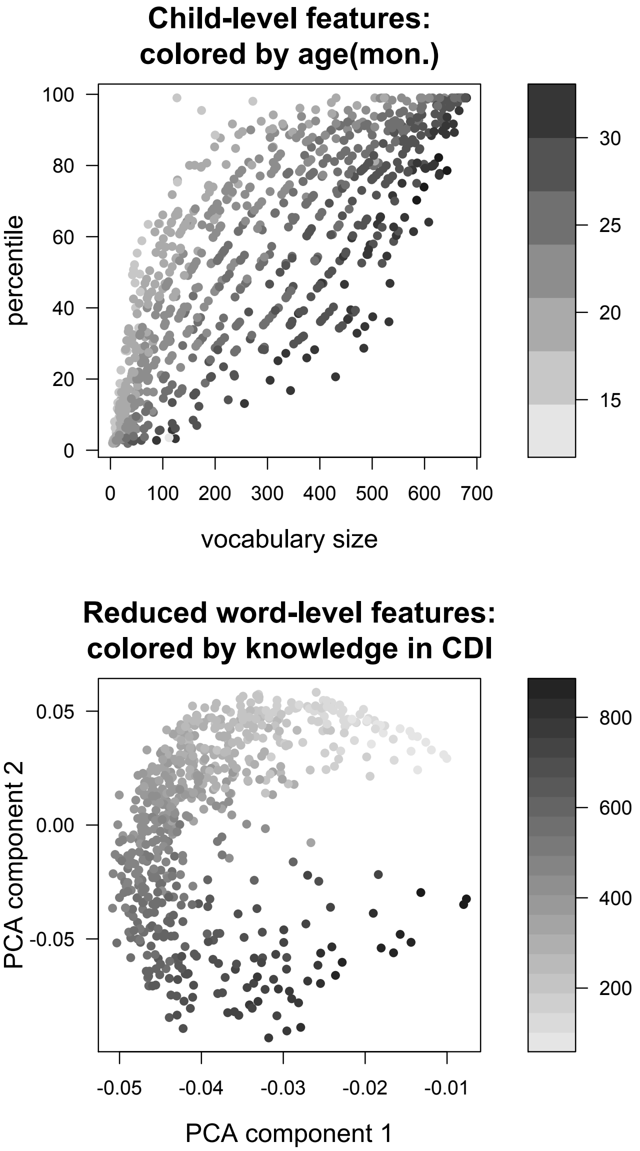

Figure 2: Graphical representation of child-level and word-level features.

The success of the word-feature representation, both with and without dimensionality reduction, indicates that the content of the vocabulary is predictive of language learning. This result is exciting because understanding how the known vocabulary supports future vocabulary learning provides a new opportunity for understanding the developmental process (Smith, 2000). Further, this type of modeling can potentially be extended to interventions: if we know how words build on one another, can we teach children certain words to create vocabularies that are more useful for future language learning?

We found that reducing the word-feature vector via principal components analysis improves model performance compared to using the original word-feature vector. This reduction is beneficial because PCA performs noise suppression when we drop non-primary components and because it reduces the opportunity for overfitting. In addition, the reduced representation may be more interpretable and psychologically relevant. Referring to Figure 2, we see that the frequency at which the word is known in the overall set of vocabularies seems to be strongly related to the first 2 components but other components might include semantic or phonological features that can tell us more about the process of acquisition. We plan to explore this research question in more detail in the future.

Perhaps our most important finding is that the child and word features are complementary. This complementarity was not evident when we constructed a single regression model with both sets of features, but stood out when we combined predictions of child-feature and word-feature models. The combination, obtained by averaging the two models' outputs, achieves a statistically reliable improvement in prediction. The resulting ensemble is proof that the child and word features contain different types of information, both of which are useful for predicting future language learning.

The key value of modeling in this domain is to help us understand the sources of information that aid in prediction the acquisition of new words. We showed in this work that both child and word features are useful, and that the nature of representation matters (e.g., unreduced versus reduced word vectors). Clearly, there are many other source of information that could be incorporated into a model, such as demographic characteristics, the linguistic environment, and cognitive and motor assessments of the child. Of course, obtaining these measures can be costly, and future modeling will be directed at determining which measures provide the most diagnostic features. One dimension we have begun exploring is the semantics and phonology of the child's productive vocabulary. In our present work, we treat the words as independent symbols, but in principle a word representation which characterizes known words and the target word in terms of semantic and phonological features could be utilized.

Beyond exploring new types of features that might be useful in modeling language acquisition, we would also like to expand the class of models used to predict acquisition. The most natural extension of logistic regression is a multilayer neural network. In a network's hidden layer, we can look for the emergence of new features that have psychological plausibility. Indeed, the success of our ensemble model suggests that an intermediate level of representational transformation may serve the prediction task. Although the models we have focused on in this work are not intended to characterize cognitive and developmental processes per se, the representations found to be useful for prediction should inform *cognitive* models of child language development.

Acknowledgments

This work was funded through an award from the John Merck Scholars Fund and by NICHD grant R01 HD067315 to Eliana Colunga. The first author was funded in part through the NSF GRFP. Data was collected by the DACS Lab at the University of Colorado Boulder with much help from the DACS Lab undergraduates and the Boulder community parents.

References

Arriaga, R. I., Fenson, L., Cronan, T., & Pethick, S. J. (1998). Scores on the macarthur communicative development inventory of children from lowand middle-income families. *Applied Psycholinguistics*, *19*(02), 209–223.

Beckage, N. M., & Colunga, E. (2013). Using the words toddlers know now to predict the words they will learn next. *Proc. of the 35th Conf of the Cog. Sci. Society*, 163-168.

Beckage, N. M., Smith, L. B., & Hills, T. T. (2011). Small worlds and semantic network growth in typical and late talkers. *PloS one*, *6*(5), e19348.

Dale, P. S., & Fenson, L. (1996). Lexical development norms for young children. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 125–127.

Dale, P. S., Price, T. S., Bishop, D. V., & Plomin, R. (2003). Outcomes of early language delay. predicting persistent and transient language difficulties at 3 and 4 years. *Journal of Speech, Language, and Hearing Research*, *46*(3), 544–560.

DeLoache, J. S., Simcock, G., & Macari, S. (2007). Planes, trains, automobiles–and tea sets: Extremely intense interests in very young children. *Developmental Psychology*, *43*(6), 1576-1586.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... Stiles, J. (1994). Variability in early communicative development. *Monographs of the society for research in child development*, *59*(5), 1–185.

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22.

Mayor, J., & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from cdi analysis. *Developmental Science*, *14*(4), 769–785.

Smith, L. B. (2000). Learning how to learn words: An associative crane. In *Becoming a word learner: A debate on lexical acquisition* (pp. 51–80).

Thal, D. J., O'Hanlon, L., Clemmons, M., & Fralin, L. (1999). Vaidity of a parent report measure of vocabulary and syntax for preschool children with language impairment. *Journal of Speech, Language, and Hearing Research*, *42*(2), 482–496.

Weizman, Z. O., & Snow, C. E. (2001). Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*, *37*, 265-279.