Quantifying the role of vocabulary knowledge in predicting future word learning

<table>
<thead>
<tr>
<th>Journal:</th>
<th>IEEE Transactions on Cognitive and Developmental Systems</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manuscript ID</td>
<td>TCDS-2017-0144.R1</td>
</tr>
<tr>
<td>Manuscript Type:</td>
<td>SI: Language Learning</td>
</tr>
<tr>
<td>Date Submitted by the Author:</td>
<td>06-Mar-2018</td>
</tr>
<tr>
<td>Complete List of Authors:</td>
<td>Beckage, Nicole; University of Kansas, EECS; Colunga, Eliana; University of Colorado Boulder, Psychology and Neuroscience Mozer, Michael; University of Colorado, Computer Science</td>
</tr>
<tr>
<td>Keywords:</td>
<td>neural networks for development, language acquisition through development, cognitive system and development</td>
</tr>
</tbody>
</table>
Quantifying the role of vocabulary knowledge in predicting future word learning

Nicole Beckage, Eliana Colunga, and Michael Mozer

Abstract—Can we predict the words a child is going to learn next given information about the words that a child knows now? Do different representations of a child’s current vocabulary knowledge affect our ability to predict the acquisition of lexical items for individual children? Past research has often focused on population statistics of vocabulary growth rather than prediction of words a child is likely to learn next. We consider a neural network approach to predicting vocabulary acquisition. Specifically we investigate how best to represent the child’s current vocabulary in order to accurately predict future learning. The models we consider are based on qualitatively different sources of information: descriptive information about the child, the specific words a child knows, and representations that aim to capture the child’s aggregate lexical knowledge. Using longitudinal vocabulary data from children aged 15-36 months, we construct neural network models to predict which words are most likely to be learned by a particular child in the coming month. Many models based on vocabulary information outperform models with child information only, suggesting that both the words a child knows—and the way those words are represented— influence prediction of future language learning. Through the exploration of these models, we gain an understanding of the role of current vocabulary knowledge on predictions of future lexical growth.

Index Terms—Language acquisition; word learning; lexical acquisition; neural networks; cognitive development

I. INTRODUCTION

What role does the current lexical knowledge of a child have in accurately predicting future word acquisition? If all children learn in approximately the same way, knowing the specific words in a child’s vocabulary should not improve accuracy at predicting what words the child is likely to learn next. Alternatively, if the idiosyncratic words a child knows at a given time influence the words they are going to learn next, this suggests current lexical knowledge influences future lexical growth. Even assuming a child’s vocabulary is predictive of the words a child is likely to learn next, it is possible that the word itself is an artifact of learning a relevant feature or category in the world. If this is the case, the knowledge of the specific lexical item may be less predictive than the concepts or features it represents. For example, a child might learn the word dog; she might have learned that her household pet is the only dog, or that only animals walked around on a leash in her neighborhood are dogs, or instead she could be learning that dogs have four legs, a tail, etc. and that dogs are somehow different than cats. All of semantic information related to the word dog capture different types of language knowledge that a child might use to learn language. Here we explore how various representations of a child’s current vocabulary predict that child’s future lexical acquisition. We compare the usefulness of different network representations by comparing predictive performance of single-layer neural networks, evaluating the models on their ability to predict the words a child will learn one month into the future.

We tackle the problem of lexical acquisition in toddlers because language learning is one of the first complex cognitive tasks humans undertake, and therefore a great way to model learning more generally. Infants start producing their earliest words around 12 months of age and within only a few months, young children have hundreds of words. Shortly thereafter, young children begin to construct sentences with complex ideas, and grammatical structure. Despite how quickly this learning comes online, much of the language acquisition process is still challenging to explain—particularly how children represent and access language knowledge. The approach of machine learning to model complex processes, such as language acquisition, can provide novel insight into the learning and representation of language. We focus particularly on how different vocabulary representations of a young child’s lexicon increase predictive accuracy. Pairing powerful statistical learning tools with observational acquisition data, we can isolate differences in individual learning in early acquisition and quantify the role of current vocabulary knowledge on future vocabulary growth.

Currently, a toddler’s lexical knowledge is often measured as the number of words they know, given their age and sex [8], [10], [31]. While there is strong evidence that this number is useful in assessing language ability, it is unclear that it is useful in predicting future acquisition. In our age of interest, the most commonly used measure is the MacArthur-Bates Communicative Development Inventory (CDI). Parents indicate which of a fixed set of about 700 words their child can say. From this vocabulary report, developmental psychologists assign each child a CDI percentile that compares the child’s CDI vocabulary size to that of their peers. This percentile value is used to flag children who are learning language at slower rates than their peers. These children, classically called late talkers, are important to monitor because many of them will continue to have language learning difficulties [14], [51]. Sometimes these language-specific difficulties will develop into wider reaching cognitive difficulties. However, not all children who have a low CDI percentile go on to have lasting language difficulties. By exploring different types of language representations in predicting future acquisition, we can help uncover relationships between current language and future learning that could help with diagnostic assessment. The relationship between linguistic
representation and CDI percentile might suggest an approach for quantifying meaningful differences in learning between at-risk children and their normally developing peers. However, before these questions can be directly studied, a working predictive model of acquisition must be constructed and studied. Here we consider a simple neural network modeling approach.

