Predictive modeling to capture the words a toddler will learn next

by

Nicole M. Beckage

B.S., B.A., Indiana University, 2010

M.A., University of California Irvine, 2012

M.S., University of Colorado Boulder 2015

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

Department of Computer Science

2016

This thesis entitled:
Predictive modeling to capture the words a toddler will learn next
written by Nicole M. Beckage
has been approved for the Department of Computer Science

_____
Prof. Eliana Colunga

_____
Prof. Michael Mozer

_____
Prof. Aaron Clauset

_____
Prof. Matt Jones

_____
Prof. Tammy Sumner

Date _____

The final copy of this thesis has been examined by the signatories, and we find that both the content and the form meet acceptable presentation standards of scholarly work in the above mentioned discipline.

Beckage, Nicole M. (Ph.D., Computer Science and Cognitive Science)

Predictive modeling to capture the words a toddler will learn next

Thesis directed by Prof. Eliana Colunga

Network models of language provide a systematic way of linking a child's current vocabulary knowledge processes to the structure and connectivity of properties of language which promote future lexical learning. Using network growth models, we explore the relational role of language and the influence of linguistic structure on language learning. Previous research has proposed that language is learned by a process of semantic differentiation that can be modeled through a network process of preferential attachment, with highly connected nodes being learned earliest. This model accounts for high-level lexical network structure and also captures empirical age of acquisition reports. Alternately, language learning may be driven by contextual diversity, or the diverse contexts and meanings of unknown words in the environment. In this thesis, we test these and other ideas by extending these models to acquisition trajectories of individual children, predicting the individual words a child is likely to learn next. We explore how the definition of a graph, the assumed network growth process, and measures of node importance affect our ability to model acquisition. We not only construct a theoretical framework for network models of acquisition but also test the ability of these models to account for learning and development. This work suggests that network models provide a framework for understanding the cognitive and developmental processes of language acquisition.

Neural network models, often called connectionist models, offer another independent approach to modeling learning and development. We focus on associations in a child's current vocabulary that might be relevant and even facilitatory to the learning process of young children by constructing predictive models. The associative learning framework of our neural network models allow for different types and timescales of learning to be captured. A key idea to data-driven neural network models of acquisition is that there are strong similarities among the way in which children learn,

but the differences between children are also predictive. Assuming that there are different types of language learners and that the vocabulary (together with child age) at any time point reflects the type of learner a particular child is, machine learning models can provide a powerful and predictive tool to aid with classification and diagnostics of a child's learning trajectory. Focusing specifically on using a child's vocabulary to predict future lexical learning. We explore a variety of representations of a child's current vocabulary knowledge, including those from a productive vocabulary report as well as representations based on natural language processing algorithms, adult norms, and phonemic content. We find that individual words in a child's vocabulary are informative in predicting future vocabulary growth using a neural network model. These results additionally suggest the need to consider differences amongst learners. Our best performing model has information not only about a child's own vocabulary knowledge but also about the normative acquisition trends of words in that child's vocabulary. These two types of information improve predictive accuracy and suggest potential diagnostic and interventional tools for helping bridge the lexical differences of language delayed children and their age-matched peers.

Acknowledgements

like to thank my advisors, Eliana Colunga and Mike Mozer, for all the helpful discussions, support, and patience over the years. I would like to thank my committee members, Aaron Clauset, Matt Jones and Tammy Sumner, for their feedback, suggestions, and especially for their time. For their support, guidance, and kindness, I would like to thank my mother, Cheryl Beckage, my father, Michael Beckage, and my partner Matthew Moore.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

The Developing Lexicon

How does a child's current vocabulary inform and relate to their vocabulary in the future? By studying the trajectory of language acquisition, we can begin to understand how children's early lexicon provides support and scaffolding for future language acquisition. Children, during the course of early lexical acquisition, are not only learning the meaning of individual words, but they are also learning about the structure of language and meaning (Seidenberg & McClelland, 1989; Bloom, 2002; L. B. Smith, 2000; L. B. Smith, Jones, Landau, Gershkoff-Stowe, & Samuelson, 2002). Differences in early learning can reflect developmental variation across children and different learning strategies (Sandhofer, Smith, & Luo, 2000), environmental factors (Marchman & Fernald, 2008), or individual interests of young learners (DeLoache, Simcock, & Macari, 2007). These learning differences may result in deficits in a child's early lexicon. These deficits are a predictor of future language difficulty (Dale, Price, Bishop, & Plomin, 2003; Marchman & Fernald, 2008; Storkel, 2009). Potentially, if researchers can predict words the child is ready and able to learn, early learning can be better understood. If models are capable of predicting words a child is likely to learn next, it is also possible that deficits in learning can be detected, and maybe even corrected. However, reliable prediction can be made only if word learning develops in a systematic way.

There is strong evidence from literature on language acquisition indicating words are learned systematically. Perhaps not surprisingly, for example, children's vocabulary is related to their parent's vocabulary (e.g. Weizman & Snow, 2001; Van Veen, Evers-Vermeul, Sanders, & Van den Bergh, 2009). That is, the child will learn the words in his or her environment. In addition, some

concepts, and therefore the words that name them, may be easier to learn than others. For example, concrete nouns are learned earlier than verbs and adjectives (e.g. Gentner, 1982; Sandhofer et al., 2000). Further, the child may bring some preferences and constraints to the task of word learning. For instance, a child may become particularly interested in dinosaurs, construction equipment, or tea sets (DeLoache et al., 2007). A child might also differ in the speed of word recognition indicating differences in cognitive ability (Marchman & Fernald, 2008). In characterizing the forces that systematically influence word learning, there is evidence that at least three distinct, but not necessarily mutually exclusive, sources of information may affect the learning of individual children. These forces are a) the structure and composition of the linguistic environment, b) the structure of the concepts and categories being named, and c) the characteristics of the learner herself. There are also interactions among these influences (C. B. Smith, Adamson, & Bakeman, 1988) and among the lexical items in a child's vocabulary (Beckage & Colunga, 2013).

With this dissertation, we intend to examine the systematicity present in word learning. Specifically, we examine the relationship between current lexical knowledge, the process of acquisition, and future vocabulary knowledge. Children do not learn new words in a vacuum. They are also learning about concepts and objects in the world around them and how words map to these concepts. Importantly, they are using the statistical regularities in the world around them to bootstrap concepts and learn new object-to-meaning mappings. We try to capture the learning process through predictive models of word learning. Specifically we focus on the use of network analysis and neural networks for predicting future lexical acquisition, modeling individual lexical trajectories of children between the ages of 16 and 36 months old. We consider various ways to represent the child's language knowledge to maximize our ability to predict what a child will learn next. These vocabulary representations are a proxy for attentional and preference differences in the child as well as the structure of language and the learning environment.

Previous theoretical and experimental work provides a framework for our predictive modeling. While we review specifics of each approach later, much work has previously been done on the use of network representations (for review see Borge-Holthoefer & Arenas, 2010; Baronchelli, Ferrer-

I-Cancho, Pastor-Satorras, Chater, & Christiansen, 2013; Beckage & Colunga, 2015) and neural network models of language (for review see Christiansen & Chater, 2001; McClelland et al., 2010; Sims, Schilling, & Colunga, 2013a). Network analysis has been extremely useful in uncovering systematic structure of language that extends across languages (Motter, de Moura, Lai, & Dasgupta, 2002; Solé, Corominas-Murtra, Valverde, & Steels, 2010; Liu & Li, 2010) and linguistic evolution (Dorogovtsev & Mendes, 2001; Ferrer i Cancho & Solé, 2001; Baxter, Blythe, Croft, & McKane, 2006). More important to our work on acquisition, network models have been used as tests for cognitive theories related to language acquisition (e.g. Steyvers & Tenenbaum, 2005; Ke & Yao, 2008; Vitevitch, 2008; Hills, Maouene, Maouene, Sheya, & Smith, 2009b). Representing language as a network can capture differences in children across the course of development (Ke & Yao, 2008) or differences of children who learn language at slower rates than their peers (Beckage, Smith, & Hills, 2011). Network approaches have also offered explanations for biases of learning such as frequency and age of acquisition effects (Steyvers & Tenenbaum, 2005; Griffiths, Steyvers, & Firl, 2007).

Neural network models also provide insight into language acquisition in young children. These models have explained behavioral data of young learners, as a test of the mechanisms that drive language learning, and as support for various philosophical debates within the language learning community (Christiansen & Chater, 2001; McClelland et al., 2010; Sims et al., 2013a). We use neural networks simply as predictive tools, but in the future we hope to extend these predictive models to explore cognitive theories directly. For example, early in word learning children are shown to gradually acquire a shape or material bias (Landau, Smith, & Jones, 1988). Neural network models have been used to categorize word learning trajectories as showing evidence of the shape bias (Colunga & Smith, 2005, 2004). These models also capture the **emergence** of this bias in young learners (Sims, Schilling, & Colunga, 2012; Sims et al., 2013a). A first step to modeling the process of lexical acquisition requires understanding of how the content of the current vocabulary directs and influences future language learning in predictable and informative ways. We therefore endeavor to build predictive models of what words a child will learn next given information about the words the child knows now.

1.1 Measuring the developing lexicon

To characterize and capture the systematicity present in word learning, we need detailed data of the words that a child learns through the course of development. One well-established way to measure and characterize toddlers' lexicons is vocabulary checklists, such as the MacArthur-Bates Communicative Development Inventory (**CDI**; Fenson et al., 1994). The CDI checklist is completed by parents indicating whether or not the child **produces** each word of a fixed set of words. These parent-reported vocabulary measures have been shown to be effective in evaluating children's communicative skills (e.g. Dale, Bates, Reznick, & Morisset, 1989; Thal, O'Hanlon, Clemmons, & Fralin, 1999) and have been shown to be related to language skills later in life (e.g. Feldman et al., 2014). The CDI: Words & Sentences Toddler Form is a checklist of approximately 700 early words, most of which children typically produce by 30 months of age. Dale and Fenson collected norms for the words on this checklist based on a cross-sectional sample of more than 1130 children between the ages of 16 and 30 months (1996). These norms are widely used to evaluate children's language development, counting the number of words that a specific child knows and comparing the count to same-aged peers (Dale & Fenson, 1996; Thal et al., 1999; Dale et al., 2003). Previous work by Heilmann and colleagues has shown that a score below the 11th percentile is predictive of low language skills and a score above the 49th percentile is predictive of normal language skills (2005), however understanding the relationship between an individual child's current and future vocabulary requires longitudinal data and predictive modeling.

CDI data from 83 monolingual toddlers (37 females) were collected as part of a 12-month study. This longitudinal study was conducted at the University of Colorado Boulder DACS lab under the direction of Dr. Colunga. Recruitment for the study included three cohorts and was biased toward recruiting children who were learning language at a slower rate than their peers (classically called **late-talkers**). We evaluate the rate of language learning by computing the CDI percentile, a function of the child's age, sex, and vocabulary size as compared to average, age matched, acquisition trends. Language ability spanned all levels with an average percentile of 37.3

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

Figure 1.1: Example longitudinal CDI data for two children.

at the start of the 12 months and 61.3 at the end. The mean starting age of the children was 17.65 months (range 15.4-19.3). Participants completed monthly behavioral tasks and a parent or guardian completed a CDI vocabulary assessment paired with the behavioral tasks. The study, and thus the CDI collection, was conducted for 12 consecutive months, with the majority of parents completing the forms each month. On average we have 10.9 months of data for each child. All together we have a total of 908 CDI forms. Figure 1.1 represents the longitudinal data. For modeling purposes, we consider the change in vocabulary, or the difference between two sequential CDIs from the same child to be a **vocabulary snapshot**. In total we have 825 vocabulary snapshots in which we can use to train, validate and test our models.

1.2 Model methodology and evaluation

For modeling, we use each snapshot, with the first CDI (containing the child's vocabulary at time t) and the related information about the child at the time of the initial CDI to construct the input representation for the model. We use this input to predict whether an individual word, not known at time t, will be learned by time $t + \Delta t$. We evaluate the model predictions by comparing to the second CDI of the snapshot. Ideally, observations are a month apart. This is not always the case, however, due to the difficulty of scheduling toddlers and their parents. Thus Δt varies slightly across observations. The mean time between observations is 1.1 months.

We build our models such that the model predicts the probability of learning each word. The model is trained on a set of input features related to the child or to the vocabulary of the child. Evaluation of the model considers predictions only for words that are currently unknown, so in

practice the model is not penalized for incorrectly predicting that known words stay known. Because of the way the CDI is collected, we know only the set of words the child learns and not the order in which the words are learned. This may be disadvantageous for certain models, namely the network analysis models that would benefit from sequential information. In future work, we hope to consider sequential learning by either utilizing more fine grained language acquisition data or inferring the order of the learned words. Before discussing model evaluation metrics, we first consider model validation.

To test the generality of each model, we divide the data into training (60% of snapshots), validation and test (each 20%). Recall that there is a bias in the data collection to oversample children with language delays. To control for this sampling bias, we consider the child's **CDI percentile** – a measure of language skill computed based on the sex of the child and the number of words that child knows, according to the CDI, as compared to their age-matched peers. Children that have a smaller vocabulary for their age will have a smaller CDI percentile. For each fold, we ensure that the distribution of CDI percentiles is reflective of the distribution of CDI percentiles in the larger sample. By training and validating on these folds, we hope to minimize model overfitting and to control for variability across children and variability across recruitment cohorts. Another key component to the cross-validation method is that the model is evaluated on unseen children. All snapshots of a particular child are tied to a specific fold. Thus, the model must not only predict vocabulary growth throughout the course of development, but must also generalize to an unseen set of children. This can be compared to a cross-validation scheme that predicts unseen observations of children in which certain sequential observations are available. We acknowledge the usefulness of that approach, but for this work we focus on the generalization to unseen children.

