This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

Quantifying the role of vocabulary knowledge in predicting future word learning

Nicole M. Beckage, Michael C. Mozer, and Eliana Colunga

***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 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 an individual child is likely to learn next. We consider a neural network approach to predict 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 likely to be learned by a particular child in the coming month. Many models based on child-specific vocabulary information outperform models with child information only, suggesting that the words a child knows**

influence prediction of future language learning. These models provide an understanding of the role of current vocabulary knowledge on 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 specific child is likely to learn next. Alternatively, if the idiosyncratic words a child knows at a given time influence the words that child is going to learn next, this would provide strong evidence that 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 learned itself is a side-effect 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 encapsulates. 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 new words in the future.

In this work we explore how various representations of a child's current vocabulary predict that same child's future lexical acquisition. We focus on high level features of the child and the child's lexicon specifically to zoom in on what the role of lexical structure is on future lexical acquisition. We acknowledge that there are many forces that could influence vocabulary growth and language acquisition besides the composition of a child's lexicon. Here we focus on the ability to use language knowledge to predict future lexical growth. We consider various vector representations that aim to capture a child's language knowledge and evaluate these representations on their ability to predict future acquisition trajectories *at the level of an individual child* to model lexical growth. We compare the usefulness of different vector 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, which is the focus of this modeling work. 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, and how the representation of that knowledge influences the ability to predict future vocabulary growth. We argue that these analyses can be informative in suggesting the relative importance of different factors in word learning, leading to specific predictions that could be tested empirically.

Currently, a toddler's lexical knowledge is often measured as the number of words they know, given the child's age and sex [8], [10], [33]. While there is strong evidence that this count of the number of words is useful in assessing language ability, it is unclear that this number alone is useful and

N. Beckage is with the Department of Psychology at the University of Wisconsin, Madison WI, 53706 USA e-mail: nicolebeckage@gmail.com.

Michael Mozer and Eliana Colunga are with University of Colorado Boulder.

Manuscript received August 31, 2017;

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

informative enough to aid in predicting future acquisition. In our age range of the individual children of interest, the most commonly used measure of vocabulary size 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 [15], [33]. Sometimes these early language difficulties will persist and be reflected in reading and other difficulties in academic settings [11]. However, not all children who have a low CDI percentile as toddlers go on to have lasting language difficulties, and to date, it is impossible to accurately predict which toddlers will have persistent difficulties and which will catch up. By exploring different types of language representations in predicting future acquisition, we may help uncover relationships between current language and future learning that could help with diagnostic assessment, providing a quantitative tool to help distinguish children who are simply learning language at a slower rate than their peers and those children who are at risk for these language-specific delays to manifest as other types of cognitive difficulties, such as specific language impairment (for more information on SLI see [18]). The relationship between lexical representation and CDI percentile might suggest an approach for quantifying meaningful differences in word 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 as an initial attempt to tackle these issues.

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 accurately 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. By understanding the influence of the representation of a childs linguistic knowledge in our models ability to predict future acquisition, we aim to isolate the effects of the role of lexical knowledge on the learning of individual lexical items. In the larger literature of neural networks, neural networks are essentially 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, allowing us to scientifically investigate the effect of the representation of lexical knowledge on the ability of these models to predict future acquisition of lexical items for individual children. 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 for individual children.

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, have 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], [21] and a specific review of semantic development here [29]. Previous research also explores neural network approaches to link neuroscience to early development [27] and to semantic cognition [22]. 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 behavioral 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 of CDI words, acquires useful word learning biases, even though the models are not directly rewarded for this bias, suggesting that the model benefits from learning these biases and using them to generalize to novel word-to-object mappings. 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 experimentally verified in young children [5]–[7], [30].

Other examples of neural network models applied to modeling early lexical acquisition include models capable of capturing word confusability and age of acquisition effects [19], and the formation and degradation of conceptual categories [22]. These, and other examples of lexical development, explain behavior with basic mechanistic accounts of associative learning. One

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

neural network model, which learns to map word-forms to object referents [23] shows a mutual exclusivity bias—a preference for novel words to map to novel objects [25], 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 model captures the acquisition of categories. Interpreting the online learning of the neural network, the authors provide evidence for a feedback loop between perceptual features and linguistics labels [38]; the linguistic labels are thought to support generalization of categories and thus facilitate learning. The model itself is able to capitalize on the relationships between category formation and language learning to provide structure and reinforcement, via the feedback loop between perceptual and linguistic features, 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 facilitate 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 under-determined or noisy. We leverage the robust learning of neural networks to provide quantification of the predictive power of different vocabulary representations in relation to future lexical acquisition.