Neural network models, often called connectionist models in psychology, provide a systematic way of extending observational findings and behavioral studies of early language learning. Single layer neural networks, while more complex than generalized linear models, still provide interpretability and insight into aspects of linguistic knowledge that may impact future language learning. As statistical learning tools, neural networks are powerful and adaptive, capable of modeling change over time, and dealing with noise and uncertainty in the data. Here we use neural network models to build predictive models of lexical acquisition. We specifically explore how the representation of a child’s current vocabulary influences our ability to predict what words a specific child will learn next.

Evaluating our predictive models on longitudinal language acquisition trajectories, we interpret the model accuracy as evidence of the importance of a specific type of lexical knowledge representation in early acquisition. In the larger literature of neural networks, neural networks are automatic and optimized feature detection systems, whose training algorithms work to find the best combination of complex features that accurately predicts the measure of interest. In our simple neural networks, the intermediate layer aggregates input features into representations that maximize predictability of future language learning. By considering only simple graphs, we allow for predictive models that are slightly more complex than generalized linear models but whose performance is still limited by the usefulness of the input representation. We compare model performance as a direct means to assess the usefulness of a particular network representation in capturing the relationship between current lexical knowledge and future language learning.

We first briefly review the state of the art in using neural networks to capture aspects of language acquisition, before turning to the methods, and detailing the longitudinal data and vocabulary representations. Next we discuss the neural network training and optimization. Finally we discuss the different predictive capacity of the various vocabulary representations and the implications of the results.

II. PAST WORK

Neural network models as applied to early learning has a long history which we review only briefly here. The interested reader may find a more extensive review of neural networks applied to the cognitive sciences here [4], [19] and a specific review of semantic development here [29]. Previous research also explores neural network approaches to link neuroscience to early development [25] and to semantic cognition [20]. We limit our literature review specifically to neural network models of lexical acquisition as our prediction task aims to capture learning of specific lexical items.

Much of the past connectionist work to model lexical acquisition focuses on capturing infant performance on behaviorial learning tasks, with the goal of providing a mechanistic explanation of language learning in children that can be verified experimentally. For example, connectionist networks have been used to understand the role of associative learning on the emergence of word learning biases [7]. Work using neural network models tasked with learning word-to-object mappings, trained on a vocabulary in the CDI, acquires useful word learning biases, even though they are not directly rewarded for this bias. These models have been shown to make novel predictions – about learning for different types of categories, learning different languages, or different language proficiency – that have subsequently been verified [5], [6], [23], [28].

Other examples of neural network models applied to modeling early lexical acquisition include models capable of capturing word confusability and age of acquisition effects [17], and the formation and degradation of conceptual categories [20]. These, and other examples of lexical development, explain behavior with basic mechanistic accounts of associative learning. One neural network model, which learns to map word-forms to object referents [21] shows a mutual exclusivity bias — a preference for novel words to map to novel objects [23], even though no training instances explicitly exhibited this bias. The model uses this bias and associative “knowledge” of other words to quickly and accurately learn new words, even in highly ambiguous contexts. Another neural network aims to model acquisition of categories. The neural network provides evidence of a feedback loop between perceptual features and linguistics labels [35]. Linguistic labels are thought to support generalization of categories. The model itself is able to capitalize on the relationships between category formation and language learning to provide structure and reinforcement during learning.

Unlike the work reviewed above, we do not focus on neural networks as cognitive models. We instead use neural networks as a means to uncover associations in the environment and language knowledge of a child that might be relevant and even facilitatory to the lexical acquisition process of young children. Neural network models are useful tools for modeling development because the associative learning framework allows for different types and timescales of learning to be captured within a single representation. This is mostly due to the ability of connectionist models to incrementally learn and to have predictive capacity even when representations are underdetermined or noisy. We leverage the robust learning of neural networks to provide quantification of the expressability of different vocabulary representations in relation to future lexical acquisition.

A key idea to data-driven neural network models of acquisition is that there are regularities in the way in which children learn. But the differences are also informative. If all children learn similarly, and the variability is not predictive, then high-level features such as the age of the child should be adequate in predicting lexical growth. But if there is variability among learners that can be assessed from the vocabulary data directly, then the data-driven approach can offer unique insights into these trends. Previous work suggests that different types of learners exist and that there are meaningful similarities in learning within these different types [14], [18], [25], [31]. For
example, network analysis approaches have found that not only are late talkers learning slower than their peers, but the resulting vocabulary is less structured than one might expect if the children were simply learning at a slower rate [1]. Assuming that there are different types of language learners and that the vocabulary at any time point reflects the type of learner a particular child is, machine learning models may provide a powerful and predictive tool to aid with classification and diagnostics of a child’s learning trajectory.

Many features of the language environment likely affect learning. We assume that the content of the child’s vocabulary contains information about the most influential forces directing the child’s learning trajectory. While we are agnostic as to which specific features influence and direct learning, we do assume that representations that accentuate relevant features will result in an improvement in model accuracy. We consider the performance of various language representations to the baseline child-feature model as a way to quantify the influence of certain language representations on predictions. With the goal of capturing the role of a child’s current vocabulary on future language learning, we explore different ways of representing the child’s current vocabulary knowledge. Our baseline model considers features of the child such as their age, total vocabulary size, and CDI percentile. If all children learn similarly, then these features should be informative and predictive of which words the child is likely to learn approximately in the future. Alternatively, if the idiosyncratic vocabulary of a child reflects the child’s interests in specific themes (for example, animals) or their language environment drive learning, then knowing the semantic content of their vocabulary will be helpful in capturing future lexical acquisition.