We use three types of evaluation metrics to develop a model. First, we use negative log-likelihood **(NLLK)** evaluations which penalizes overestimation as well as underestimation of learning. We also consider a ranking algorithm which measures the **percent overlap** between the top k words predicted and the actual k words that were reported as learned by a child in a particular month[1] .

[1] This is an unweighted version of normalized cumulative gain, except that we vary the number of words considered

Since each month a child learns a varying number of words, the percent overlap measure is useful in knowing how similar the ranking of the model prediction is to the observed set of learned words. We also consider **Receiver Operating Characteristics** (ROC) measures such as area under the ROC curve (**AUC**). These ROC measures allow us to assess the trade-off between true-positive and false-positive rates. We also fit a threshold to convert our probabilities to binary outcomes, defining the threshold to be the point where the model predicts the same number of learning events as observed in the data. With these binary predictions, we compute **accuracy** and a discriminability measure **d-prime**. With these methods of evaluation we can directly compare model performance across models and frameworks.

While we want to construct predictive models, we are also interested in the role of development and individual differences. It is possible that a model that accurately captures learning of 18 month old children would not capture learning of 25 month olds. Similarly, it might be that a model that is highly accurate for children with a vocabulary of 10 words may not accurately predict the vocabulary growth of a child with 400 words. We train all models across the course of development and for children with very different vocabulary profiles, but we are still interested in variance in predictive accuracy. We consider the NLLK estimates predicted snapshots, sorting snapshots by features we consider potentially relevant. For example, we order snapshots based on the age of the child and compare performance of certain models when predicting younger vs. older children. Similarly, we consider the average age of acquisition of individual words using the normative acquisition data. For example, the word 'mommy' is learned earlier than the word 'table'. We may find that certain models are more accurate at predicting early (or late) words learned in the course of development. In the future, we hope to use these trained models to predict not only the words a child is likely to learn next, but also to provide insight into the learning process itself. Considering developmental effects is the first step towards this larger goal.

---

to match the number the child has learned during the snapshot.

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

Table 1.1: Normative age of acquisition rates can be utilized to construct a baseline model in predicting future language learning.

1.3 Baseline models

We construct a baseline evaluation to test if a particular model is accurate in predicting future language learning. One source of information that predicts when words are likely to be learned is the population-level CDI age of acquisition (AoA) norms. The most comprehensive study (Dale & Fenson, 1996) reports productive vocabulary for over 1130 children between the ages of 16 and 30 months, based on parent reports for 680 words. For example, 78.7% of children produce the word dog by age 18 months. Table 1.1 shows an example of the **CDI AoA norms**. These norms, when aggregated over word categories, are typically used to assess a child's vocabulary in relation to her peers, as quantified by a CDI percentile, given age and vocabulary size. However, the CDI population statistics can be used to predict an individual's learning of a given word at a given age.

The accuracy of future vocabulary predictions for any individual child depends on the nature of variability within the population of learners. Any prediction model based on normed data assumes that children learn in a fundamentally similar fashion to one another. For example, implicit in a prediction model based on normed data is that **late talkers** (children below the 20th CDI percentile) have the same vocabulary trends as **early talkers** (children above the 80th percentile). The aggregation of AoA norms, if accurate, would essentially suggests that these late talkers do not learn words in a different order, just that they learn words later. This suggestion has been directly examined and shown to be false: typical and late talkers learn not only at different rates but they learn different lexical items (e.g. Thal et al., 1999; Beckage et al., 2011). Additionally, knowing what words a child knows now may provide information as to the specific words the child is most likely to learn next, a hypothesis we have explored (Beckage & Colunga, 2013; Beckage, Aguilar, &

Colunga, 2015; Beckage, Mozer, & Colunga, 2015), and will explore in detail in later chapters. More generally, limitations of the norms have been noted by many researchers. For example, the norms do not generalize to all populations (e.g. Arriaga, Fenson, Cronan, & Pethick, 1998; Thal et al., 1999) and the norms mask idiosyncrasies in an individual's learning (e.g. Mayor & Plunkett, 2011).

Despite their shortcomings, the CDI norms may be useful for characterizing an individual child's lexical growth. For a baseline, we compare predictions based on the CDI norms with predictions based on child-specific sources of information pertaining to the child's current vocabulary. For comparison, we construct a few different baseline models. One simply considers the probability of knowing a word to be equal to the proportion of children who know the word. We call this our **baserate** model. This model is absent of information pertaining to the child for whom we are predicting, such as the child's age or the words the child already knows. It should, however, be able to capture specific trends such as nouns being learned before verbs (Gentner, 1982). If all models fail to outperform this baseline model, it is likely that there is no systematicity in word learning across children.

The second version of our baseline models includes age-dependent word predictions. In our **age baseline**, instead of considering the probability of learning equal to knowledge of that word in the larger sample, we condition on the age of the child. Here, we evaluate the predictive ability of a model that assumes children of a particular age tend to have similar vocabularies. This model also assumes that a specific child is likely to learn words similar to the words their peers of the same age will learn.

In practice, we also know that male and female children learn at different rates (Fenson et al., 1994; Dale & Fenson, 1996). To account for this difference, we construct a CDI age model that considers the individual age-specific predictions for children of the same sex as the child for whom we are predicting. This **m/f age baseline** model may suffer more than other models from the small samples size of some age-sex pairs. For example, we only have 3 observations for females 16 months of age in our empirical data set.

Finally, we consider two ways of collecting normative acquisition data for our models. The

	NLLK	% overlap	AUC	acc	d-prime
CDI baserate	.645	.138	.498	.774	.001
CDI age	.453	.307	.691	.804	.057
CDI m/f age	.524	.317	.614	.785	.042

Table 1.2: Baseline model performance using the published age of acquisition norms.

first is to consider the published **CDI** norms discussed above (Dale & Fenson, 1996). The other uses longitudinal data to construct empirical norms from the children in our study. The longitudinal **LCDI** responses includes only 49 children (the number in our training data) as compared to over 1000 in the norming study. The LCDI data, however, samples from the same population as the children in the test set and thus may perform better. Performance of these models is included here. Table 1.2 is baseline performance using the CDI norms. Table 1.3 uses the training LCDIs of the longitudinal study. From the tables we can see that the age specific model returns the highest predictive accuracy in terms of NLLK, % overlap and ROC measures. We can also see that, in general, the CDI norms do not perform as well as using a subset of the CDI data collected as part of the study (Beckage & Colunga, 2013). The best performing baseline is the empirical age-specific model that does not consider the sex of the child.

Our final (informed) baseline is a trained logistic regression. In this case, we trained an individual logistic regression for each word separately, meaning that we have in total 677 models. The goal is to predict, given a child's current vocabulary, if the child learns a specific word in the next month. Aggregated to predict the whole vocabulary of a child, we find a negative log-likelihood score of 0.39 and an AUC of .816. We cannot compute percent overlap in the same way we do for the above baseline models because in the logistic regression model we are making individual word

	NLLK	% overlap	AUC	acc	d-prime
LCDI baserate	.639	.139	.496	.776	.002
LCDI age	.456	.333	.701	.805	.064
LCDI m/f age	.496	.341	.628	.788	.045

Table 1.3: Performance of baseline models using the training data of our longitudinal CDIs to construct empirical age of acquisition norms.

predictions instead of jointly predicting the whole set of words. See Beckage, Mozer, and Colunga for more detail on this logistic regression model (2015).

The baseline models provide initial estimates of the performance we might expect from our vocabulary informed models. We also see from these models that certain information, like the baserate knowledge of a word, even in the same population as the children for whom we are predicting, is not relevant. The poor performance of the age-specific baseline model indicates that we need models capable of capturing differences in the learning of individual words and also the differences across individual children. We use these baseline models as we explore questions about how the child's current vocabulary relates to the child's future vocabulary. We also consider how a lexicon might grow over the course of development.

1.4 Dissertation overview

To investigate the influence of a child's current vocabulary on what words they are likely to learn next we turn to predictive models of early lexical acquisition. By assuming that the child's current vocabulary reflects the combined influence of important linguistic and environmental factors, we can evaluate the representation of the current vocabulary in the ability to predict future language learning of individual children. We assess the ability of a certain model to predict future acquisition of a specific child given information about the child's current language state. The difference in performance, and the ability to combine individual models in an ensemble model, will provide insight into how the child's cognitive and linguistic state affects their future language ability.

We begin the work by considering cognitively informed models that consider a child's lexicon as a growing semantic network. Using network analysis techniques, we evaluate potential growth processes on their ability to model network growth, approximating acquisition trajectories of individual children. Within this network analysis approach, we investigate the role of local interactions and global structure of emerging language graphs as related to the acquisition trajectories of individual children. Here, we begin by replicating previous network analysis approaches as applied to language acquisition (Steyvers & Tenenbaum, 2005; Hills et al., 2009b; Beckage, Aguilar, & Colunga, 2015).

We extend this work by modeling individual learning trajectories as opposed to normative acquisition trends. We then extend the current models to representations of different language features that might be relevant to early language learners. These results offer a way to explore how the (network) structure of vocabulary changes throughout development and how this structure impacts our ability to predict lexical acquisition of individual children.

Machine learning tools are particularly adept at predictive modeling and so we also explore neural network models. We consider a neural network model that jointly predicts the full vocabulary of the child one month in the future. We explore the best way of representing the child's current vocabulary knowledge in order to gain the most accurate predictions of words the child is likely to learn next. We also consider the best way of combining these multiple representations to improve overall performance through ensemble models. The effects of performance may be related to developmentally relevant features such as the age of the child and the child's vocabulary size, which we also explore.

Finally, we conclude the dissertation by comparing the network growth models to the neural network approach. These two approaches capitalize on different aspects of cognition and vocabulary development. We conclude by exploring ensemble approaches, combining the model estimates of these different models to further improve predictive ability.

Chapter 2

Predicting lexical acquisition with network growth models

Children learning words cannot do so in isolation. Instead they must learn the meanings and relationships of words to other words. These same relationships, which make learning initial words challenging, likely offer scaffolding and context that help children make sense of the world around them. The connections and relationships between words may aid in future learning of new words. How the structure of language develops through the course of acquisition and how this structure facilitates future language learning is of great interest to developmental psychologists and language researchers.

Here we set forth to build a predictive model of the words a child is likely to learn next based on the emerging structure of the child's current vocabulary. We represent the child's current vocabulary as a network, with the words as nodes, and edges based on linguistic similarity. Words are learned or added to the graph through proposed mechanisms of growth. These mechanisms assume that growth is driven by either the child's current vocabulary knowledge or the structure in the language environment. With this approach, we have a systematic way of linking possible learning mechanisms to the structure and connectivity of a child's language network.

While it is unlikely that language is represented in the mind as a network, it is probable that the structure of a child's current vocabulary or the structure of the language learning environment facilitates language learning and lexical acquisition. Networks provide a method for abstracting this proposed structure. By modeling language growth specifically, we can explore the influence of linguistic structure as related to the emergence of language. We do not review network terminology

or applications here, (but for introductory terms and applications of network science to language and cognitino see Borge-Holthoefer & Arenas, 2010; Baronchelli et al., 2013; Beckage & Colunga, 2015).

The structure of language, as captured by pairwise relationships among words, has been considered in many different domains. In computer science and cognitive psychology, Quillian (1967; 1969) modeled how semantic knowledge might be stored in, and accessed from, long term memory. His model defined concepts as words, and relations were defined as pointers to word-specific features and relations. Collins and Quillian (1969) extended this work to semantic memory[1] in humans. If semantic knowledge is stored in such a hierarchical, pointer-based fashion, longer distances between relevant concepts should require longer retrieval time. Experimental evidence not only supports this claim, but this representation can capture psychological effects. For example retrieval time scales linearly with the number of edges or levels between words within this representation (Collins & Quillian, 1969).

These early results provided motivation for a model of semantic processing (Collins & Loftus, 1975) known as **spreading activation**. In the spreading activation model, a concept is heard, thought, or otherwise "activated" within the network; with time, this activation spreads, in a decaying fashion, along the edges between words, activating other nearby words to varying levels. If neighbors of the activated words are themselves connected, activation will build within the cluster. The difference between the target word's activation and the activation of nearby words has been shown to be related to ease of retrieval (Collins & Loftus, 1975; Vitevitch, Chan, & Roodenrys, 2012), confusability (Chan & Vitevitch, 2009; Vitevitch, Ercal, & Adagarla, 2011), and speech production (Chan & Vitevitch, 2010). Important to our work, this spreading activation model may capture learning effects as well. For example, this model can account for differences in second language learning (Stamer & Vitevitch, 2012) and as aspects of language learning such as semantic differentiation (Steyvers & Tenenbaum, 2005) and contextual diversity (Hills et al., 2009b; Hills,

[1] In psychology, semantic memory includes properties of language storage and retrieval. It is called semantic memory to contrast episodic memory, not to constrain the type of language properties it includes.

Maouene, Riordan, & Smith, 2010). This type of model has also been used to investigate lexical acquisition in the domain of phonology (Vitevitch, 2008).

More generally, network science, as a discipline, has shown a great deal of interest in quantifying and exploring the topological structure of language networks (for a review see Borge-Holthoefer & Arenas, 2010; Beckage & Colunga, 2015). One robust finding of language network analysis is **"small-world"** properties (Watts & Strogatz, 1998), in which a network has high local clustering but maintains low average shortest paths (geodesics) between words. Language networks exhibit evidence of small-world structure (Watts & Strogatz, 1998; Ferrer i Cancho & Solé, 2001; Dorogovtsev & Mendes, 2001; Motter et al., 2002; Steyvers & Tenenbaum, 2005; De Deyne & Storms, 2008; Vitevitch, 2008; Solé et al., 2010). As discussed in a previous review of language networks (Beckage & Colunga, 2015), the cognitive relevance of this structure is unclear. If language is used as efficiently and quickly as required for conversation, small world structure, at least at the semantic and syntactic level, is a necessity. Even so, research suggests high clustering supports similarity within contexts. Short average paths between words supports smooth transitions between words and ideas. Another key feature of (semantic) language networks is a scale-free (or approximately scale-free) degree distribution (Barabási & Albert, 1999; Steyvers & Tenenbaum, 2005). While possibly not a truly scale-free network (Clauset, Shalizi, & Newman, 2009), the presence of high degree nodes and a heavy tail in the degree distribution may suggest an important process by which connections between concepts are formed, one that may play a role in modeling language development.