A key assumption to this type of data-driven neural network model of acquisition is that there are regularities in the way in which children learn. But the differences are also informative and predictive. If all children learn similarly, and/or 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 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 of learners [15], [20], [28], [33]. 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 [3], and in the lab, late talkers seem to learn new words differently than their typically developing peers [6], [36]. 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 note that a large body of work focuses on the aspects of the language and linguistic environment that affects language acquisition. Here, we instead assume that the content of the child's vocabulary embodies much of the relevant 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 thus infer that models with higher accuracy are capitalizing on representations or aggregation methods that accentuate those aspects of the child's language or characteristics that are relevant to their acquisition process. We consider the performance of various language representations to the baseline child-feature model to approximately quantify the influence of certain language representations on predictions, and thus as a proxy for the relevance of this type of linguistic information on the acquisition process.

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 only 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 one month in the future. Alternatively, if the lexical items in the vocabulary of a child captures predictive information that influences future acquisition such as the child's interests in specific themes (for example, animals) or their language environment, then knowing the semantic content of the child's vocabulary will be helpful in predicting 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 at approximately monthly intervals for a year. On average, children in our study had 10.9 visits. 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 reports which form 825 CDI vocabulary *snapshots* (i.e., two sequential vocabulary reports). We define a CDI *snapshot* as a sequential set of CDI's where the first CDI is the (transformed) input to the neural network and the second CDI is the output (target) 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

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

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 when their first CDI was collected 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.

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

Fig. 1: Example of longitudinal CDI data used as untransformed input and output of the neural network. Note that only the productive knowledge of the individual words is the output the neural network models.

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. 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 (α), number of hidden units (hu), batch size, number of epochs until learning rate is effectively zero (α 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 decent 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 not penalized, nor are the weights updated, for incorrect predictions on words that are already known by the child.

Most of the step-wise optimization procedure 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 to 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 vocabulary representations 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? Different 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. The performance of our models with various definitions of lexical knowledge may provide insight into the types of information that is guiding language acquisition. We use two broad types of representations to capture a child's lexicon. First, we explore representing vocabulary knowledge by *decomposing* individual words into lower level units, for example breaking down the sounds of the words to capture phonemic level information. 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:

1) *CDI child* feature model based demographic information of the child,
2) *CDI word* model based on the CDI vocabulary report of a child's productive vocabulary,
3) *Semantic* model based on the semantic features of particular words in the child's vocabulary based on the McRae feature norms [24],
4) *Phonology* model which considers the child's phonological composition of their productive vocabulary,
5) *CDI label* model which captures the production of words within particular categories as labeled on the CDI, and
6) *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 representations below.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

TABLE I: Hyper-parameter for neural network models. (α) is the learning rate, hu is hidden units, $\Delta\alpha \approx 0$ is epochs when α is nearly 0, (m) is momentum and $avg?$ is whether word features were averaged. Input +6 shows inclusions of CDI child features.

model	input	α	hu	batch	$\Delta\alpha \approx 0$	m	$avg?$
CDI child	6	0.3	800	25	500	.7	
CDI word	677+6	0.8	500	25	200	.9	
Semantic	30+6	0.7	300	50	500	.5	T
Phonology	37	0.5	200	10	400	.7	F
CDI label	22+6	0.8	500	25	200	.7	T
Word2Vec	200	0.2	500	25	650	.7	F

A. Lexical knowledge and input representations

In our first simulation 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) CDI 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 standard information researchers usually use to assess a child's lexical knowledge and approximate their language ability.

We create the *(2) CDI 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 but learns via training on other snapshots the general trend of the order in which words are acquired, 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 word-level model may perform less accurately than a model that clusters words based on semantic 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 two classes of representations that 1) break down the words into specific features and 2) 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 features or more high-level abstract information about a child's lexicon. This may help direct future developmental research focused on understanding the role of different kinds of linguistic features that may influence early lexical 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 [24] 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 word features such as word length, binary vector representing whether a word has the feature, 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*. Because a child may have multiple words that share a specific feature, we aggregate together all of the individual McRae feature vectors for all words that the child knows in order to represent the child's vocabulary knowledge. Previous work has found the McRae representation has minimal predictability in accounting for acquisition of young children using network analysis [16]. Here we test the usefulness of this representation within a neural network model.