III. METHODS

A. Longitudinal vocabulary data

To train and evaluate the neural network models, we use data collected as part of a 12-month longitudinal study in the Colunga Lab at the University of Colorado Boulder. The data were collected over three cohorts. Parents and children visited the lab monthly for a year. Visits were timed near monthly intervals and, on average, we had 10.9 visits for each child. Overall, we included 83 monolingual children (37 female) in our current analysis. At each visit, parents completed a vocabulary report indicating which, of a fixed set of words, their child produced. The parental vocabulary report was collected using the MacArthur-Bates Communicative Development Inventory (CDI) [8] for children between 16 and 30 months. Our modeling work includes 677 of the CDI’s 680 early learned words. Three words (grass, slide (noun) and work (noun)) were excluded from our analysis due to missing data. Figure 1 includes an example of what the CDI data look like. Across all recruitment phases we have a total of 908 CDI vocabulary snapshots which form 825 CDI vocabulary snapshots (i.e., two sequential vocabulary reports). We define a CDI snapshot as the first CDI report transformed to be the input to the neural network and the output vocabulary. In all cases, our model is given information pertaining the content of the first CDI report in the snapshot and is tasked with predicting the vocabulary as measured by the later CDI. While the time between CDIs is usually one month, there is some variability due to scheduling issues. We attempt to control for this variability by including the time between CDIs as an input feature to all neural network models.

The longitudinal study represents many different types of language learners with the age of the children ranging from 15.4 to 32 months of age during the course of the study. The median age of children on their first CDI is 16.4 months. We also have a full range of language ability represented, as estimated via the CDI percentile measure. This measure is calculated based on the size of a child’s productive vocabulary as compared to the child’s age-matched peers. The range of the CDI percentiles represented in the longitudinal snapshots is between 3 and 99, with a median percentile of 54. We note that recruitment of participants in the longitudinal study was biased to over-represent late talkers, or children in the bottom 20th percentile, as late-talkers are a population of particular interest in language acquisition research.

<table>
<thead>
<tr>
<th>age</th>
<th>sex</th>
<th>voc. sz</th>
<th>dog</th>
<th>house</th>
<th>zoo</th>
</tr>
</thead>
<tbody>
<tr>
<td>16.2</td>
<td>F</td>
<td>32</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>17.1</td>
<td>F</td>
<td>49</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18.9</td>
<td>F</td>
<td>132</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>19.3</td>
<td>M</td>
<td>257</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>20.5</td>
<td>M</td>
<td>345</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

Fig. 1. Example of longitudinal CDI data used as input and output of the neural network. Note that only the individual words are part of the output level of the neural network.

B. Neural network training

Neural networks were constructed and fit using Torch7, a scientific computing framework for LuaJIT. Models are trained via stochastic gradient descent and have a single hidden layer, optimized in size for each trained model. Models were trained using adaptive stochastic gradient descent; weight decay and momentum were used to determine how long to train the model. The network architecture had a variable number of input features based on the vocabulary representation, a single hidden layer and a logistic transformation on the output layer such that the probability of learning a specific word was returned by the model. Learning rate \( \alpha \), number of hidden units \( hu \), batch size, number of epochs until learning rate is near zero \( \alpha \) decay, and momentum \( m \) were optimized via step-wise optimization (e.g. learning rate was optimized first, followed by the number of hidden units etc. with momentum optimized last.) Table I shows the neural network hyper-parameters for each model. Dropout rate of the hidden units was fixed to 0.5. We note that there may be better neural network architectures and gradient descent parameters that could be uncovered by more sophisticated optimization procedures but the greedy-search procedure was effective for the comparison of interest. During training, the gradients are only back-propagated for those words that are learned by the model, thus the model is no penalized, nor are the weights updated, for incorrect predictions on words that are already known by the child.

Most of the step-wise optimization was used to determine the neural network architecture that best suited the particular representation of current lexical knowledge. However, some
of the parameters directly affected the update of the internal model weights. We review them quickly here as this provides increased interpretability to model optimization. Learning rate decay allows models to quickly learn initial patterns but also adapt later in training to more nuanced patterns and negates the need to determine stopping criterion since the learning rate asymptotes to zero. Momentum ensures each update is a combination of the current error gradient and the error gradient accumulated from previous time steps. Dropout was used to minimize overfitting and was fixed at 0.5; so during training, the model only had access an expected 50% of the hidden units. During model evaluation, all hidden units were available. Overall parameter selections (including input feature size) are presented in Table I. We note that optimization happened via 5-fold cross-validation at the child level such that all data for a particular child was in the same fold. Thus, model performance is based not only on generalization to unseen vocabularies, but also to unseen children.

IV. NEURAL NETWORK MODELS

We ask two main questions with this work. 1) To what extent does the vocabulary knowledge of a child increases predictability of which words the child will learn next? It is possible that children generally learn words in a certain order, and that knowing the specific lexicon of a child is not helpful for our predictive models. 2) Assuming the set of words a child knows is predictive of the words they will learn next, how can we best represent the lexical knowledge of a child to our simple neural networks? It is possible that different feature representations of a child’s vocabulary knowledge may allow for a more robust and accurate predictive model of the words the child is likely to learn next. We use two broad types of representations. First, we explore representing vocabulary knowledge by decomposing individual words into lower level units, for example breaking down the sounds of the words to get a phonemic level description. Second, we consider representations that aggregate word knowledge, for example aggregating latent space vectors to capture a multidimensional description of the words a child knows.