In the domain of language learning, this small-world structure is important and may be related to language and cognitive ability. Previous work on small-world structure in early language learners suggests that the absence of small-world features is correlated with late-talking children – those children at highest risk for severe language impairment (Beckage et al., 2011). Similarly, the lack of small world structure in verbal fluency tasks is correlated with the onset and extent of Alzheimer's disease (Goñi et al., 2011). Small-world structure of the language network may be important not only for language use but also for language learning. By modeling the emergence of this small world structure through processes of network growth, we may be better able to predict future acquisition

trajectories of individual children.

In this direction, previous work has linked topological features of language networks to that of the acquisition process. Steyvers and Tenenbaum (2005) proposed language is learned by a process similar to **Preferential Attachment**, with highly connected nodes being learned earliest. Under their model, aimed at capturing semantic differentiation, new words or concepts are learned in relation to already known words (C. Smith, Carey, & Wiser, 1985; Clark, 2002; Waxman & Leddon, 2002). In a network growth framework, a learned word attaches to highly connected nodes in the current vocabulary graph and then forms edges with the neighbors of the attachment node, modeling a process of differentiation. The resulting network maintains scale-free structure found in some semantic networks, and also maintains high local clustering. This growth process can account for the observed topological network structure of the emerging adult language network. This model also shows a strong relationship between the model's ordering of learning and self-reported age of acquisition (Steyvers & Tenenbaum, 2005).

Hills and colleagues (Hills, Maouene, Maouene, Sheya, & Smith, 2009a; Hills et al., 2009b) suggest, instead, that language learning is driven by contextual diversity, or, in network terms, the connectivity of unknown words in adult language. Connectivity of words may be related to structure in the environment (e.g. ball and catch), close proximity in spoken language (e.g. cat and dog), or even close proximity in physical space (e.g. chair and table). The connectivity of individual words in the full language graph approximates the number of contexts and meanings of an individual word. Under a contextual diversity hypothesis, a word is more likely to be learned if it appears in multiple contexts since, with multiple exposures, the ambiguity of meaning will decrease (Saffran, 2003; Yu & Smith, 2007). Experimental and theoretical work has shown many ways in which multiple contexts and exposures can increase the likelihood of learning, whether through cross-situational learning (e.g. Yu & Smith, 2007), or attentional mechanisms (e.g. Frank, Goodman, & Tenenbaum, 2009). This network growth model, meant to model language acquisition through contextual diversity, is known as **Preferential Acquisition**.

Here we propose a network modeling framework to predict the individual words a child is

likely to learn next. Whereas previous models have focused on normative acquisition, we apply these growth models to the language trajectories of **individual children**. We additionally formalize the relationship between network analysis and language acquisition. The use of network models to explain acquisition requires 1) a clear definition of edges or similarity between words, 2) a systematic influence of the network structure on future vocabulary acquisition, and 3) a way of relating network structure to the acquisition of individual words. In the next section we define our framework before evaluating performance of models based on our framework. We find that each aspect plays an important role, impacting our ability to accurately capture the language acquisition of individual children. Our framework offers novel insight into the acquisition process as related to network growth models. We also see applications of network models to capture cognitive mechanisms. Most promising is that our models outperform the CDI age of acquisition (AoA) age-specific baseline especially when predicting children who are learning language slower than their peers. We find a strong relationship between this improvement in performance and assumptions about growth mechanisms underlying our model.

2.1 Network growth modeling framework

When explaining vocabulary acquisition with a network growth model, we should consider all aspects of the network structure that might influence predictability and interpretability. Towards this goal, we propose three levels of analysis in which to frame and understand our models. This framework also helps us quantify the ability of our model to accurately predict learning.

(1) **Macro level**: The definition of a graph in terms of nodes and edges: a criterion specifying when an edge to exist between two nodes.

(2) **Mezzo level**: Measure of (changing) structure on a chosen graph: structural properties that capture influence and change.

(3) **Micro level**: The interaction of nodes with other nodes: centrality measures as a quantification of node importance.

In the case of child language acquisition, the strongest form of the hypothesis assumes that if we have 1) a meaningful definition of relationships between words, 2) a growth process that correctly approximates learning, and 3) a cognitively relevant measure of word importance then we can accurately model the acquisition trajectory. We evaluate our assumptions on our ability to predict the specific words an individual child is likely to learn next, given the child's current language network.

Beyond the goal of accurate prediction, we can also investigate performance of our models at each of these levels of analysis. For example, at the macro level, we can ask whether, given the acquisition trajectories of individual children, semantic features, phonological similarity or co-occurrence are more predictive in capturing the learning of new words. The difference between phonological similarity and semantic feature similarity has previously been considered, assuming that words are added based on a "rich get richer" or preferential growth model (Beckage, Aguilar, & Colunga, 2015). Assuming the ability of a network representation to model acquisition trends is related to the importance of that linguistic feature in language learning, we investigate what types of information are accessible and relevant to young children. We consider changes through the course of development as measured by the age of the child, the vocabulary size of the child and the child's CDI percentile, which quantifies the vocabulary size of the child compared to their age matched peers. We also consider the emerging structure of these various network representations across development.

After exploring the macro level, we turn to the mezzo level and micro level simultaneously. Here we ask whether different models and different centrality measures affect the accuracy with which we can predict the trajectory of individual children. We consider the models of Hills and colleagues, explained in detail below, with the various macro level network representations. We evaluate the ability of the models to predict, for an individual child, the words the child is likely to learn next. Important to our framework, we consider the role of each learned word, not only in terms of growth process but also in terms of centrality in the network. If we assume that learning a word has a specific utility and that the utility is related to network structure, we can directly

explore how structure influences that utility, and the influence of the probability, of learning a specific word. Here we can explore questions related to why certain words are learned before others and what mechanisms influence the evolving network structure.

By considering each level of analysis, we can begin to understand the interaction of different levels and the role of representation. Further we consider the ability of the macro, mezzo and micro levels to explore the acquisition trends of young children. We evaluate the role of each level in a predictive modeling task of word learning. We use the resulting predictive performance to inform our understanding of the complex process of early acquisition. These network models are not only predictive models but they also suggest mechanisms and attentional influences that alter network growth trajectories, capturing young learner's acquisition patterns.

2.2 Previous work

Initial research questions relating to network representations and language acquisition have focused on proposed mechanisms and processes that account for the structure of the graph (Steyvers & Tenenbaum, 2005; Vitevitch, 2008). To capture or account for a cognitive process that could give rise to the observed structural features of adult semantic networks, Steyvers and Tenenbaum suggested a process of semantic differentiation that could be modeled by a preferential attachment mechanism (2005). Analysis of the topological structures of adult semantic graphs was based on three different definitions of semantic relations between words: 1) Nelson free association norms (Nelson, McEvoy, & Schreiber, 2004), 2) Roget's thesaurus (Roget, 1911) and 3) WordNet (Miller, 1995). The Nelson free association norms were collected by asking individuals to respond to a word with the first semantically related word that came to mind. For example, the prime 'cat' generates responses such as 'dog', and 'mouse'. The other two network representations were based on explicitly defined similarity relationships in terms of shared meaning (thesaurus) or labeled linguistic relationships (WordNet).

Topological analysis of network structure indicated that all three of these semantic networks had similar large scale structure, with high local clustering and short average path lengths between

words. The authors also found evidence of a power-law in the degree distribution across these three networks (Barabási & Albert, 1999). To account for the topological structure, they construct a model where nodes are learned and attached to a specific attachment node proportional to the degree of the attachment node. Once a word is added to the graph, edges are also drawn to the immediate neighbors of the newly added node. The direction of these was biased to go from the newly added node to the near neighbors, further increasing the attachment probability of the previously known words (Steyvers & Tenenbaum, 2005). This step, of adding edges within the local neighborhood, ensures local clustering. Coined 'preferential attachment', we call this model **Preferential Growth** to distinguish it from the original preferential attachment model which considers probability of the nodes' attachment point rather than probability of a node and respective edges being added.

Using their model of Preferential Growth, Steyvers and Tenenbaum strong correlation between reported age of acquisition and global network structures (specifically in the degree distribution) of early language networks. Nodes with a higher degree in the grown network are acquired earlier as reported by self-assessed age of acquisition (AoA). Further, when the observed semantic networks were reduced to the vocabulary of young children, the simulated Preferential Growth showed similar structural properties as the acquisition networks (Steyvers & Tenenbaum, 2005; De Deyne & Storms, 2008). One shortcoming of the work is that it considers an unlabeled graph. Summary of the emerging structure through such measures as clustering coefficients and the degree distribution of the growing networks may miss important features of the learning process.

Hills and colleagues modeled trajectories using empirical CDI age of acquisition norms, extending these models to language acquisition. Models used only 130 words of a network based on the McRae features or 532 words of the Nelson network as opposed to the 680 words on the full CDI. The authors considered the preferential Growth model as well as two additional models called **Preferential Acquisition** and **Lure of the Associates** (Hills et al., 2009b, 2010). Figure 2.1 offers a visualized representation of the growth processes as related to the acquisition of individual words under these models. The resulting model performance, based on normative age of acquisition results, investigated the relationships between lexical items and normative word learning a month

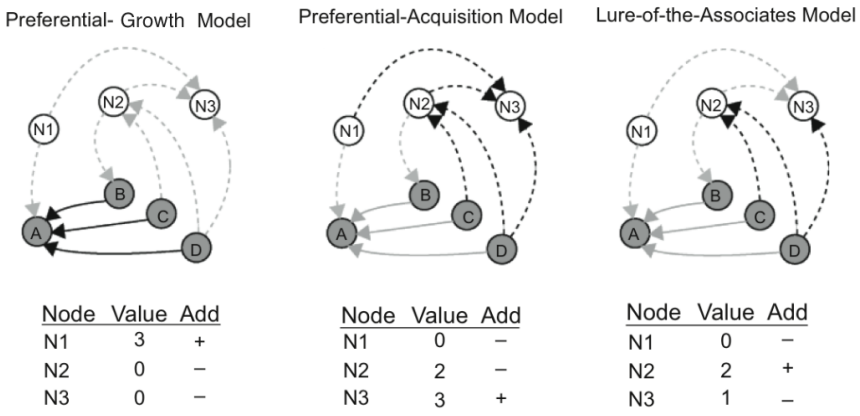

Figure 2.1: The three growth models depicted in a simplified network. From Hills et al. 2009b

into the future. Specifically, the authors investigated the influence of the child's current vocabulary as compared to the structure of the full language network.

In figure 2.1, the underlying network is consistent but the role of the individual edges differs according to the assumed model. Black edges indicate which edges matter to the given network model. Solid edges indicate relationships between known words and dashed edges indicate that at least one of the words along the edge is unknown. The network structure, and the model assumptions, differ such that the probability of learning the unknown (white) words differs under these models. With **Preferential Growth**, words are more likely to be learned if they connect to well-connected words in the **known** vocabulary. Here word N1 is most likely to be learned. Alternatively, **Preferential Acquisition** assumes connectivity in the **language learning environment** or the full language network drives learning, making the acquisition of word N3 most likely. The **Lure of the Associates** also considers how connected the unknown word is, but conditioned on the edge connecting from an already known word, and would most often select N2 for learning.

Assuming these growth models capture acquisition, it is possible to define the probability of learning an individual unknown word i as:

$$P(Y_i) = \frac{e^{\beta\delta_i}}{\sum_j e^{\beta\delta_j}}, \tag{2.1}$$

where j is the set of unknown words and δ_i is the **growth value**, or utility of learning a particular

word, defined based on the assumed growth process. β is a free parameter that allows for variation on how sensitive the model is to the growth value δ_i. Each model defines the growth value of a particular word based on network structure. For example under the Lure of the Associates model, the growth value δ_i is the indegree of word i of the specific child's known vocabulary graph when i is added to the graph. Each additional model can similarly be defined to produce a mapping of network structure to a single δ value for each word.

Here we extend this work to look at the learning trajectory of individual children rather than normative acquisition. This introduces additional challenges for models as individual differences might influence not only which **growth processes** are most accurate but also how accurate a specific **network representation** is at accounting for acquisition. Another challenge is that the number of words learned in a given month no longer varies only according to age but also according to differences in learning and differences between observation time. These models are sequential in so far as learning an individual word alters the structure of the network graph which in turn may alter the probability distribution over words that are likely to be learned next. While we consider batch rather than sequential orderings due to data limitations, we do attempt to understand how development and individual differences affects our ability to model acquisition. We analyze the resulting models using the framework discussed above, specifically exploring the interaction of network levels with development.

Even with these limiting assumptions and the challenges of modeling individual acquisition, we are able to predict words learned by an individual child. We find especially strong evidence that the type of growth process used affects and influences our ability to accurately predict. We also find that, for particular populations of learners, namely **late talkers**, we are able to improve greatly over CDI AoA age-specific baseline predictions. The type of centrality measures considered also influences predictability, suggesting that learning differences may be quantified with this approach. We now turn to the specifics of our methods before turning to the modeling results.

2.3 Methods

We extend the models of Hills and colleagues to our multilevel network growth modeling framework. We also model the trajectories of individual children, explaining variation across learners. We use the success and failure of particular **network representations** to understand differences and variance in children's acquisition patterns. We consider many different network representations to explore the macro level network representation. We explain the specifics of the network definitions in detail below. Considering the same **processes of growth** as discussed above, we additionally consider different types of **network centrality** (beyond indegree) to explore the interaction of the mezzo and micro levels of analysis. With these predictive modeling results, we consider differences in learning during the course of development and across children.