We then consider the phonemic composition of individual words. Past work shows the sounds of words play a significant role in learning [31], [32] and that computational models can capture this effect [34]. 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/or the most common transcription. We took an approach of broad transcription, ignoring subtle and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

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 (CDI) vocabulary of the child. Research related to phonological importance in early learning suggests there is a strong effect of word onset and word rhyme [14] but other work has instead suggested phonemic awareness is a better predictor [17]. 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" 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 the language input the child receives from the parent [35], 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 as well as or better than the CDI child model and the other models we consider. 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 [26]. We explore the use of this representation to capture child language acquisition. Using the Word2Vec algorithm, which considers co-occurrence frequency of words and the neighborhood of the word in text/speech, we constructed a 200 dimensional vector representation of nearly all words in the CDI using a Word2Vec representation trained on a large GoogleNews corpus. We assume this is an aggregate representation rather than a decomposition because Word2Vec requires co-occurrence information as well as information about those word's nearby contexts, resulting in words that have both 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 vector representation [26]. 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 holistic 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 since larger vocabularies will have more instances of each feature (e.g. the phoneme ɪ is more frequent in larger vocabularies). Beyond the method of aggregation, we also consider whether we see an improvement in predictability when we include the child specific features of the CDI 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 (llk) 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, which are chosen via cross-validation, do we consider the withheld test set. The test set includes unseen children and all their respective CDI snapshots.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

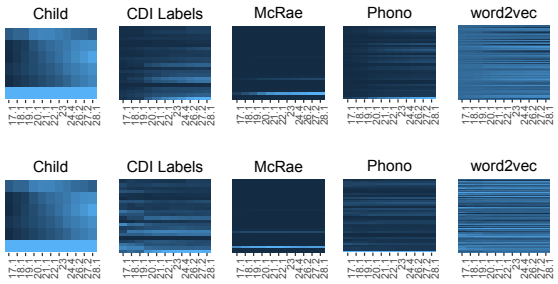

Fig. 2: Input representations to the neural networks with different methods of aggregation for a child. Top row represent summing the individual word vectors in the child's vocabulary, bottom row indicates averaging individual word specific vectors. The x-axis indicates CDI time points. Lighter color indicates higher activation.

We evaluate performance by averaging the negative log-likelihood (llk) 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' words that are predicted as most likely to be learned 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 learned and unlearned varies. To capture model performance in relation the ROC curve, we present the area under the ROC curve (AUC) as well as summary measures of accuracy and discriminability (d-prime). We assume for both accuracy and d-prime the threshold is the point in which learning events are predicted with equal frequency to what is observe within the particular CDI learning snapshot. Also included is the t-statistic from a paired t-test on average llk of the specific model compared to the CDI child model for the unseen CDI snapshots in the test set.

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 neural 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 [24] 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 [12], 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 [20].

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 (llk) 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 llk scores for comparison. If the model always returned 0.5 as the probability of learning a word, the average llk 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 llk 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 llk of 0.456 and 0.453 respectively. Values closer to zero indicate better model performance.

Our final (informed) baseline llk 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. With this result we can now compare 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 other than to remind the reader that we did step-wise optimization for each free model parameter and for each model individually. We first consider the CDI-based models and then we turn to the feature based models.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

B. CDI models

As discussed in section IV-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 negative log-likelihood (llk) for all 171 snapshots (17 children) in the test data. Table II shows that the *CDI word* representation performs better than the CDI child model in log-likelihood, and accuracy which is near 84.3%, suggesting that knowing the individual words a child knows increases overall performance of our predictive models. However, looking at aggregate performance, it is difficult to capture where the gains of the word-based model are. For example, it is possible that this model is a better model for every child in the testing data set. Alternatively the CDI word model may see gains for a sub-population of learners or snapshots capture specific populations of learners, performing better for a subset of the snapshots explored. To examine this idea, we turn to figure 3. In this figure, 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. From left to right, we order snapshots by the age of the child, 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. We include a density plot (right of the scatter plot) to indicate the number of snapshots that are best fit by each model for the range of child features we consider.