This leads to 6 models: the (1) CDI child feature model based demographic information of the child, a (2) CDI word model based on the CDI vocabulary report of a child’s productive vocabulary, a (3) Semantic model based on the semantic features of particular words in the child’s vocabulary, a (4) Phonology model which considers the child’s phonological composition of their productive vocabulary, the (5) CDI label model which captures the production of words within particular categories and a (2) Word2Vec representing the child’s productive vocabulary as a combination of vectors in a high-dimensional linguistic space. Finally, we construct ensemble models as a way to explore whether the types of language representations are redundant or whether the various representations increase model predictability. We further motivate these questions below.

A. Lexical knowledge and input representations

In our first experiment, we explore whether vocabulary knowledge is helpful in predicting future language learning. To this end, we train two neural network models, one that only has access to information related to the child’s developmental stage and another neural network with the additional information as to the specific words on the CDI that are currently in the child’s vocabulary. We call the demographic model the (1) child feature model. This model, with a total of 6 features, includes the child’s age (both at time of CDI collection and time at CDI prediction), vocabulary size, sex, number of visits to the Colunga lab, and CDI percentile. This model is the simplest model and contains all the information researchers usually use to assess a child’s lexical knowledge and approximate their language ability.

We create the (2) word feature model by combining the features in the child model and a 677 binary word vector indicating if the child reportedly produces each specific word on the CDI or not. It is not clear that knowing the child’s current productive vocabulary will outperform the child model which has access only to the child features, as there is much more individual variability in the words a specific child knows. The variability may wash out meaningful signals from which the neural network would learn. In fact, previous work on logistic regression models found that the child-features outperformed a model based on the individual words the child knows [2]. In the neural network approach, we explore this question again, asking whether the content of the child’s vocabulary improves model accuracy in predicting future language learning.

Intuitively, it is also possible the CDI word-feature model will be the best performing model. The neural network has access to input that may allow for the learning of individualized trajectories for each word, capturing both temporal dependencies (like boat is usually learned later than car) and relational dependencies (such as red is usually learned in relation to blue). Further, the neural network model, even with only one hidden layer, has internal states that may allow the model to aggregate this information in useful ways, increasing predictive accuracy. Alternatively if there is systematicity in word learning at a level different than the individual words, the predictability of this model may be less than other vocabulary representations. For example, if the number of animal words a child knows is important for predicting future learning of animal words, this model may perform less accurately than a model that clusters words based on phonological or syntactic categories.

Turning to our final question, we explore how representing lexical knowledge in different ways may affect the predictability of future language learning. Here we introduce a few representations that consider language knowledge at a different scale.
than individual words. We consider types of representations that try to break down the words into specific features and those that aggregate the words into categories or higher level representations. We choose this perspective as a means to assess whether the neural network can more accurately predict future word learning from lower-level or more abstract information about a child’s lexicon. This may help direct future work in understanding the role of different kinds of features in driving early language learning.

We first consider two ways of representing the child’s current vocabulary in a more fine-grained way—one based on semantic features, and the other based on phonological information. We considered the McRae feature norms [23] as an approximation of features related to concrete nouns that might bolster early lexical acquisition. These norms were collected based on adult judgments in which individuals were asked to list features of concrete nouns. Features were aggregated to capture general types of features such as taxonomic and encyclopedic features (e.g. taste, animacy, fact, description) [1]. We use the McRae features (e.g. planes have wings) and the number of each type of feature (e.g. number of taxonomic features of a plane) as input to the neural network. The McRae feature vector representation is 30 continuously valued input features from the McRae feature dataset (and include things like word length, number of taxonomic features etc.). This particular representation only overlaps with about 200 of the 677 CDI words, namely the concrete nouns. To approximate the whole vocabulary knowledge of the child, we consider the average of the individual features of the child’s productive vocabulary assuming that word is in the McRae feature data set. Even though the input representation is only based on nouns, we still evaluate the model on the prediction to the whole set of CDI words. We call this representation (3) Semantic. Previous work has found this representation has minimal predictability in accounting for acquisition of young children using network analysis [15]. Here we test the usefulness of this representation in training a neural network model.

We then consider the phonemic composition of indi have a single hidden layer, optimized individual words. Past work shows the sounds of words play a significant role in learning [29], and that computational models can capture this effect [32]. Here we consider the individual words a child produces and construct a vector representation of how many times a given phoneme appears in the child’s current vocabulary. IPA transcription is done using lingorado.com. For words with multiple transcriptions, we consider the form related to the American accent and also the most common transcription. We took an approach of broad transcription, ignoring subtle and dialectical variants. In total, we consider 37 different phonemes (including diphthongs). Each word is a vector representation of the 37 phonemes indicating the count of the number of times each phoneme appears in the word. Each word is aggregated together in order to represent the whole vocabulary of the child. The vocabulary representation includes the distribution of phonemes in a child’s productive (CDI) vocabulary. Research related to phonological importance in early learning suggests there is a strong effect of word onset and word rhyme [13], but other work has instead suggested phonemic awareness is a better predictor [16]. While this approach of modeling acquisition with neural networks could provide some insight to this debate, for this work we consider only phonemic content and ignore location of the phoneme in the word. We call this representation (4) Phonology.