We begin by redefining the probability of learning an individual word Y_i to be

$$P(Y_i) = \frac{1}{(1 + e^{-\beta(\delta_i - x_0)})} \tag{2.2}$$

where i is the word to be learned and δ_i is the **growth value** for word i. Under this model, the growth value is defined based on an assumed growth process and an assumed centrality. The centrality measure can be either local (e.g. degree) or global (e.g. betweenness) and is affected by the choice of network representation. The **growth value** δ_i, which is minimally zero, is mapped to a probability through a logistic transformation characterized by a scaling parameter β and an intercept x_0. At x_0 the model returns a probability of 0.5 for learning that specific word. Note, that equation 2.1 has an additional constraint that the probability of all known words sums to 1. Here we relax this constraint because we are interested in simultaneously predicting the set of words learned by an individual child, not the single most likely word. β and x_0 are optimized across training snapshots for each network, growth process, and centrality we compare.

Because network growth models cannot generalize to words outside of the network representation, we use the CDI AoA age-specific baseline model, which estimates the likelihood a child of a specific age learns a specific word, to augment the network predictions. This means that if the specific model does not predict the learning of a word in the CDI norms, we will use the baseline

prediction. If the model does predict a specific word, we learn a weight of the CDI model and the network model, combining the baseline and network predictions.

2.4 Longitudinal CDI data

To evaluate our models, we need detailed data of the words that a child learns through the course of development. One well-established way to measure and characterize toddlers' lexicons is to use vocabulary checklists, such as the MacArthur-Bates Communicative Development Inventory (**CDI**; Fenson et al., 1994). The CDI checklist, completed by parents, indicates whether or not the child **produces** each word of a fixed set of words. These parent-reported vocabulary measures have been shown to be effective in evaluating children's communicative skills up to 30 months of age (Thal et al., 1999; Arriaga et al., 1998). The CDI: Words & Sentences Toddler Form is a checklist of approximately 700 early words, typically produced by the majority of children by 30 months of age.

Longitudinal CDI data from 83 monolingual toddlers (37 females) were collected as part of a 12-month study, conducted at the University of Colorado Boulder, Colunga lab. Recruitment for the study was done in three phases and was biased toward recruiting children that were learning language at a slower rate than their peers (classically called **late-talkers**). Language ability, evaluated based on **CDI percentile**, spanned all language learning levels with an average learning percentile of 37.3 at the first of 12 visits and 61.3 at the end. The mean age of children was 17.5 months (range 15.4-19.3) at the first visit. On average, we have 10.9 CDIs (minimum of 2, maximum of 12) for each child. All together we have a total of 908 CDI forms. Figure 2.2 represents the type of longitudinal data utilized for modeling. For modeling purposes we consider the change in vocabulary, or the difference between two sequential CDIs from the same child, to be a **vocabulary snapshot**, with the first CDI being the initial CDI and the second being the prediction CDI. In total we have 825 CDI vocabulary snapshots.

One goal of our modeling approach is to predict the individual words a child is likely to learn next. To this end, we model how the network of an individual language learner changes over time by predicting when nodes will enter the graph. Figure 2.3 visualizes four network graphs of the

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

Figure 2.2: Example longitudinal CDI data for two children.

same child through the course of development. The edge list (discussed more below) is based on McRae feature norms. Our research goal is to construct a network growth model that captures the **evolving networks structure** through accurately predicting the individual words the child is likely to learn next.

2.5 Network construction

To characterize the changing network structure over time, we discuss explorations of model performance at each level of the network growth framework. First we introduce the various types of network representations considered in our modeling. We analyze the topological structure of these network representations as these topological structures have previously been shown to relate to cognitive aspects of language learning and use (e.g. Collins & Loftus, 1975; Chan & Vitevitch, 2009; Beckage et al., 2011; Goñi et al., 2011). We compare performance at the mezzo level by considering the three network growth models discussed in detail above. We evaluate performance of these growth models on their ability to correctly predict the acquisition trajectories of individual children. Finally, we explore what types of relationships are important to young children by approximating the **growth value** δ_i using measures of node centrality. Table 2.1 indicates the types of networks (macro), growth processes (mezzo), and centralities (micro) explored in the work to follow.

Turning first to the macro level of the modeling framework, we evaluate the role of a specific network representation based on the ability to accurately predict the words learned by an individual child beyond the CDI AoA age-specific baseline model. Because the **growth value** of each word is related to the word's connectivity in the graph, the definition of our network directly impacts

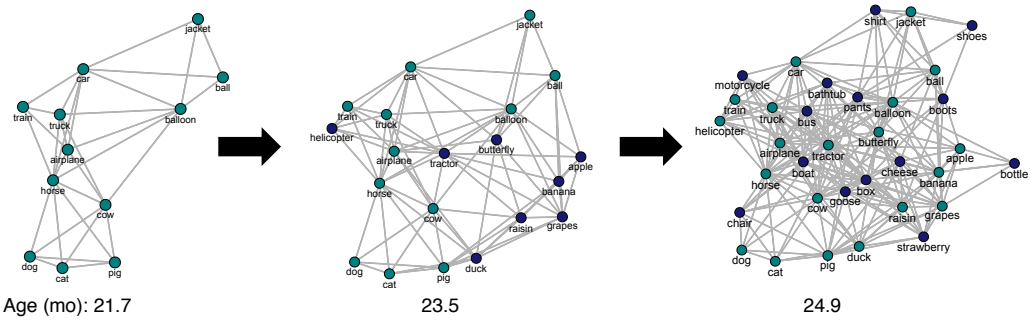

Figure 2.3: Network representation of a child's growing productive vocabulary.

Macro level Network Representations	**Mezzo level** Growth Processes	**Micro level** Centrality measures
Co-occurrence in CHILDES	Preferential Growth	betweenness
Howell Sensory Motor Feat.	Preferential Acquisition	closeness
Nelson free association	Lure of the Associates	degree
McRae feature norms		eigenvector
Phonological edit dist.		
Phonological overlap		
Word2Vec using GoogleNews		

Table 2.1: Levels of analysis for network growth modeling predicting individual acquisition trajectories of young children in our longitudinal study.

our ability to model word learning. Previous work has found differences in predictability can be attributed to network representation. For example, both Steyvers et al. and Hills et al. found that the Nelson free association norms network was more accurate than a network based on the Roget's thesaurus (Steyvers & Tenenbaum, 2005) or McRae feature norms (Hills et al., 2009b). Here we compare seven different types of weighted networks.

The networks we use are:

(1) **Co-occurrence** in CHILDES corpus (MacWhinney, 2000). Edge weights indicate the count of how often word a follows word b within a sliding window of five words.

(2) Cosine similarity of **Howell** feature norms. The feature norms are based on human ratings of how much a particular word embodied a particular feature (Howell, Jankowicz, & Becker, 2005). Participants were asked to respond with the frame of reference of a pre-school child.

(3) Count of shared features from the **McRae** feature norms. The feature norms are based on features listed by participants of a specific set of items (McRae, Cree, Seidenberg, & McNorgan, 2005). A weighted edge exists indicating the proportion of shared features between two items. Construction mimics the network construction of Hills et al. (2009a, 2009b).

(4) **Nelson** free association norms (Nelson et al., 2004). An edge represents the proportion of individuals who, given a cue word a, responded with word b. For example the cue of *dog* frequently elicits the response *cat* and in the network an edge exists from dog to cat.

(5) Phonological Levenshtein Distance (**Phono. Dist**) (Vitevitch, 2008). The inverse of the number of substitutions, insertions or deletions required to transform the phonological form of one word to another word. For example *dog* and *hog* have an edit distance of 1 (by substitution).

(6) Phonological Overlap (**Phono. Overlap**) (Beckage, Aguilar, & Colunga, 2015). The number of shared phonemes between two phonological forms. Normalized by the length of

the word.

(7) Cosine similarity of **Word2Vec** vectors. Training with the GoogleNews corpus, 200 dimensional vectors were created using an online tool and the Word2Vec algorithm (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013). Adult words were used as copora related to child language were not large enough for the application of the Word2Vec algorithm.

Bold words in the above description indicate the label used in the results section. All networks are initially weighted networks that have been rescaled so that the minimal edge value is 0 and the maximal is 1. We use a standard optimization method to discover the best threshold, for each network representation, converting the network to an unweighted graph. We chose to use binary networks because it is difficult to assess what the edge weights would be for young children or even if young children have access to the gradated information of edge weights. Once thresholding is performed, any edge with a weight below the fitted value is excluded, resulting in a network that is binary and, except in the case of cosine similarities, directed.

2.6 Network growth modeling

For training individualized network growth models, we utilize cross-validation. Training data consists of 60% of the children and a total of 484 snapshots. Validation and test sets were divided evenly with 170 and 171 snapshots respectively. Note that models were evaluated not only on performance of unseen snapshots but also on unseen children. Each training example consisted of a **vocabulary snapshot** or a paired set of CDI reports collected at approximately one month intervals. The initial CDI was used to construct the child's known vocabulary network. The **growth value** was then computed based on the network representation, growth process, and centrality, conditioned on the known vocabulary network. The growth value is then converted to a probability through a logistic transformation, with fitted scale and intercept parameters. The resulting probability indicates, for each word not known by the child at the initial CDI, the probability that the word will be learned by the next CDI for that child. Observations are approximately a month apart, but

the time between snapshots varies slightly across observations.

Each model has a total of four parameters. The network threshold can range between 0 and 1. A value of 0 results in a binary network where any edge is present in the full graph. A threshold of 1 results in a fully disconnected graph. From the thresholded network representation, we calculate the **growth value** for each snapshot conditioned on the centrality measure and network growth process. For example, the **Preferential Growth** model has a growth value conditioned only on the connectivity between words the child knows, whereas growth values based on **Preferential Acquisition** is computed as connectivity in the full network representation. We normalize the growth values across individuals. Note that the normalization helps performance of all models. This is due to the correlations between number of words learned and the size of the child's current vocabulary. If a child does not know many CDI words initially, they are unlikely to suddenly learn a bunch of words, whereas if they already know many words, they are more likely to continue to learn more words, including those on the CDI. Normalization helps capture this trend. Once growth values are computed, the additional two parameters which scale and center the logistic transformation are optimized. The final parameter weights the CDI AoA age-specific baseline model with our network growth models. Because the size of the network varies based on information related to the network representation, we assume words on the CDI but not in the network representation are learned according to the normative CDI AoA age-specific baseline model (c.f. section 1.3).

2.7 Growth modeling equations

Here we briefly formalize the growth models to provide mathematic intuition and understanding of the mechanisms of growth. See Figure 2.1 for a visual representation of these models. For the mathematical formalization, we assume that each network graph can be represented as a square matrix N of size $|V|$, where $N_{i,j}$ can be either 0 or 1, indicating the presence or absence of an edge from i to j. We represent the network **induced** by the set of known words K to be N^K where $K \subseteq V$. We let c_j^N represent the "importance" of node j, given network N. All models make predictions only for unknown words. The resulting growth value δ_i is a function of c_j^N and is

standardized before being mapped to a probability through equation 2.2. Node "importance" simply means that there is some utility that an individual (unknown) word has for a specific child. In our modeling we will consider different types of centralities as a way of operationalizing the importance of individual nodes. Other types of word-specific measures, such as frequency or word length, could also be used.

Preferential Growth assumes that words are more likely to be learned if they connect to nodes that are themselves well connected in the graph. A word is added to the graph, or learned by the child, in proportion to the sum of the importance of each **known** word it attaches to. For example if the child knows many animal words, with the word *dog* being the most "important" word in the child's known vocabulary, the Preferential Growth model model would predict that words are more likely to be learned if they attach to the word *dog*. For this model, and all models, we presume the edges are fixed by the chosen network definition. The edges form the basis for the calculation of a word's importance and growth value. We additionally presume that if a word is learned, the word and the word's respective edges to known words are added to the graph of the child's known vocabulary. We do not consider words to be learned in order but instead predict the joint set of words. Under this definition, the Preferential Growth model can be defined as:

$$\delta_i = \sum_{j \in K} N_{i,j} \; c_j^K \tag{2.3}$$

Letting c_j^K be the degree of word j in the known graph, the **growth value** for word i (δ_i) is set to be the sum of the "importance" of the known words, given that an edge exists between nodes i and j. This suggests that words are more likely to be learned if they connect to high degree, already known words. One possible cognitive mechanism that could drive this model is semantic differentiation, in which words are learned if they are similar to already known concepts (Steyvers & Tenenbaum, 2005). One key feature of this model is that predictions of word learning are driven only by the words that the child currently produces.

The second model, called **Preferential Acquisition**, assumes that words are learned based on their connectivity in language. This connectivity is approximated by the **'full language graph'**.

We consider this full language graph to be the graph constructed when all of the 677 CDI words are known. In the original paper exploring this model, two types of full language graphs were used (Hills et al., 2009b). One contained all words that were part of the network whether or not the words were on the CDI form. The other version considered only the words that are on the CDI form (Hills et al., 2009b, 2010). There was a small improvement in predictive accuracy when using the CDI word graphs so here we only consider this variant. In the case of the networks considered here, we usually only have network connectivity for a subset of the CDI words. In this case we consider the full language graph to be the maximally overlapping set. Mathematically, the growth value of each word is defined as follows:

$$\delta_i = c_i^N \tag{2.4}$$

Here N is the full network as specified at the macro level. This model relies on the idea that the more important a word is, the earlier it is learned. Additionally, this model assumes that the full language graph approximates the **language environment** and **linguistic context** that is important to child learning. The centrality of a word varies based on the specific network definition, but for some networks, such as the Howell feature norms, success using this growth process would indicate that early learned words share many of the same (sensory-motor) features with other words. In the case of the Nelson free association norms, high importance would indicate that the word is the response to a variety of different cue words and central in free associations. This, in turn, could indicate that specific words appear in many different contexts. If this model outperforms other models, we might conclude that words with diverse contexts are learned earliest. This model is capable of capturing large scale acquisition trends such as the fact that *dog* is often learned before *duck*, possibly because the number of contexts (or degree) of dog is higher than that of duck. Regardless of the specific interpretation, this model suggests that all words are learned in effectively the same order, with minimal individual differences in the order in which words are learned. The **growth value** of this model is not based on the individual words the child knows but only on the connectivity of words in the language graph.