TABLE II: Neural network performance of 6 different models.

model	llk	% overlap	AUC	acc	d prime	t-stat.
CDI child	.312	36.9	.816	.840	.167	—
CDI words	.311	36.2	.816	.843	.167	7.4
Semantic	.314	36.3	.809	.837	.165	-7.8
Phonology	.312	37.1	.814	.840	.167	-8.3
CDI label	.307	37.6	.820	.843	.170	20.5
Word2Vec	.312	37.3	.815	.841	.166	6.4

We find that the *(1) CDI Child* model performs on par with the CDI model for children with higher percentiles but not as well for children with small percentiles. This result is in line with other work that finds higher variability and more heterogeneity in the lexicon of children with lower CDI percentiles e.g. [10]. In practice this means our CDI word specific model has the greatest improvement over the baseline CDI child model for children who have a CDI lexicon that is smaller than average. The fact that the CDI word model shows strongest gains for a population that is known to be variable in their learning strategies suggest that this data driven approach can leverage meaningful trends to predict future acquisition. In future work we aim to explore what these trends are and if they can provide direct insight into different strategies or environments that might impact a toddler's language ability.

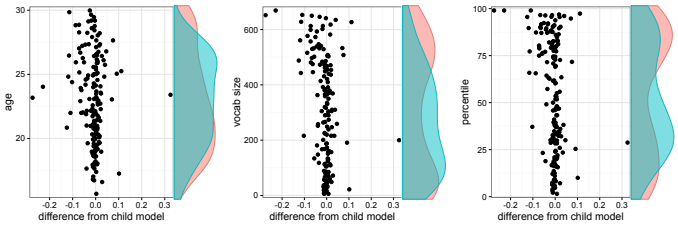

Fig. 3: Performance differences of the CDI word model compared to the CDI child model. Zero means the models perform equally. Histogram is the frequency of an individual snapshot being better fit by the CDI child model (pink) or word model (green). Data is sorted by the child's age, then vocabulary size, and percentile.

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 *(3) Semantic* feature norms and the *(5) CDI label*) 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.

The fact that there are some representations that do not benefit from the inclusion of child level features may suggest that there are a subset of words that are learned systematically [6], [15] which can be used to mark development and other child specific features, resulting in a redundancy between vocabulary representations and 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. We find that when the features are averaged, child features increase accuracy further highlighting the fact that summing the features preserves information about the child's age and potentially about their language ability as vector 'activity' increases as vocabularies grow. Table I details what models used the child features as well as what models were averaged (as opposed to summed) in

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

order to aggregate the individual representations of the words. For the rest of our analysis we choose the best aggregation model for each vector representation of the lexicon.

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 as indicated by a large negative t-statistic. The decrease in predictive performance of the phonemic model when compared to the CDI word model suggests that children are not only learning words based on their ability to pronounce and parse the phonemes. While it is a necessary condition for children to understand the word they are learning to produce, phonemic production is not enough to predict what words a child will learn next. While not a surprising result, the performance loss of this representation compared to that of the CDI word model validates the ability of model performance to provide an indirect means to quantify the usefulness of various representations in predicting future lexical learning of individual children.

We also see that the neural network using the *(3) Semantic* feature representation is unable to outperform the child feature model. Previous work has found that the feature norms themselves do not adequately capture the relevant features to small children (e.g. [16]). This result is not unexpected as the McRae feature norms include features, such as encyclopedic, that are learned much later in development [24]. We summarize performance of these decompositional models in Table II. 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 see table II. In fact these models tend to outperform the CDI word model, implying 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. The improvement of these aggregated vocabulary representations implies that what is most important to predicting future 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

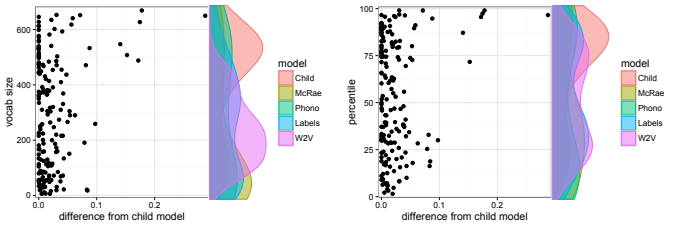

Fig. 4: We consider difference of performance of the CDI word model as compared to the CDI child model. Zero means the model perform equally well. The histogram represents the frequency of an individual being better fit by the CDI child model (pink) or the CDI word model (green). From left to right, and along the diagonal axis, the data is sorted by vocabulary size and percentile