To represent the aggregated lexical knowledge, we first consider a measure of categories such that input to the neural network includes the number of items in a particular category that the child knows. On the CDI form itself, words are classified into 22 different linguistically informed groups, capturing semantic themes such as “animal”, and “people”, and grammatical classes like “action words”, and “helping verbs” [30]. There is also a class that contains sound effects, including words like “owie” and “woof”. Using these classes, we represent the child’s current vocabulary as counts of the number of words the child produces from each class. Each class does not have equal representation in the CDI and we do not normalize by the size of the class. Instead, we let the network learn both the frequency of each class and the predictability of that class in future language learning simultaneously. This representation suggests what word a child learns next is related to the collective categories of words the child knows now. For example, this model may more easily pick up on a child’s preference for learning food words. This preference could be due to specific interests of the child [9], the parent to child speech [33], or other features of the environment. We withhold judgment as to what aspects of learning might motivate the accuracy of this model, testing instead whether or not this type of vocabulary representation can capture future learning language. We consider this to be the (5) CDI label representation.

For our final representation, we consider an aggregate representation that has been particularly useful in modeling adult language. Word2Vec uses a large corpus of data to build up a rich representation of words [24]. We explore the use of this representation to capture child language acquisition. Using the Word2Vec algorithm, which considers co-occurrence frequency of words, we constructed a 200 dimensional vector representation of nearly all words in the CDI using the large GoogleNews corpus to build our representation. We assume this is an aggregate representation rather than a decomposition because Word2Vec requires co-occurrence information, resulting in words that have context and relational information. Vector representations of compound words, like peanut butter, are constructed by averaging the individual representations of the component words. Natural language processing models using Word2Vec representations have found syntactic, co-occurrence, semantic, and even phonological information embedded in the complex representation [24]. We consider this representation as input to the neural network under the assumption it captures the complexity and relationships of the language children eventually learn. We call this input representation (6) Word2Vec.

We believe that by extending the representation of each word to a vector representation, rather than a single value, the model will more accurately capture the language learning of individual children. We also suspect that some of these representations will fail to account for language acquisition. This failure may suggest features which are not readily available to young children. One additional consideration of these representations...
is how to aggregate the word specific vectors to accurately represent a child’s productive CDI vocabulary knowledge. We consider both averaging and summing the individual word vectors. In the case of averaging, vocabularies are size-invariant and have the same relative activation across all children and age. When summing the individual vectors, information regarding the child’s age and vocabulary size is indirectly measurable based on the activation level. Beyond the method of aggregation, we also consider whether we see an improvement in predictability when we include the child specific features of the (1) child model. We consider both a model with and without the child features because many features are highly correlated with child demographic information. With limited data, the high correlation among features can negatively affect model performance.

All in all, we construct four different variants for each feature representation. One averages the individual word representations and one sums the word representations. We also consider the effect of adding in the child features to each of the input representations. We compare the performance of these models to the CDI features directly. In practice, these models contain different amounts and types of information. Figure 2 visually represents the vocabulary of child under the input variants discussed above. The top row assume individual word representations are summed, while the bottom row illustrate the averaging of word representation for each child. The age of the child at the time of the CDI is indicated along the x-axis. Words are roughly sorted based on parts of speech (e.g. noun, adjectives, verbs).

V. EVALUATION

We first evaluate models based on their performance in terms of minimizing negative log-likelihood (NLL) error on the validation set. Error is computed only for words that were unknown by the child at the beginning of the snapshot, thus we do not penalize the model for incorrectly predicting that known words stay known. Once network architecture is optimized for each representation, we select a single model to investigate. Only after fixing the network architecture and hyper-parameters do we consider the test set.

We evaluate performance by averaging the negative log-likelihood (NLL) error of all predictions. Note, this more heavily penalizes the vocabulary snapshots of children with smaller vocabularies as we only predict unknown words but gives us more insight into the ability of the model to predict learning. We also estimate predictive accuracy based on percent overlap and receiver operating characteristic (ROC) measures. Percent overlap measures the overlap between the \( k \) words reported as learned by the child and the \( k' \) most likely learned words as predicted by the model. The percent overlap measure approximates how accurate the model is at correctly predicting which words are learned but does not consider correct predictions of words that are not learned. We report the median percent overlap across children. ROC curves compute the trade-off between true positives and true negatives as the cutoff for converting probabilities into binary varies. Summary measures of accuracy and discriminability (d-prime) are reported. We assume for accuracy and d-prime the threshold is the point in which learning events are predicted with equal frequency to what we observe. Area under the ROC curve (AUC) is reported and summarizes the accuracy of the model in a single number. Perfect accuracy would have an AUC value of 1 and chance would be 0.5. Also included is the t-statistic from a paired t-test on average NLL of the specific model compared to the CDI child model.

Accurately predicting individual word learning has many applications. But simple predictive assessment may mask developmental changes. For example, assume children attend to phonological features early in language learning only then to attend later in development to semantic features. We would then expect the phonological feature neural network to be particularly adept at predictions of young children or children with small vocabularies. We would also predict semantic networks to capture changes in productive vocabularies of older children with higher accuracy. Thus, we consider performance variability related to the child’s language ability, age, and vocabulary size.

Just as we consider the effect of performance on individual children, we can also compare performance across individual words. It is possible that the representation based on the (3) Semantic feature norms [22] will be extremely accurate at predicting the acquisition of concrete nouns but generalize less well to action verbs or abstract nouns. We investigate this by considering the performance of models based on the average age at which a word is learned. Because earliest learned words are often concrete nouns [11], we might expect the models with semantic information to perform best early in development. Further, if certain words are predominantly learned by children of a certain age, and other words are learned based on individual differences, we can expect overall accuracy differences when considering individual word acquisition patterns [18].