The final model, **Lure of the Associates**, again proposed by Hills and colleagues (Hills et al., 2009b) bridges the gap between a model based only on the words in the child's **known vocabulary** and a model based only on the connectivity of words in the **language environment**. Here, the word is learned proportionally to node "importance" but conditioned on links existing between the to-be-learned word and known words. Here we compute the "importance" of a word if it was added to the known graph. If the word has higher importance than other unknown words, then this word would be more likely to be learned. For example, if a child has many animal words and many water words, learning words like *duck* and $fish$ might be more 'important' under such measures as betweenness because it provides a bridge between the water and animal concepts. This model presumes that the words that are most likely to be learned are words that will become most important in the productive vocabulary graph (in comparison to other unknown words) once learned.

$$\delta_i = c_i^{K \cup \{i\}} \tag{2.5}$$

Allowing for word importance to be based only on pairwise relationships in which at least one element of the pair is known suggests that children may need context and understanding to ground learning. The reason that this type of model might perform best is because children may learn new words by finding distinguishing, but relevant, differences between a new concept and one the child already has. Another possibility is that children (or caretakers) could have explicit interests (such as in animals) that causes animal words, but the most 'important' animal words, as approximated by graph centrality in our models, to be learned first.

We train the slope and intercept terms in order to minimize negative log-likelihood (nLLK). For our individual word predictions, we combine the network-based predictions with predictions using the CDI AoA age-specific baseline model after optimizing the weight of the baseline model. We include the CDI AoA age-specific baseline model because the network representations vary in size and the growth models cannot generalize to words not in the network. If the word is not in the network representation, the baseline CDI AoA age-specific model estimate is used.

For the current analysis, we define "importance" to be a measure of graph centrality. Centrality measures calculate the role of each node in the graph. Although there are many different types of centrality capturing different types of node importance, we consider only four types of centrality. The first is in-degree centrality as put forth and considered on normative vocabulary snapshots by Hills and colleagues (Hills et al., 2009b). We also consider betweenness, closeness and eigenvector centrality. Though these measures are correlated, there are differences in terms of interpretation when using these different centrality measures. **Degree** centrality models presume that the number of neighbors a word has is relevant to making a prediction of future word learning. **Betweenness** centrality is calculated in order to consider the number of shortest paths that contain a particular word. Thus words are more likely to be learned if they provide new and/or shorter paths between currently known words. **Closeness** centrality considers the total distance between a particular node and all other nodes. Here vocabularies are assumed to grow based on minimizing shortest paths between words. Finally, **eigenvector** centrality approximates the influence of a node by considering not just the number of neighbors but also how important and influential those neighbors are in the graph. Word learning according to eigenvector centrality, similar to PageRank (Brin & Page, 1998), would imply that words are more likely to be learned if they are connected to words that are themselves well connected and influential at the moment of learning. Of possible relevance to modeling early acquisition, these measures weigh local and global information differently. In the case of degree centrality, one need know nothing about the additional structure beyond a node's immediate neighbors. This is considered a local measure of centrality. Global centrality measures, such as betweenness, require information not only about a node's neighbors but also a node's neighbor's neighbors and beyond. The trade off between local and global connectivity of language learning may be important in modeling infant language acquisition.

2.8 Model evaluation

Model selection is based on the average negative log likelihood (nLLK) for each unknown word in the training set. This measure penalizes both overestimation and underestimation of learning. We

use our validation set to perform model selection at the macro, mezzo and micro levels of analysis. All together we are comparing a large set of models, as we believe that each of these levels has a separate but important influence on modeling acquisition of young children. Only after selecting models of highest predictability over paired random network models, do we extend our models to the test set. This procedure may miss out on some of the population differences of individual learners but does provide a systematic procedure for considering a large population of models simultaneously. We select models based on their prediction to words in the network only, and do not consider the weighted average of the network predictions and the CDI AoA age-specific baseline until applying models to the test set.

After deciding which models to investigate further, we retrain the models using the combined validation and training data. We conclude model training by using the validation results to learn the best linear weighting of the baseline CDI AoA age-specific predictions and the network based predictions. We include these baseline CDI AoA predictions so that our models predict the learning of all words on the CDI as opposed to just those words in the network representation. Once fully specified, we evaluate our test models on nLLK, percent overlap and ROC measures. The nLLK tells us the overall accuracy in correctly predicting which words are learned and also which words are assumed not to be learned. Percent overlap considers the overlap between the top k' words predicted by the model and those k words that are actually learned by the child. This measure does not, however, penalize the model for incorrectly suggesting words for learning that were in fact not learned. Finally we consider performance of the model with ROC measures of discriminablility and accuracy, calculating the area under an ROC curve, accuracy and d-prime. The curve considers the proportion of true positives to true negatives as the threshold for converting the probability estimates to binary outcomes varies.

We are also interested in developmental changes and trends in the data. We compare the best performing model as a function of child-specific features such as age, vocabulary size and percentile. We compute a smoothed average of nLLK (on the validation data set) sorting by these child level features. Here we hope to see effects of age or language learning abilities more clearly.

We conclude by constructing basic ensemble models that aggregate predictions of different network representations.

2.9 Network topology

We begin by training each combination of macro, mezzo and micro levels, cf Table 2.1. In total 84 models were trained. We compare performance of these models initially to a network baseline model that assumes a fully connected language graph. Because our theory of acquisition includes the influence of network structure to support and direct language learning we first examine the topology of the full network representations and the developmental language networks as well.

As discussed above in the methods section, we use a binary graph. We optimize the edge-weight threshold for individual network representations in conjunction with the scaling parameters. The topological structure of these binary networks may be related to model accuracy, providing insight into the structural mechanisms that result in predictive accuracy. Here we look at the full network graphs based on maximal overlap with the words on the CDI.

Because not all network representations contain the full set of CDI words, the networks themselves vary in size (V). In table 2.2 summary statistics of these binary networks are reported. As can be seen, the networks also vary in network connectivity or density (den), making direct comparison difficult. Still, we see unifying trends across all network representations. All of the language networks considered here are characterized by high clustering coefficients (CC) as compared to the density of the graph, a feature commonly attributed to small-world and language networks (Watts & Strogatz, 1998; Ferrer i Cancho & Solé, 2001; Dorogovtsev & Mendes, 2001; Motter et al., 2002; Steyvers & Tenenbaum, 2005; De Deyne & Storms, 2008; Baronchelli et al., 2013; Beckage & Colunga, 2015). We consider the mean degree ($\langle k \rangle$) of these networks as it is relevant to one of the random graphs we consider below. We also see that the mean geodesic distance ($\langle d \rangle$) of language networks are nearly the same as what one would expect if edges were drawn with a fixed probability as in an Erdős-Rényi (ER) random graph ($\langle d_r \rangle$). This suggests that while these networks have local structure, global paths are still available and easily accessible. Also reported is the diameter (D) or

maximum geodesic distance of the network.

Nearly all nodes in the graph are weakly connected with at most two isolates. The giant component of these networks are close to including all words in the graph, indicating that not only is the graph weakly connected but there is one connected component that dominates (giant component size not reported). Considering the degree distribution, the best fitting power-law exponent (γ) is reported in table 2.2, asterisks indicate evidence of scale-free structure. We perform a significance test using bootstrapping as discussed by Clauset et al. (2009) and find reliable super-linear relationships for the co-occurrence and Nelson networks. While we do not consider whether these networks are best fit by exponential or power-law distributions, other language network research explores this question in more detail (Dorogovtsev & Mendes, 2001; Motter et al., 2002; Steyvers & Tenenbaum, 2005).

Finally, we report the correlation of the degree of a word with its neighbors, also called the assortativity coefficient (a). This is a topological feature that is difficult to capture using random networks as we show below. In the co-occurrence network, we find that highly connected nodes tend to be connected to nodes that are themselves not highly connected, as characterized by a negative assortativity coefficient. Alternatively, in the case of phonological distance, nodes are likely to be connected to nodes of similar degree, a common feature of phonological networks (Vitevitch, 2008; Stella & Brede, 2015) Both **Howell** and **Phono Dist** network are very dense. Even so, we find this network structure still provides predictive information about individual acquisition trajectories, especially for snapshots with small CDI vocabulary sizes.

We now consider whether the topological features of the observed language graphs can be replicated through random networks. Previous work has shown that the Erdős-Rényi random models, where edges are generated with a fixed probability, do not characterize language graphs well (Motter et al., 2002; Borge-Holthoefer & Arenas, 2010; Beckage & Colunga, 2015). We do not consider this model beyond the report of geodesic distance ($\langle d_r \rangle$) in table 2.2. We consider the configuration model (**CM**) where similarity in topology of the CM model to the observed networks indicates that the network structure is due to the network's degree distribution. In the CM model,

Network	$\|V\|$	den	CC	$\langle k \rangle$	$\langle d \rangle$	$\langle d_r \rangle$	D	γ	a
Cooc	635	0.195	0.535	124.1	1.85	1.80	4	2.34*	-0.322
Howell	334	0.626	0.807	208.6	1.36	1.29	3	1.32	0.084
McRae	133	0.289	0.551	38.2	1.74	1.69	3	1.72	0.005
Nelson	545	0.009	0.153	5.3	4.65	3.86	12	3.32*	0.012
P. Dist.	677	0.689	0.926	466.1	1.32	1.28	6	1.91	0.081
P. Overlap	677	0.495	0.657	334.9	1.50	1.50	2	2.00	0.042
Word2Vec	626	0.415	0.590	259.5	1.58	1.58	3	1.74	0.148

Table 2.2: Network summary statistics of the seven models used in our analysis below. Reported is graph size ($|V|$), density, average degree ($\langle k \rangle$), average clustering coefficient (CC), geodesic distance of the observed network ($\langle d \rangle$) and the geodesic distance of an ER random graph ($\langle d_r \rangle$). Also reported is the network diameter (D), best fitting power-law exponent (γ) and the assortativity coefficient (a).

an edge between i and j is generated with probability $(k_i^{out} * k_j^{in})/m$ where m is the number of edges in the graph, k_i^{out} is the outdegree of node i and k_j^{in} is the indegree of node j. We control for self loops and multiple edges between i and j. We also consider the growth model proposed by Steyvers and Tenenbaum (**STM**; Steyvers & Tenenbaum, 2005). In this model, an attachment node is chosen proportional to the degree; the new node is connected to the attachment node. Letting M be the average degree in the network, $M-1$ other edges are randomly added between the new node and the neighbors of the attachment node. Edges are directed, with a bias for edges of the newly learned word to point to already known, well-connected, words. Previous analysis has suggested that this network growth process accounts for emerging structure during language learning. It can also account for a super-linear degree distribution while maintaining the observed local clustering of language networks. Standard deviations across 10 repeated runs of the random models are small and do not alter conclusions drawn. They are not reported for clarity.

Table 2.3 contains information about the performance of the configuration model (CM) and the Steyvers and Tenenbaum (STM) model. For high density graphs, the Steyvers and Tenenbaum model could not be run because the required initialization condition has a fully connected set of nodes equal in number to the mean degree. Table 2.3 suggests that the random models cannot capture certain characteristics of the observed network structure. None of the random models can account accurately for the observed assortativity coefficients. Another difference between the random models and the observed networks is that the clustering coefficient of the observed networks is hard to recover. STM comes closest to capturing observed clustering but it still over- or underestimates clustering. The configuration models tend to result in a clustering coefficient that is less than the original network and closer to network density.

Topological analysis suggest that the observed network properties cannot be explained by the degree distribution or the Steyvers and Tenenbaum model of growth. We find that clustering and assorativity of the language graphs are especially difficult to capture with the random network models explored here. This local structure of high clustering and assortativity may be important to early language learning. But the topological structure of the full language networks may not be

Network	den	CC	$\langle k \rangle$	$\langle d \rangle$	D	a
Cooc CM	0.182	0.496	115.3	1.85	4	-0.087
Cooc ST	0.195	0.550	123.8	1.80	3	-0.122
Howell CM	0.996	0.996	331.9	1.00	2	-0.004
McRae CM	0.285	0.375	37.6	1.72	3	-0.018
McRae STM	0.285	0.644	37.7	1.70	3	-0.123
Nelson CM	0.009	0.027	5.3	4.00	9	-0.042
Nelson STM	0.009	0.113	4.9	3.05	6	-0.142
P. Dist CM	0.952	0.983	643.8	1.04	3	-0.001
P. Ov CM	0.492	0.596	332.7	1.50	3	-0.012
Word2Vec CM	0.413	0.518	258.4	1.58	3	-0.008
Word2Vec ST	0.415	0.750	259.5	1.58	2	-0.096

Table 2.3: Performance of the configuration model (CM) and the Preferential Growth model of Steyvers and Tenenbaum (ST). Clustering coefficients and assorativity of degree are difficult for the random models to capture.

related to the emerging networks during the course of acquisition. Our predictive models utilize the emerging network structure based on the words that a child knows at a particular time in development. To understand how network structure changes over the course of development, we consider some basic summary statistics for the **induced sub-graph**, or subset of known words and edges between those words, of an individual child's current vocabulary.

In Figure 2.4, we plot the density, clustering coefficient, % of vocabulary in the giant component and the assortativity coefficient of the vocabulary network as a sorted by the child's age (top) or CDI percentile (bottom). Note age and percentile are highly correlated since age and vocabulary size jointly determine percentile. At a high level, the age plots and the percentile plots look quite similar but some interesting differences in network structure can be seen. For example, we see that the giant component is generally slow to emerge in the Nelson network (due to low density edges), but is slower for children with low percentile. We also see that the Nelson network has higher clustering in low percentile snapshots, a result that cannot be fully explained by age alone. In the Howell network, the lowest percentile snapshots have less density than we'd expect considering age. Finally, assortativity scores are lower for snapshots of late talkers (low percentile) as compared to their typically developing peers. These differences could suggest that late talkers or younger children may learn differently. However, it may instead suggest that network models do not have enough structure to predict learning early in development. We return to this finding more in the discussion.