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 4. 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 the high-level trend is predictive. We also see here that the *(6) Word2Vec* model and the *(2) CDI word* model are particularly good for children with low percentiles and children with small vocabularies. This suggests that the Word2Vec representation, or the individual words on the CDI may be capitalizing 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 weigh 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*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCDS.2019.2928023, IEEE Transactions on Cognitive and Developmental Systems

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 fact that this ensemble model does not outperform the CDI child model suggests that the types of information each representation is using to predict are not equally relevant and the success of certain models is effectively canceled out by the poor performance of other models. 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. Even though this ensemble model is not more predictive than the best performing single feature model, the weighting of each model indicates what features of language learning are most predictive of future lexical acquisition. The fact that the CDI label model and the CDI word model are significant contributors to the final predictions suggests that knowing the child's vocabulary as well as higher level category information is useful in predicting future acquisition. We believe that this ensemble model could be made more powerful with more data and suggest that researchers interested in modeling acquisition consider both category information and word level information.

TABLE III: Performance of ensemble models on test data

model	llk	% overlap	AUC	acc	d prime
Avg. Ens.	.307	36.1	.820	.843	.169
Wgt. Ens.	.306	36.5	.821	.843	.171
Word Vote	.310	36.7	.817	.842	.168

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. In the future we aim to further investigate how to integrate these different representations to build a more accurate and robust predictive model of lexical acquisition.

VI. Conclusions and discussion

We find evidence that developmental changes as captured by child level features and the individual words a child knows now have an impact on which words a child will learn next. Individual words in a child's vocabulary are informative in predicting future vocabulary growth. The *(2) CDI word* model, which contains the words produced by the child 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 clinical practice. 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. The predictive accuracy of our network models 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 [13], [14], [17].

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 the lexical 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 and that these vector representations can capture all of the information available in the high-level information about the learner even though the information available to the neural network excludes this infoormation. 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 from the perspective of a toddler 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 children who have higher CDI percentiles [3], [6], [15], [37]. 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, might allow us to distinguish late talking children who will catch up to their peers from those late talking children who will 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

[1] LW Barsalou, WK Simmons, AK Barbey, and CD Wilson. Grounding conceptual knowledge in modality-specific systems. *Trends in cognitive sciences*, 7(2):84–91, 2003.

[2] N M Beckage, M Mozer, and E Colunga. Predicting a child's trajectory of lexical acquisition. *Proc. of the 37th Conf. of the Cog. Sci. Society*, 2015.

[3] NM Beckage, LB Smith, and TT Hills. Small worlds and semantic network growth in typical and late talkers. *PloS one*, 6(5):e19348, 2011.

[4] MH Christiansen and N Chater. Connectionist psycholinguistics: Capturing the empirical data. *Trends in Cognitive Sciences*, 5(2):82–88, 2001.

[5] E Colunga and C E Sims. Early-talker and late-talker toddlers and networks show different word learning biases. In *Proceedings of the 34th Annual CogSci Conference*, pages 246–251, 2012.

[6] E Colunga and CE Sims. Not only size matters: Early-talker and late-talker vocabularies support different word-learning biases in babies and networks. *Cognitive science*, 41(S1):73–95, 2017.

[7] E Colunga and L B Smith. From the lexicon to expectations about kinds: a role for associative learning. *Psychological review*, 112(2):347, 2005.

[8] PS Dale and L Fenson. Lexical development norms for young children. *Behavior Research Methods, Instruments & Computers*, 28:125–7, 1996.

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

[10] L Fenson, PS Dale, JS Reznick, E Bates, and et. al. Variability in early communicative development. *Monographs of the society for research in child development*, 59(5):1–185, 1994.

[11] Rebecca R Fewell and Barbara Deutscher. Contributions of early language and maternal facilitation variables to later language and reading abilities. *Journal of Early Intervention*, 26(2):132–145, 2004.

[12] D Gentner. Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In *Language development: Language cognition and culture.* 1982.