We also consider ensemble models where we combine the individual predictions of the neural networks to increase predictive accuracy. We aim to further capture what types of vocabulary representations are most useful in predicting future lexical acquisition. We describe these ensemble models after discussing the results and performance of the current neural network models mentioned above.

A. Baseline performance

Negative log-likelihood (NLL) is a useful and efficient metric for training neural networks; but as a measure, it can be
difficult to interpret. To understand performance of these neural network models, we orient the readers by introducing a few NLL scores for comparison. If the model always returned 0.5 as the probability of learning a word, the average NLL score of predictions would be 0.631. If we condition on words such that the model returns the probability of learning a given word proportional to the empirical data, the result is a NLL score of 0.496. We can further improve this basic prediction by conditioning on the age of the child. Here we can use two independent predictions. One is from the published CDI norms [8] which indicate the proportion of children at a given age who reportedly produce a specific word. We can also estimate the learning rate of individual words directly from the data. Using the published CDI norms and the empirical age of acquisition results, we get a NLL of 0.456 and 0.453 respectively. Values closer to zero indicate better model performance.

Our final (informed) baseline NLL measure uses logistic regression models for prediction. Training an individual logistic regression for each word, we predict, given a child’s current vocabulary, if the child learns a specific word. Aggregated to predict the whole vocabulary of a child, we find a negative log-likelihood score of 0.391. See Beckage, Mozer and Colunga for more detail on the modeling framework and results [2]. Any model that contains useful information to word prediction must clearly outperform this logistic regression model by attaining a score smaller than 0.391.

All neural network models outperform logistic regression models. Of the models tested, the model with the worst performance still had a negative log-likelihood error of 0.32. We turn to comparing neural network models directly. As mentioned above, all models were individually optimized for learning rate, batch size, number of hidden units, momentum, and learning rate decay. We ignore the specifics of optimization in favor of comparing the results using the best architecture. First we consider the CDI-based models and then we turn to the feature based models.

B. CDI models

As discussed in [V-A], we construct and train a global (1) Child model that includes only the child features as our baseline model. We then compare performance of this model to the (2) Word model which includes the individual vocabulary words that the child can produce for a total of 683 features. In Table[II] we report summary performance of these CDI representations by average NLL for all 171 snapshots (17 children) in the test data. Table[II] shows that the CDI word representation has the lowest negative log-likelihood error, and performs with accuracy near 84.2%, suggesting that neural networks are capable of predicting future word learning for individual children. These models may capture specific populations of learners, performing better for a subset of the snapshots explored. To examine this idea, we plot the difference between the (1) CDI child model and the (2) CDI word model along the x-axis. We then consider features of the child along the y-axis.

We find that the (1) CDI Child model performs well for children with higher percentiles but not as well for children with small percentiles. In Figure[III] from left to right, we order snapshots by the vocabulary size, and percentile. We normalize the x-axis so positive values indicate that the (2) CDI word model is performing better and negative values indicate that the (1) CDI child model is performing better.

C. Feature-based models

While the individual words a child knows as recorded by the CDI are useful in predicting the words the child will learn next, we are also interested in whether the CDI is the best representation of the content and structure of a child’s current productive vocabulary. It may be that by representing a child’s vocabulary as an aggregate set of word-feature representations, we can outperform the CDI models. In this section we discuss the resulting model performance when using our (3) Semantic features, (4) Phonology, (5) CDI label and (6) Word2Vec representations. As mentioned above, we also consider whether averaging or summing the individual word features produces the best predictions. We also briefly discuss whether adding additional child specific information such as age improves performance of the language representations.

We find different aggregation processes of the vocabulary, even within a specific representation, have large effects on the ability for a model to predict future lexical acquisition. Across all representations, there is no best aggregation method. Two of the models (the (5) CDI label and the (3) Semantic feature norms) performed significantly better when including child specific features, suggesting that the child information of age, percentile and vocabulary size are not independently useful for some representations. The other models, including phonology, saw no reliable improvement when including the child features. Additionally, half of the representations were most predictive when the individual word representations were summed across the whole productive vocabulary. The remaining showed increased accuracy when the individual word features were averaged. For the rest of our analysis we choose the

---

TABLE II
Neural network performance of 6 different models.

<table>
<thead>
<tr>
<th>model</th>
<th>NLL</th>
<th>% overlap</th>
<th>AUC</th>
<th>acc</th>
<th>d prime</th>
<th>t-stat</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDI child</td>
<td>.312</td>
<td>36.9</td>
<td>.816</td>
<td>.840</td>
<td>.167</td>
<td>—</td>
</tr>
<tr>
<td>CDI words</td>
<td>.311</td>
<td>36.2</td>
<td>.816</td>
<td>.843</td>
<td>.165</td>
<td>7.4</td>
</tr>
<tr>
<td>Semantic</td>
<td>.314</td>
<td>36.3</td>
<td>.809</td>
<td>.837</td>
<td>.165</td>
<td>-8.3</td>
</tr>
<tr>
<td>Phonology</td>
<td>.312</td>
<td>37.1</td>
<td>.814</td>
<td>.840</td>
<td>.167</td>
<td>-8.3</td>
</tr>
<tr>
<td>CDI label</td>
<td>.307</td>
<td>37.6</td>
<td>.820</td>
<td>.843</td>
<td>.170</td>
<td>20.5</td>
</tr>
<tr>
<td>Word2Vec</td>
<td>.312</td>
<td>37.3</td>
<td>.815</td>
<td>.841</td>
<td>.166</td>
<td>6.4</td>
</tr>
</tbody>
</table>

---

Fig. 3. Performance differences of the CDI word model compared to the CDI child model. Zero means the models perform equally. Histogram is the frequency of an individual snapshot being better fit by the CDI child model (pink) or word model (green). Data is sorted by the child’s age, then vocabulary size and far right percentile.
best aggregation model for each feature representation. Table I details what models used the child features as well as what models were averaged (as opposed to summed) in order to aggregate the individual representations of the words.