2.10 Network growth modeling

We now turn to performance when using our network growth models to capture the words a child is likely to learn next. Recall that we consider the usefulness of the network representation as well as the role of different growth processes. We finally consider the definition of the word "importance" in the graph by considering network centrality measures (cf. Table 2.1). In total, we consider 84 combinations of network, growth process and centrality as well as seven network growth baseline models that assume a fully disconnected network. Due to the large amount of possible models, we do not consider each model individually, instead we mainly discuss results at levels of

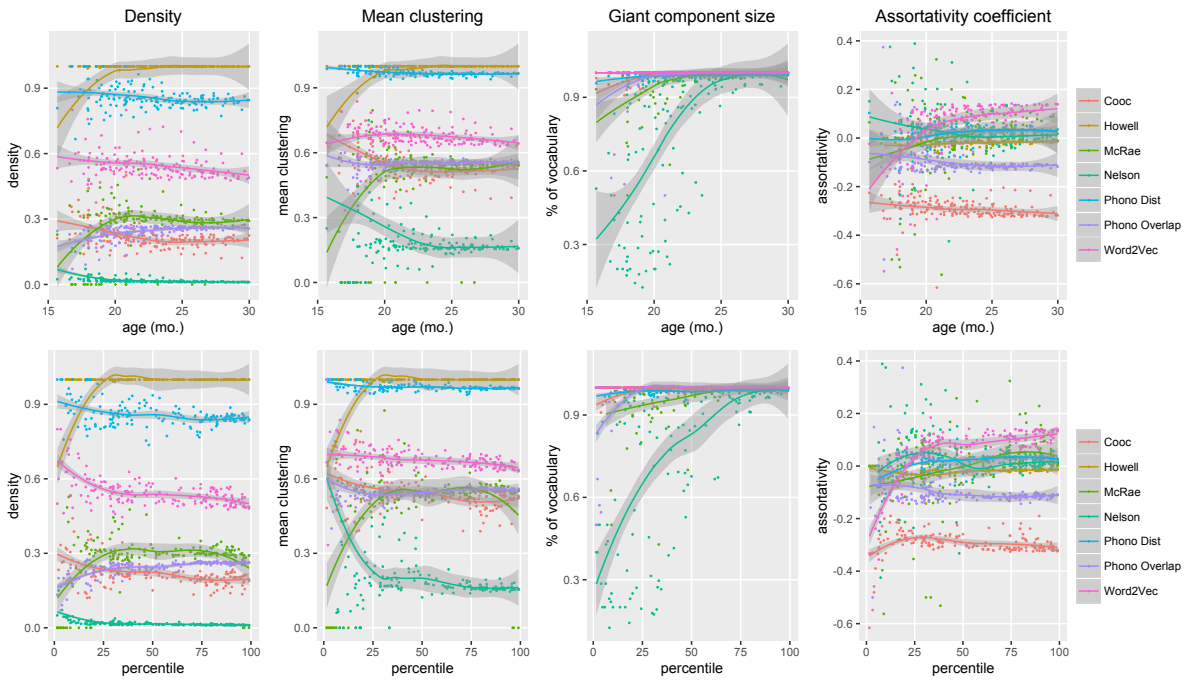

Figure 2.4: Plots of network measures as a function of the child's age (top) or vocabulary size (bottom) for validation snapshots. Different network representations (colored) indicate that the snapshot structure varies.

analysis mentioned above. We analyze performance of each network representation (macro level) in terms of predictive accuracy presuming specific growth mechanisms (mezzo level) and centrality measures (micro level).

Figure 2.5 indicates model performance of network variants organized by growth mechanism (top row) and centrality measures (bottom). Plotted along the y-axis is the average nLLK difference of each prediction between a given network model and the paired network baseline model. The paired network baseline model is defined by assuming that the underlying network is fully disconnected and thus each word within a snapshot is equally likely to be learned[2] . This model still undergoes standardization such that across snapshots, some words are more likely to be learned than others. The slope and intercept parameters are optimized for the network baseline models as for the network models. Positive values in figure 2.5 indicate the model outperforms the network specific baseline model. Note that a specific model can perform worse than our network baseline because we estimate parameters on training data but report on validation performance. The validation data, as well as the test data, are snapshots for unseen children. Performance worse than network baseline may suggest overfitting or high parameter sensitivity. Though different models predict a different number of words, the average error for each prediction, indicated along the y-axis, is roughly the same across all models. The order of the models along the x-axis is simply a ranking, based on either growth mechanism (top) or centrality (bottom), within a particular network representation.

Analyzing this figure and the matching nLLK scores, we find that certain network representations perform quite poorly, and occasionally below the network specific baseline model. This is an important finding as it shows that these network models are not always expressive enough to capture acquisition trends. Failure of certain models may suggest that features relevant to early language learners are not present in a specific network representation. It also suggests that we have not found the correct way of quantifying relationships that impact acquisition trajectories of young children, influencing which specific words are likely to be learned next. We further reduce

[2] Most network models, including the network's baselines, outperform the CDI age model. We choose a fully disconnected network baseline because it is a stronger test.

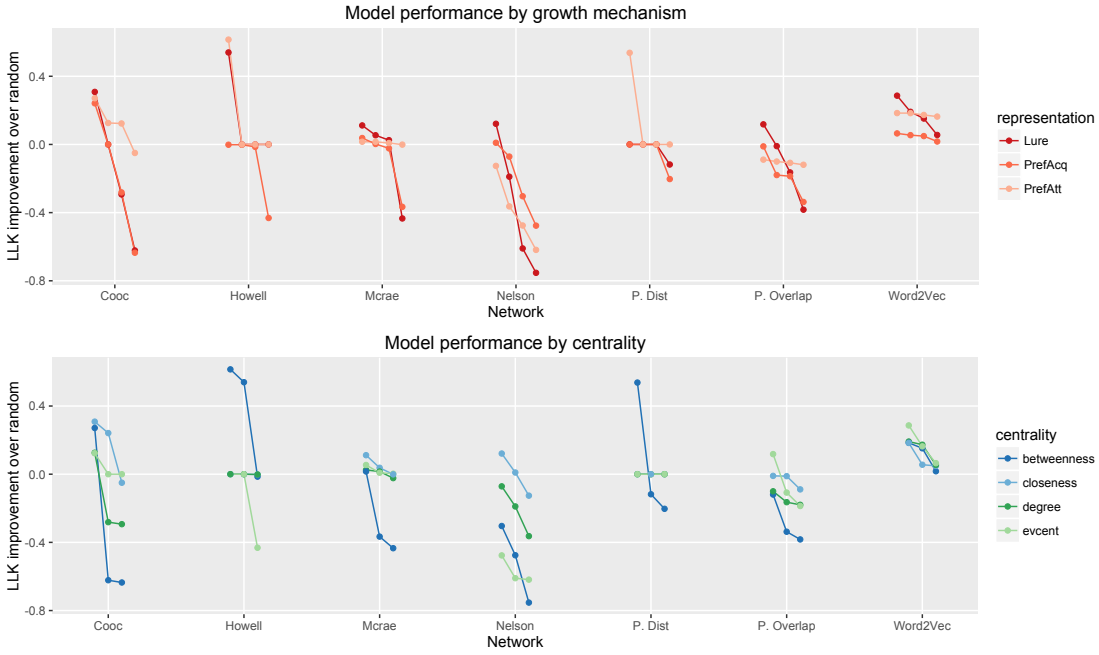

Figure 2.5: Average log-likelihood of predictions compared to random. Performance is clustered by growth mechanism (top) or centrality (bottom). Positive y-values indicate improvement over random, position along the x-axis is for better legibility and not reflective of performance.

the set of models we consider by only considering models that outperform the network baseline model on a paired t-test of individual snapshot predictions. Using Scheffe's correction for multiple comparisons, we exclude 20 models for not being statistically different than the random model in predictive accuracy of individual snapshots. This is in addition to the other 44 models that were excluded by performing worse than the network baseline model in terms of average nLLK.

Figure 2.5 and the significance test results show that the McRae feature network and the Phonological Overlap representations perform very near their network baseline counterpart, regardless of the growth model or centrality measure used. The failure of the McRae feature norms to account for growth trajectories has been previously reported (Hills et al., 2009b, 2009a). The network based on phonological overlap also fails to yield a model that reliably outperforms random, but other phonological representations do provide predictive accuracy. This suggests that the definition of a network impacts our ability to model acquisition trends and to predict specific words that a child will learn next. This also supports findings suggesting children have access to some types of semantic and phonological information but not to others early in language learning. The co-occurrence models, while on average they perform better than the network baseline, do not outperform the baseline in a paired t-test of snapshots. This suggests that the average performance improvement for this particular model is due to a small number of snapshots as opposed to general predictive accuracy. We leave these representations behind in terms of predictive models and consider it future work to compare these networks to networks that increase predictive ability.

Considering the mezzo level, or the different mechanisms of growth, rarely do we see **Preferential Acquisition** outperforming the network baseline model, suggesting that it is not a useful mechanism for predicting the vocabulary growth of an *individual* child. Recall that Preferential Acquisition assumes that words are learned proportionally to their centrality in the full vocabulary graph. Previous results found this mechanism of Preferential Acquisition (using degree centrality) to be the most accurate model when accounting for normative acquisition trends (Hills et al., 2009b)[3] . The failure here to account for individual vocabulary growth is likely because this mechanism does

[3] We confirm this finding using our models on normative acquisition, replicating their results.

not adapt well to individual differences. The poor performance of this growth model suggests two main findings 1) normative acquisition is quite different than the acquisition of any particular child, and 2) the content of the child's vocabulary is important and predictive of what words a child is likely to learn next.

Closer examination of the remaining 24 models which reliably outperform their network baseline counterpart suggests not only that the content of the child's vocabulary matters, but network density also affects model performance. **Lure of the Associates** considers the connectivity of the new word as related to connectivity from known words, whereas **Preferential Growth** considers only the content of the child's vocabulary. We find that Preferential Growth dominates in predictive accuracy for dense network representations, whereas Lure of the Associates better captures the networks that have less connectivity. The influence of the known vocabulary could reflect systematic structure in the child's environment, parent's child-directed speech, or individual interests of the child. It is clear that knowing about the child's current productive ability is useful in predicting the words the child will learn next.

We now consider the effects of various definitions of node importance. We approximate importance of a node by centrality measures of degree, betweenness, closeness, and eigenvector centrality. Figure 2.5 shows that, for certain network representations, only one type of centrality performs better than random. For example, predictive models for Howell require **betweenness** centrality. But the predictive accuracy for the Word2Vec representations does not vary much based on which centrality measure is used. The Nelson model shows a clear change in performance when different centrality measures are used and we can even rank which centrality measures are best for this representation. Still no clear centrality measure is the best summary of word importance across all the growth models. Centrality even varies within a network representation based on the growth model used. This suggests that the definition we give to word importance (e.g. centrality) is interconnected.

Collectively, the role of network centrality suggests that *global* centrality measures – measures determined and influenced by the structure of the full graph rather than local neighbors – are much

more accurate than local measures – measures determined only by local or immediate neighbors. In the case of global measures, the addition of a node may change the centrality measures of some or all of the other nodes in the graph. The most local centrality measure we consider is **degree** whereas the most global measure is **betweenness**. Eigenvector and closeness centrality emphasize a combination of local and global features. Global centrality measures particularly affect Lure of the Associates as this model selects words for learning that are most likely to alter connectivity of the known vocabulary structure the most. Even though the results do not clearly explain the role of centrality, the chosen definition of centrality does affect performance. Predictive performance benefits from access to global connectivity measures such as betweenness and closeness, regardless of the network representation.

2.11 Predictive accuracy on unseen children

For evaluation on the testing data we select one model from each of the representations where the network baseline model was outperformed. This includes Howell, Nelson, Phono Dist., and Word2Vec representations. Final models for each of the four network representations, listed in Table 2.4, were the best performing models when the estimated parameters were extended to the validation set. We evaluated performance based on accuracy on all snapshots of a given child rather than considering each vocabulary snapshot as independent. In the case of the Phono Dist representation and the Nelson representation, only one combination of growth algorithm and centrality resulted in performance above network baseline for the majority of the individual children. For Word2Vec and Howell, multiple models preformed quite similarly. In this case, the best fitting models were chosen based on the sensitivity of the slope and intercept parameters of the logistic transform to perturbation. After model selection, the logistic transform mapping the network **growth values** to probabilities was optimized using the combined training and validation data. We then optimized the weighting of the CDI AoA age-specific baseline model, fitting a linear combination of network predictions and normative acquisition trends as discussed in section 1.3. This last step allows all models to predict all words in an individual CDI snapshot, as opposed to just

Net	Growth	centrality	p	t	β	x_0	CDI w	nLLK	nb nLLK
Howell	Pref G.	betw.	.028	.320	81.01	.023	.253	.390	.426
Nelson	Lure	close	.011	.011	129.70	.014	.331	.381	.413
Phono Dist	Pref G.	betw.	.020	.100	197.95	.009	.332	.367	.402
Word2Vec	Lure	betw.	.040	.011	147.25	.013	.301	.377	.415

Table 2.4: Best performing models on each of four network representations. All models reliably outperform random when applied to validation data and when extended to the test set as reported (p-value, abv. p). The network threshold (abv. t), the fitted logistic transform (β, x_0) and the influence of the CDI AoA age-specific baseline are reported. Also reported is the nLLK of the model and the network baseline (nb nLLK).

the words in a specific network representation. The threshold to convert the network representation to a binary network was based on the original training data only. The resulting average nLLK of the test set is shown in Table 2.4. Also shown is the contribution of the CDI AoA age-specific baseline model and the nLLK error of the network specific baseline model, abbreviated as nb nLLK. The optimized parameters show a positive correlation (as seen by a positive slope) between the centrality of a word and the probability that the word is learned. Words that have a higher growth value, are those words that are more likely to be learned by the child.