[13] JA Gierut, ML Morrisette, MT Hughes, and S Rowland. Phonological treatment efficacy and developmental norms. *Language, Speech, and Hearing Services in Schools*, 27(3):215–230, 1996.

[14] U Goswami and P Bryant. *Phonological skills and learning to read*. Wiley Online Library, 1990.

[15] J Heilmann, S E Weismer, J Evans, and C Hollar. Utility of the macarthur-bates communicative development inventory in identifying language abilities of late-talking and typically developing toddlers. *American Journal of Speech-Language Pathology*, 14(1):40–51, 2005.

[16] T Hills, M Maouene, J Maouene, A Sheya, and L B Smith. Longitudinal analysis of early semantic networks: preferential attachment or preferential acquisition? *Psychological Science*, 20(6):729–39, 2009b.

[17] C Hulme, P J Hatcher, K Nation, and et. al. Phoneme awareness is a better predictor of early reading skill than onset-rime awareness. *Journal of experimental child psychology*, 82(1):2–28, 2002.

[18] Laurence B Leonard. *Children with specific language impairment*. MIT press, 2014.

[19] P Li, I Farkas, and B MacWhinney. Early lexical development in a self-organizing neural network. *Neural networks*, 17(8):1345–1362, 2004.

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

[21] J L McClelland, M M Botvinick, D C Noelle, D C Plaut, T T Rogers, M S Seidenberg, and L B Smith. Letting structure emerge: connectionist and dynamical systems approaches to cognition. *Trends in cognitive sciences*, 14(8):348–356, 2010.

[22] J L McClelland and T T Rogers. The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, 4(4):310–322, 2003.

[23] B McMurray, J S Horst, and L K Samuelson. Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological review*, 119(4):831, 2012.

[24] K. McRae, GS Cree, MS Seidenberg, and C McNorgan. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–59, 2005.

[25] William E Merriman, Laura L Bowman, and Brian MacWhinney. The mutual exclusivity bias in children's word learning. *Monographs of the society for research in child development*, pages i–129, 1989.

[26] T Mikolov, I Sutskever, K Chen, G S Corrado, and J Dean. Distributed representations of words and phrases and their compositionality. In *NIPS conference*, pages 3111–3119, 2013.

[27] Y Munakata and J M Stedron. Neural network models of cognitive development. In *Handbook of developmental cognitive neuroscience*, pages 159–172. MIT Press, 2001.

[28] CM Sandhofer, LB Smith, and J Luo. Counting nouns and verbs in the input: differential frequencies, different kinds of learning? *Journal of Child Language*, 27(3):561–85, 2000.

[29] CE Sims, SM Schilling, and E Colunga. Interactions in the development of skilled word learning in neural networks and toddlers. In *2012 IEEE ICDL*, pages 1–6. IEEE, 2012.

[30] CE Sims, SM Schilling, and E Colunga. Beyond modeling abstractions: learning nouns over developmental time in atypical populations and individuals. *Frontiers in psychology*, 4, 2013.

[31] SF Stokes, T Klee, CP Carson, and D Carson. A phonemic implicational feature hierarchy of phonological contrasts for english-speaking children. *Journal of Speech, Language, and Hearing Research*, 48:817–33, 2005.

[32] HL Storkel. Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. *Applied Psycholinguistics*, 25(02):201–221, 2004.

[33] DJ Thal, L O'Hanlon, M Clemmons, and L Fralin. Validity 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, 1999.

[34] M S Vitevitch and H L Storkel. Examining the acquisition of phonological word forms with computational experiments. *Language and speech*, 56(4):493–527, 2013.

[35] S R Waxman and E M Leddon. Early word learning and conceptual development: Everything had a name, and each name gave birth to a new thought. *Blackwell handbook of childhood cognitive development*, pages 102–126, 2002.

[36] Susan Ellis Weismer, Courtney E Venker, Julia L Evans, and Maura Jones Moyle. Fast mapping in late-talking toddlers. *Applied Psycholinguistics*, 34(1):69–89, 2013.

[37] ZO Weizman and CE Snow. Lexical output as related to children's vocabulary acquisition: Effects of sophisticated exposure and support for meaning. *Developmental Psychology*, 37:265–279, 2001.

[38] C Yu, D H Ballard, and R N Aslin. The role of embodied intention in early lexical acquisition. *Cognitive science*, 29(6):961–1005, 2005.