We now turn to the performance of the models we classified as being a decomposition of the vocabulary knowledge—the (3) Semantic feature norms and (4) Phonology. In Table II we see that the (4) Phonology reaches comparable performance with the child feature model even though this model has no direct information about the words in the child’s vocabulary or features of the individual learner but only information about the phonemic composition of the words in the child’s vocabulary. However, the paired t-test suggests that this model performs worse than the child feature model on a kid-by-kid measure. We also see that the neural network using the (3) Semantic feature representation is unable to outperform the child feature model. We summarize performance of these models in Table II.

Closer inspection of these models suggest the performance of the (3) Semantic feature representation is mostly due to the addition of the child features. Removing these features from the representation results in performance with much greater NLL scores than even the child feature model. In general, these results suggest that considering words, rather than their constituent parts is more useful in predicting future language acquisition trajectories.

We now consider representing the vocabulary knowledge through aggregating individual words into a higher-level representation of vocabulary knowledge. Here we consider the category labels from the CDI in our (5) CDI label model and the (6) Word2Vec representation of the child’s vocabulary. Both of these models outperform the (1) CDI child model on most measures and has similar performance as the (2) CDI word feature model. This implies some representations of a child’s vocabulary can provide additional information, beyond predictions based on the individual words a child knows. Our findings suggests that we gain improvement in predictability of individual acquisition when the model has access to category information or more general information about the words a child knows. This suggests that what is most important to predicting word learning is what categories and linguistic structure the child has in their productive vocabulary rather than the individual words the child currently knows.

Collectively, the results suggest the (5) CDI label model and the (6) Word2Vec model increase predictive capabilities of our models to accurately predict what words are likely to be learned next by specific children. We consider if this is conditional on a specific point in development or specific words in Figure II. Here we normalize such that zero indicates that the child model is the best of the language feature models we consider. We then plot the difference from zero and show the density of the best performing model as a function of vocabulary size and age. Unlike the previous plot, we do not consider each model compared to the CDI child model but instead compare all jointly. This masks the fact that some models (particularly the (5) CDI label model) outperform the (1) CDI child model much more frequently than other models, but instead highlights the statistics of the children in which the (1) CDI child model consistently outperforms other models. As before, the child model does well for children who are in general good at learning language. This suggests that the (1) CDI child model is learning only high-level trend and where there are only a few words left to learn. We also see here that the (6) Word2Vec model is particularly good for what we’d consider late-talkers and children with small vocabularies. This suggests that the Word2Vec representation may be picking up on systematicity in early language learning that is not easily interpreted by humans, a finding that requires more investigation.

D. Ensemble models

The above results help us to capture the role of a particular type of language representation in predicting future language learning. Now we explore whether the information contained in these predictions are independent or redundant. To explore this question, we construct ensemble models that weight various model predictions in hopes of training a more powerful predictive model. We now consider a few ensemble models based on the individual predictions of the 4 language representations, models (3)-(6). We also include the (1) CDI child and (2) CDI word models in our ensembles. We note that it is possible to train neural network models that include multiple representations as input, but we instead focus on averaging the final predictions.

The most basic ensemble model simply considers each of the best performing models equally. In this Avg. Ensemble model, we combine prediction across the (1) CDI child and (2) CDI word model as well as the language representation models (3)-(6). The performance of this ensemble model, as reported in Table III is comparable to the CDI child model. This ensemble does not outperform the best input representations discussed above. The second Wgt. Ensemble uses the combined training data and validation data to learn the optimal contribution of each model. This learned weighting is then applied to the test data. Table III shows that this model performs better than simply averaging all predictions together but still does not outperform our best single feature model of (5) CDI label. Looking at the weighting of the individual representations, this model suggests that the vocabulary representations that are most useful are the (5) CDI label and the (2) CDI word representations, accounting for 32% and 67% of the total estimates respectively. The (1) CDI child model, the (3) Semantic feature model and the (4) Phonology model are almost completely ignored in the optimal weighting.
The method of combining our individual neural network models affects predictive ability. Surprisingly, many of these models, especially the models that take into account general features of the child, fail to perform as well as the (2) CDI word feature model, suggesting that the information contained in these representations may be redundant or have less predictive information as compared to the CDI word model. We find that the Wgt. Ensemble model performs the best of our ensembles. This weighted ensemble model considers the (2) CDI word feature model most heavily but also weighs the (5) CDI label representation. These results suggest that there may still be unique information in the some of representations that could aid in predictions of individual word acquisition of unseen children but that due to limitations in data availability or model architecture these ensembles are not able to easily capitalize on this information. The similarity between the weighted ensemble model and the average ensemble model suggests that the benefit of these various representations can be accessed nearly as simply by averaging predictions of all models as opposed to optimizing current model weights.

### VI. Conclusions and Discussion

We find evidence that developmental changes and learning have an impact on which words a child will learn next. We considered this by capturing differences between a linguistically informed model and a child feature model when sorting by certain features we think are relevant to language learning in young children. We also considered child-based ensemble models that weigh the similarity of individual children on different features, as well as ensembles that consider features of the learner in order to generalize to unseen children. These results suggest that CDI percentile is the most relevant feature in predicting language learning and that certain models perform best on a particular subset of children.