Table 2.4 indicates that these trained models are performing better than the network baseline model. We also note that this model outperforms the CDI age-specific model alone. The CDI norms resulted in a model with a nLLK score of 0.45. We compare performance of the individual network based models to each other in table 2.5. Here we present the average nLLK scores on the test set. This measure alone does not tell us about performance on individual snapshots since for certain snapshots more words are known initially. Averaging over all words weighs those snapshots with larger initial vocabularies less. In table 2.5, column 4, we report the percent overlap between the k words learned by a child in each snapshot and the k' most likely words to be learned by the model. We also report the area under the curve, accuracy using a threshold where the number of learning events is equal to the observed data, as well as the d-prime score. Finally we report a paired t-test computing the difference in performance between the network model and the network baseline model.

model	$\|V\|$	nLLK	% overlap	AUC	acc	d-prime	t-stat.
Howell	355	.390	31.0	.728	.807	.110	114.1
Nelson	534	.381	29.8	.734	.814	.113	14.6
Phono Dist	677	.367	35.0	.765	.823	.122	12.6
Word2Vec	626	.377	32.9	.741	.813	.114	57.9

Table 2.5: Evaluation of model performance based on a variety of measures. Also reported is the number of words in the network representation (sz) and the weight of the network representation in the final predictions. t-stat. is based on a significance test between the network representation and the network specific baseline.

These results show that our network growth models are capable of predicting future word acquisition of individual children with higher accuracy than our network baseline models. We also see that the nature of the network representation can affect our ability to outperform the network baseline model. We find strong evidence for an influence of the child's current vocabulary on the accuracy of our future predictions as captured by the improved performance of the **Lure of the Associates** and **Preferential Growth** models. Global centrality measures also provide an increase in performance. Collectively this suggests that language acquisition benefits from the inclusion of interactions among words.

2.12 Modeling development

Beyond simply predicting future language learning, we are interested in a developmental perspective. It is possible that these models are more accurate during certain periods in development. To consider this, we order the predictions by specific features we believe might be relevant for learning. We then compute a smoothed average nLLK and compare performance. A value of 0 indicates performance equal to the CDI AoA age-specific baseline model. Negative values indicate improvement of our models over the AoA baseline. Plotted in figure 2.6 we show performance differences as compared to the CDI AoA age-specific baseline. We construct orders based on the 1) age of the child at time of prediction (top left), 2) vocabulary size of the child when the first CDI is collected (top right), 3) CDI percentile of the child at the initial CDI of the snapshot and 4)

average age a word is learned (AoA, average age of acquisition) as calculated based on normative acquisition trends. The colored bar indicates the specific model that is the best performing model at that point in development. Interestingly, we note that the network models, while differing slightly in their ability to predict, have general consensus on snapshots that are easy (low nLLK) or hard (high nLLK) to predict.

Figure 2.6 suggests interesting developmental effects. First, we note that there are certain snapshots in which the CDI AoA age-specific baseline model outperforms our network models. This could be due to the variability in learning trajectories or the types of learners in our sample. It's also possible that during specific months, normative learning is more likely to align with individual learning trajectories resulting in accurate CDI baseline predictions. More hopeful is that the network models outperform the CDI age of acquisition models for the population of children we are most interested in modeling, specifically children with a low CDI percentile. We also find that the network models are better than the age of acquisition norms for very small CDI vocabularies and very large CDI vocabularies. The network models generally perform well for children between 22 and 26 months, coinciding with the ages where we have the largest amount of data. Here we can see some differences between individual model performance as well. Generally, the phonological Levenshtein distance (Phono Dist) model is the most accurate, but for snapshots of older children, a semantic representation is better. In the future we hope to understand what exactly increases our predictability over the CDI AoA age-specific baseline model, specifically for late talking children, a population that has been challenging to model.

2.13 Ensemble models

Finally, we consider ensemble models to improve accuracy of prediction and to assess the importance of each model. Our first ensemble model weighs each network prediction equally averaging over all predictions from our 4 best performing models. We call this our **Average Ensemble** model. Our second model, instead, optimizes the weight of each of the four network models. We optimize the parameters of our **Weighted Ensemble** to minimize training error. We

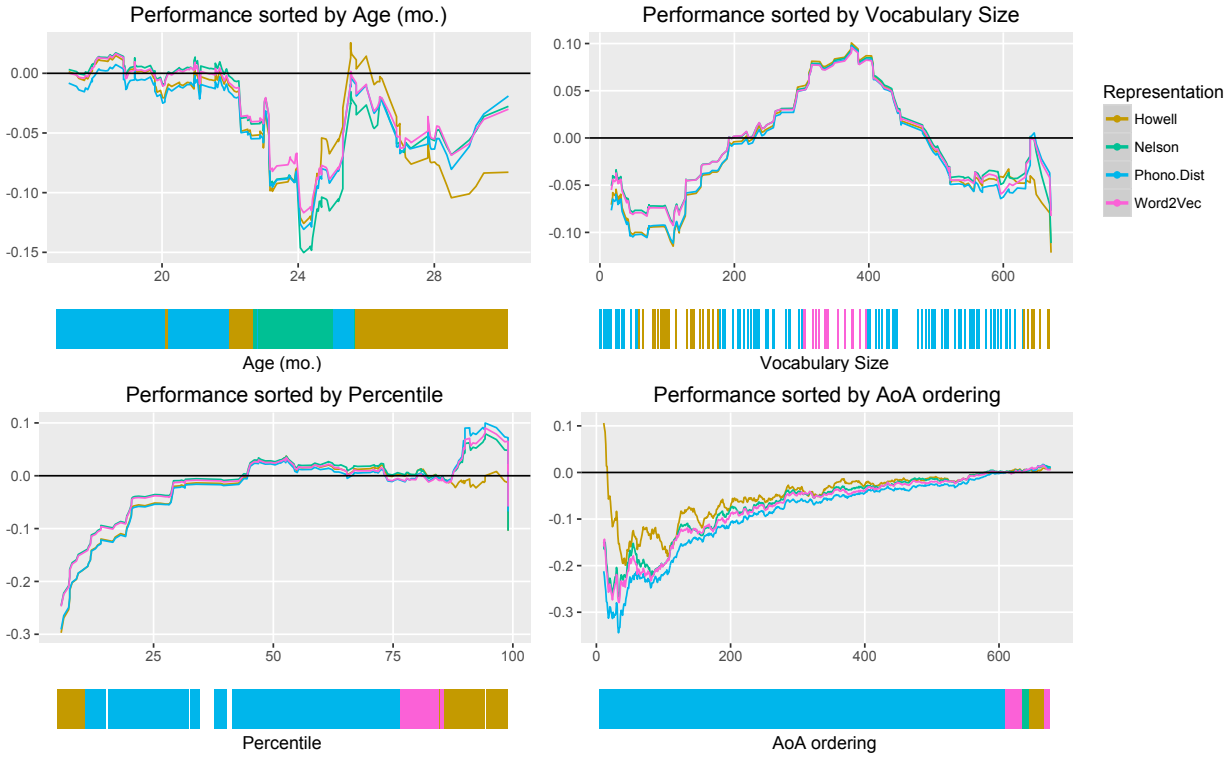

Figure 2.6: We consider performance of the best performing network growth models as related to the child's age (top left), vocabulary size (top right), CDI percentile (bottom left) or average age of acquisition (AoA, bottom right). The line plots indicate smoothed negative log-likelihood, the colored bars indicate the best performing model at that moment in time.

also consider an ensemble model that is a simple **Word Voting** model, where we select the best performing model on each word. For each prediction made to an unseen child, we consider the best performing model for that word given our training data.

We also consider a voting model based on child specific features. We consider the features of age, CDI percentile, and vocabulary size. In this model, we select the best performing model for each individual in the training set. When generalizing to an unseen individual, we pick a training snapshot that is most similar to the child specific features of the new snapshot. We then use the best performing model for the chosen snapshot to predict the unseen snapshot. We consider similarity based on a single feature such as age (**Child Vote Age**), percentile (**%**), or vocabulary size (**voc.**). Additionally, we consider a model that determines similarity based on the euclidean distance of all features simultaneously **Child Vote Avg**.

model	nLLK	% overlap	AUC	acc	*d*-prime
Avg. Ensemble	.371	33.2	.863	.811	.129
Wgt. Ensemble	.367	35.9	.873	.818	.136
Word Vote	.371	31.9	.847	.799	.117
Child Vote (age)	.374	33.5	.749	.811	.132
Child Vote (voc.)	.370	33.3	.755	.813	.134
Child Vote (%)	.367	33.1	.757	.818	.135
Child Vote Avg	.366	33.3	.761	.815	.135

Table 2.6: Validation performance on a variety of ensemble models.

Table 2.6 indicates performance of these ensemble models, when applied to the validation set. The results suggest that the weighted ensemble model performs the best. We extend these three models to the testing set, using the combined training and validation sets for the child voting models, see table 2.7. The trained weighting of the **Wgt. Ensemble** heavily considers the Phonological Levhenstein distance (Phono Dist) network representation with a weight of .87, the remaining weight is given to Howell. This model completely ignores the Nelson and Word2Vec representations, suggesting that the relational features captured in the Howell sensory motor features and phonological similarity are those relationships that can best explain network growth. The child

voting model, considering similarity based on **Percentile**, performs slightly better than the other child voting models, suggesting that percentile is the measure of child language that is the most informative as to what words a child is likely to learn next within network growth models.

model	nLLK	% overlap	AUC	acc	*d*-prime
Wgt. Ensemble	.363	31.5	0.738	.822	.124
Child Vote (%)	.368	30.3	0.732	.818	.123
Child Vote Avg	.368	30.1	0.733	.818	.122

Table 2.7: Performance on test set for best performing ensemble models.

2.14 Discussion and future direction

The network-based approach to modeling acquisition provides a unifying framework for studying the complex process of acquisition. We find that the definition of the edges in a network dramatically affects our ability to predict future acquisition. In fact, under certain types of network definitions – such as a feature overlap network, a network based on co-occurrence in child directed speech and under some measures of phonology – our network growth approach cannot account for acquisition trends. Even when networks are able to increase our ability to predict the learning of new words, we found evidence, at particular points in development, that one network representation is clearly more accurate. The Phonological Edit Distance (**Phono Dist**) network captures acquisition trajectories of younger children and children with lower percentiles. Later in development, for children with higher percentiles, we find instead that the **Howell** network representation accounts for language acquisition trends with higher accuracy. We also found that at certain points in development the CDI AoA age-specific baseline model outperformed our network base approach. This suggests that there are attentional changes during the course of learning and, potentially, that later talkers or younger children learn differently than their peers whereas typically developing children can be well captured with the CDI AoA age-specific baseline results. This type of network modeling framework may allow for us to model differences in these groups and to go further by explaining the process of acquisition that leads to these differences.

The results suggest that phonological and semantic features are both important and relevant to language learning. While phonological network structure varies greatly from semantic network structure (Vitevitch, 2008; Gruenenfelder & Pisoni, 2009; Stella & Brede, 2015), both of these models are useful in predictive models of language in the domain of lexical acquisition. In the future we hope to jointly consider the effects of these representations in predictive modeling by using a multiplex network approach.

In this work, we find strong evidence that some of our network models cannot account for acquisition trends above our network baseline model. Part of this is due to the flexibility of our definition of a network baseline model, it also suggests the need for network scientists to carefully consider the definition of language networks. Not all networks can capture the cognitive and learning aspects related to language. Specifically, we found that **Co-Occurrence** networks are not predictive at the level of individual words. We also found the **McRae** feature norms build a network representation that captures relationships that are either not accessible to young children or not the most meaningful relationships.

We found a strong interaction of performance within the mezzo or process level. In the case of normative language acquisition modeling, it has been shown that **Preferential Acquisition** outperforms models based on the child's current vocabulary knowledge. Here, however, we found that a model of Preferential Acquisition is unable to account for individual language growth. We think this might be due to the individual differences and specifically that early and late talking children learn using different mechanisms. All accurate predictive models of individual child trajectories use information about the child's productive vocabulary. We found strong evidence for two different growth processes–**Preferential Growth** and **Lure of the Associates**. Specifically, Lure of the Associates is the predominant model when the language network is large enough to have informative structure, whereas Preferential Growth is useful for modeling early language learners or language learners that are slower than their peers. The importance of the child's productive vocabulary in predicting future language learning confirms much of the earlier findings (e.g. Sandhofer et al., 2000; Weizman & Snow, 2001; Marchman & Fernald, 2008; DeLoache et al., 2007) suggesting that

individual differences in learning can be related to environmental and preference differences. While our models are ambiguous to the cognitive and environmental underpinnings influencing why a child learns a specific word, the words themselves may have important and useful cues as to relevant features that influence a learning trajectory. These cues may be related to the physical and linguistic environment that the child is learning in or to the child's specific interests. Either way, our modeling results strongly suggest an important contribution of the known vocabulary on future vocabulary growth.

In the best fitting models for each network representation, global centrality measures perform with the highest accuracy. This may suggest that the global structure of the emerging language graph is most important for supporting future language learning as opposed to just immediate neighbors of the learned word. This may be especially important for correcting language learning delays such as those of **late talking** children. The improvement of global centrality measures over local ones suggests that instead of teaching just a few words to help get children back on track, we may need to alter the connectivity of the lexical graph instead. We also note that we only consider network centrality measures here but there are other ways of quantifying a node's importance. In the future we hope to combine centrality measures with other non-network related measures such as frequency or concreteness. While network connectivity seems to be useful in modeling language acquisition, other measures may provide additional support and insight.