Individual words in a child’s vocabulary are informative in predicting future vocabulary growth. The (2) CDI word model with just the produced word information reliably outperformed the (1) CDI child feature model. This confirms our intuition that the individual words a child knows contains relevant information beyond that provided by knowing the child’s age and vocabulary size. This is an interesting result when placed in the context of current diagnostic and intervention techniques in developmental psychology. Many vocabulary assessment tools rely only on information pertaining to the size of the child’s vocabulary, with little attention to the specific words known by the individual learner. These results suggest that we can improve our assessment of children’s development by looking at the individual items in a child’s productive vocabulary. The success of the (5) CDI label model also suggests that the category structure of a child’s vocabulary may be important to understanding their language learning ability.

These modeling results additionally suggest the need to consider differences in learners. The content of the vocabulary significantly improves our ability to predict future acquisition, suggesting that an individual’s vocabulary has relevant and predictive information about the type of learner — and the learning trajectory — of a particular child. While we remain agnostic as to the nature of the relationship between known words and future learning, we find strong evidence of the importance of the current vocabulary both in the (2) CDI word feature model and the (5) CDI label model. However, (3) Semantic features (capturing semantics) and the (4) Phonology model performed significantly worse than the (1) CDI child model possibly because these representations do not aggregate the child’s current vocabulary knowledge in a meaningful way. In later work it may be interesting to consider why these models fail. For example, the chosen phonological representation may fail to capture features relevant to young learners, such as phonemic onset, rhyme, sound similarity, or the difficulty of pronouncing individual phonemes [12], [13], [16].

More interesting than the failure of individual representations is the feature representations that perform on par with, or better than, the individual word representations. The feature aggregation using the (5) CDI label or word class labels is reliably the best performing model. This suggests knowing something about the category of words a child knows can help in predicting acquisition of *individual words*. This representation is possibly capturing important features in linguistic knowledge of young learners. The (6) Word2Vec model, which did not include the child features in the neural network representation, performs on par with the (1) CDI child model, suggesting that representing the child’s vocabulary knowledge as various features that are themselves difficult to interpret still allows for the model to learn. Future work should consider how this (6) Word2Vec representation could be tailored to capture information that may be more relevant to our young learners—for example instead of training on the GoogleNews corpus, we could train on children’s books or child-directed speech.

The performance of these aggregated knowledge representations begs further investigation— is the success of these representations the result of the model’s ability to capture different learning styles which allow for easy detection of a learner’s trajectory? Or are these representations abstracting vocabulary content in a way that represents language knowledge in a more useful way? Capturing the learning trajectory may be the most reliable prediction of future growth, allowing our models to accurately predict future acquisition and provide additional insight into important differences in learning trajectories. Classification of these trajectories and different learning styles may also be possible. These aggregated lexical knowledge models perform especially well for a particular group of children commonly known as *late talkers*, children who know fewer words than their age-matched peers, raising important future directions in the diagnosis and intervention design for children with language learning difficulties. In future work, we plan to use these representations to model the *language trajectory* as opposed to individual word learning. Previous work has suggested that children with lower CDI percentile have more variance in the words they learn than.

### TABLE III

<table>
<thead>
<tr>
<th>Model</th>
<th>NLL</th>
<th>% overlap</th>
<th>AUC</th>
<th>acc</th>
<th>d prime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. Ens.</td>
<td>.307</td>
<td>36.1</td>
<td>.820</td>
<td>.843</td>
<td>.171</td>
</tr>
<tr>
<td>Wgt. Ens.</td>
<td>.306</td>
<td>36.5</td>
<td>.821</td>
<td>.843</td>
<td>.171</td>
</tr>
<tr>
<td>Word Vote</td>
<td>.310</td>
<td>36.7</td>
<td>.817</td>
<td>.842</td>
<td>.168</td>
</tr>
</tbody>
</table>

*IEEE Transactions on Cognitive and Developmental Systems*
children who have higher CDI percentiles [3], [6], [14], [34]. An implication of these results is that children who are having more difficulty with language have widely variable learning strategies. Given this idea, the success of some of the models for this population of late talkers is hopeful. It offers a model to predict future word learning and may also suggest what types of attentional mechanisms may differ between late talkers and their peers. Our work shows that we can quantify differences in the vocabulary of these children in ways that aid in prediction of future language learning. We hope to use this insight to explore the attentional and learning mechanisms that result in learning differences between these groups in future work.

By considering the developmental aspects inherent in this type of modeling, we can make predictions (and evaluate those predictions) of how a specific child’s vocabulary will grow. This type of modeling approach will allow us to capture and explain the effect of certain features in language learning as related to development, and in turn, allow us to distinguish late talking children who catch up to their peers from those late talking children who do not, allowing for targeted and early interventions. Given that the neural networks are able to predict future acquisition with some degree of accuracy, we can begin to predict further than one month, assessing language ability throughout the course of early development. We can also use these different language representations to further tease apart different types of learners and the acquisition process of late- and typically-developing children.

It is still an open question whether the performance of these models can be increased with more data — which is time intensive and challenging to collect. If instead we can use insights from machine learning to direct researchers to specific lexical features of relevance, we may improve the ability of developmental psychology to expand their understanding of learning without having to do exploratory data-intensive investigations.

REFERENCES