The complexity of these network modeling results suggests the depth of the challenge that comes with modeling individual acquisition trends. While this is a first step in building up an accurate predictive network growth model, much more work is needed to explain why certain models perform in disparate ways, for different words and different types of individual language learners. In the future, with more data and more sophisticated network models, we hope to capture language learning with higher accuracy. Improving accuracy will also allow for the exploration of mechanistic models to understand why certain children have a specific language acquisition trajectory. We are particularly encouraged by the strong improvement of the network model over the CDI AoA age-specific baseline model specifically for individual children who are learning language at a slower

rate. While the accuracy in predicting future acquisition is important, one benefit of network analysis models over machine learning models is that we have a possible mechanism of learning to explore in terms of improving the vocabulary structure of late talking children. In the future, we hope this work and other similar models can pave the way for diagnostic and intervention tools capable not only of modeling, but also explaining and maybe even influencing acquisition trends of individual children.

Chapter 3

Neural Network models predicting individual word learning in young children

Neural network models, often called **connectionist models**, provide a way of extending findings of observational and behavioral studies by capturing and capitalizing on relevant relationships that may guide and influence development or learning. As statistical learning tools, neural network models are powerful and adaptive. The models can model sequential patterns as well as uncertainty found in data. One challenge for neural networks is that it can be difficult to model noisy and small data sets. However many important cognitive questions rely on learning from sparse and/or few examples. For example, even young children can do "one-shot learning," acquiring a new word from only a single example.

Here we use neural networks as statistical tools to explore features of language acquisition trajectories. Insight form these neural network models will allow for prediction of what words a child is likely to learn next. Additionally comparing different models, we will gain theoretical insights in to the learning of young children. Language learning is one of the first complex cognitive and linguistic tasks humans undertake. Infants start producing their earliest words around 12 months of age. Within only a few months, young children have hundreds of words. Shortly thereafter, toddlers begin to construct sentences with complex ideas, and grammatical structure. This process of learning is both surprisingly quick and startlingly complex. Despite how quickly this learning comes online, much of the language acquisition process is still challenging to explain. The approach of machine learning and data science to model complex processes, such as acquisition, can provide novel insight into how young children learn. Pairing powerful statistical learning tools with observational

acquisition data, we can model how the child's current vocabulary relates to their vocabulary at some point in the future. This knowledge may help us isolate differences and variability in individual learning early in acquisition.

When assessing language ability, experimental psychologists often compute a "CDI percentile" which considers the number of words the child knows, given the child's age and sex, as compared to their peers (Dale & Fenson, 1996; Fenson et al., 1994; Thal et al., 1999). Using a vocabulary assessment form known as the MacArthur-Bates Communicative Development Inventory (CDI), parents indicate which of a fixed set of words their child produces. The CDI 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 catch and monitor because many of them will continue to have perpetual language learning difficulties (Thal et al., 1999; Heilmann et al., 2005). Sometimes these language-specific difficulties will develop into wider reaching cognitive deficits. Predictive models of acquisition could help with diagnostic assessment, and potentially suggest approaches for correcting differences, between at-risk children and their normally developing peers. However, before these questions can be directly explored, a working predictive model of acquisition must be constructed and studied. Here we consider a connectionist or neural network modeling approach to predict the words a child is likely to learn given information about the child's current vocabulary.

3.1 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 (Christiansen & Chater, 2001; McClelland et al., 2010) and a specific review of semantic development here (Sims et al., 2012). Work has also used neural network approaches to link neuroscience to early development (Munakata & Stedron, 2001) and to semantic cognition (McClelland & Rogers, 2003). Our goal, however, is one of predictive accuracy not cognitive plausibility. Nonetheless, we quickly review neural networks as cognitive models here because successes in this area motivated the direction and scope of the current work.

Much of the past work on neural network models of lexical acquisition captures performance of infants on behavioral learning tasks. These neural network models tend to offer a mechanistic explanation of language learning in children. One such behavioral phenomena that benefits from models of neural networks is the emergence of the **shape and material bias** (Landau et al., 1988). Earlier learning of the preference to generalize solid objects based on shape, and non-solid objects based on material has been shown to increase vocabulary growth in toddlers, specifically for nouns exhibiting this shape bias (L. B. Smith, 2000; Gershkoff-Stowe & Smith, 2004). Work using neural network models tasked with learning word-to-object mappings, were able to replicate the shape bias and the influence of this bias on future lexical learning (Colunga & Smith, 2004, 2005).

For modeling shape bias emergence, the child's current vocabulary (specifically the nouns) are used as input to the neural network. These nouns are represented sequentially to the model as a vector of features, including information pertaining to the shape, material, and solidity of each noun. The network is trained to learn a mapping between the vector representation of the noun and the object label (word). During testing, the presence or absence of a shape bias is probed. Given that the neural network model learns incrementally and is underspecified early in learning, the authors show the presences of a shape and material bias only after significant training. Probing the neural network with word recall, the authors are able to mimic the toddler behavioral results tied to the shape bias as well. Extensions of this work are also able capture the **emergence** of these shape and material biases (Colunga & Sims, 2012; Sims et al., 2013a; Sims, Schilling, & Colunga, 2013b). These modeling result are then used to implement and design experiments around capturing and understanding the emergence of learning biases in young learners (Colunga & Smith, 2005). For example the models suggest that children gradually acquire the bias rather than undergo a "conceptual shift," characterized by a period of transition rather than sudden change in generalization behavior.

Other examples of neural network models applied to modeling early lexical acquisition include models capable of capturing age of acquisition effects (Li, Farkas, & MacWhinney, 2004), and the formation and degradation of conceptual categories (McClelland & Rogers, 2003). These, and other

examples of lexical development, explain behavior with a basic mechanistic account of associative learning. But this simple mechanistic account of associative learning often provides added insight into cognitive behavior. One neural network model, which learned to map word-forms to object referents (McMurray, Horst, & Samuelson, 2012) showed a mutual exclusivity bias – a preference for novel words to map to novel objects (Merriman, Bowman, & MacWhinney, 1989) – even though no training instances explicitly exhibited this bias. The model used this bias and its "knowledge" of other words to quickly and accurately learn new words even in highly ambiguous contexts. Another neural network, modeling acquisition of categories, showed evidence of a feedback loop between perceptual features and linguistic labels which supported generalization of categories (Yu, Ballard, & Aslin, 2005). The model itself was able to use the relationship between category and language learning to provide structure and reinforcement during learning.

Unlike the work reviewed above, we do not focus on cognitive interpretability here. We choose not to focus on studying the power of associative learning to acquire knowledge, but rather to focus on the ability of neural networks to uncover patterns evident in learning trajectories. This approach may provide novel insights and hypotheses as to how the language learning progresses for an individual child. Here we use neural network models, not as a proposed process of learning, but as a means to uncover associations in the environment that might be relevant and even facilitatory to lexical acquisition. 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. This is mostly due to the ability of connectionist models to incrementally learn and to still find predictive capacity even when representations are underdetermined or noisy. Accuracy of predicting the individual words a specific child is likely to learn next may benefit from this type of approach. By understanding how to best predict future word learning, we may additionally be able to capture types of associations that are relevant to early lexical acquisition.

A key idea to data-driven neural network models of acquisition is that there are strong similarities among the way in which children learn, but the differences are also important and predictive. If all children learn similarly, then high-level features such as the CDI percentile should

be adequate in classifying the language learning ability of children. But if there is variability among learners that can be assessed from the vocabulary data directly, then the data-driven approach can offer unique insight into these trends. There is a large amount of work suggesting there are different types of learners (e.g. Thal et al., 1999; Sandhofer et al., 2000; Heilmann et al., 2005; Mayor & Plunkett, 2011). For example, network analysis approaches have found that not only are late talkers learning slower than their peers, the resulting vocabulary is less structured than one might expect if the children were simply learning at a slower rate (Beckage et al., 2011). 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 can provide a powerful and predictive tool to aid with classification and diagnostics of a child's learning trajectory.

Here we focus specifically on using the child's vocabulary to predict future lexical learning. Experimental and theoretical work from developmental psychology documents the effect of cognitive, social, and environmental aspects that affect learning. We aim to build a model that considers the child's current vocabulary to predict future learning, assuming that a neural network model can use the structure of the current vocabulary to extract relevant aspects of the linguistic, social and environmental structure. We know from past work that certain systematic trends alter learning trajectories. For example, nouns are learned earlier than other syntactic classes (e.g. Gentner, 1982). But individual variation may also be relevant for capturing future acquisition. The child's environment – both the social (e.g. Arriaga et al., 1998) and language (e.g. Weizman & Snow, 2001) environment – influence when and what words the child is likely to learn. Differences of individual children, such as their cognitive ability (e.g. Marchman & Fernald, 2008) and their individual interests (e.g. DeLoache et al., 2007) also affect language learning. A highly predictive model will have to represent acquisition differences and be able to recognize different characteristic trajectories of acquisition. Further, the model will need to assess what trajectory a particular child is on based on the child's vocabulary alone.

With the goal of capturing the role of a child's current vocabulary on future language learning, we explore a variety of representations of the child's current vocabulary knowledge. If all children

learn similarly, then population features should be both informative and predictive of what words the child is likely to learn approximately one month into the future. Alternatively, if children's interests in specific themes (for example, animals) drive learning, then knowing the individual words of the vocabulary will be helpful in capturing future lexical acquisition. Many features of the language environment could 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 what specific features direct learning, we do assume that representations accentuating relevant features will result in an improvement in model accuracy. We consider the performance of various language representations compared to the baseline child-feature model to quantify the influence of certain language representations on predictions.

3.2 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 recruitment phases. Parents and children came to the lab for recurrent visits over 12 consecutive months. Visits were timed at near monthly intervals and, on average, we have 10.9 visits for each child. Overall, we included 83 (37 female) monolingual children 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**, Dale & Fenson, 1996) for children between 16 and 30 months. The report included 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. We define a CDI **snapshot** as a pair of input features and output vocabulary, used to train our neural network model. Across all recruitment phases we have a total of 908 CDI checklists which form 825 CDI vocabulary **snapshots**. Figure 3.1 includes an example of what the data look like. We are particularly interested in predicting learning moments–the time when a child goes from not producing to producing a specific word. In all cases, our model is given information pertaining to

	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

Figure 3.1: 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.

the content of the first CDI in the snapshot and is tasked with predicting the productive 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 the neural network models.

The longitudinal study represents many different stages of early 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 a child across all the CDIs is 22.4 months. We also have a full range of language ability represented. To approximate language ability, we utilize the CDI **percentile** measure. This measure is calculated based on the size of a child's productive vocabulary as compared to the child's same-sex 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 should note that recruitment of participants in the longitudinal study was biased to over-represent **late talkers**, or children who are learning slower than their peers, as late-talkers are a population of particular interest in language acquisition research.

3.3 Neural network training

Neural networks were constructed and fit using Torch7, a scientific computing framework for LuaJIT. Models have a single hidden layer, optimized in size for each trained network. Model were trained using adaptive stochastic gradient descent as defined in equation 3.2, with a batch size of b. Y indicates the observations, $\hat{Y}$ the estimates and X the input data. $\sigma(z_j)$ indicates the logistic transform on the activation of layer j. We also add weight decay (α_d) and momentum (m)

as indicated in equation 3.3 and 3.4 respectively. Our final model is shown in equation 3.4. Initial learning rate ($\alpha^{(0)}$), learning rate decay (α_d), momentum (m), batch size (b) and the number of hidden units were optimized for each input representation.

$$C(Y,\theta) = \mathcal{L}(Y,\hat{Y}) = \frac{1}{n}\sum_{i=1}^{n} y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i)$$

$$\Delta W = \sum_j \frac{\partial C}{\partial w_j} = \frac{1}{n}\sum_{i=1}^{n} x_i(\sigma(z_j) - y) \tag{3.1}$$

$$\theta^{(s+1)} = \theta^{(s)} - \frac{1}{b}\sum_{i=1}^{b} \alpha \Delta W \tag{3.2}$$

$$\alpha^{(s)} = \frac{\alpha^{(0)}}{1 - s\alpha_d} \tag{3.3}$$

$$\theta^{(s)} = \frac{1}{b}\sum_{i=1}^{b} \alpha^(s)\Delta W + m * \theta^{(s-1)} \tag{3.4}$$

Learning rate decay (α_d, equation 3.3) allows network models to quickly learn weights to capture initial associations, but also adapt later in training to more nuanced patterns. It also allows for combining training and validation data without the need to consider stopping time. Momentum (m, equation 3.4) defines each update as a combination of the current gradient and the gradient previous time steps. Dropout was employed on the hidden unit at a rate of 0.5; during training only about half of the hidden units were used for prediction. During validation and test, all hidden units were available.

We discuss the input features below in more detail but the output layer of all models is a vector of size 677, indicating how certain the model is that a particular word is learned or known by the child about one month in the future. Negative log-likelihood (nLLK) was the training criterion and validation metric. Only words that were reported as unknown at the start of the snapshot contribute error to the model. This ensures that the model is not penalized for incorrectly predicting that known words stay known.

The neural networks in this paper have a large set of parameters that are individually optimized for each set of input features. Parameters were optimized based on repeatedly dividing the dataset

into a training, testing and validation set, using a 4-fold cross-validation scheme. The validation and test set each consisted of all vocabulary snapshots tied to a particular fold containing 20% of children, including about 170 snapshots each. Model performance is based not only on generalization to unseen vocabularies, but also to unseen children. Only after fixing all parameters, was the test set considered. The final set of models were run 10 times to compute variance in testing performance.

3.4 Predicting from the CDI forms

One goal of the paper is to understand if the content of a child's vocabulary improves our ability to predict words learned by a specific child. Much of the work in the developmental literature considers the size, rather than the content, of the child's vocabulary in assessing language ability. Here we instead try to leverage the particular words a child knows in order to determine the words the child will learn next. In all cases, we use a neural network model trained on a sample of children to predict an unseen child's individual word acquisition. We vary the nature of the input representation capturing the child's current language knowledge. Establishing an initial baseline of prediction, we quantify the information that is provided directly from the full CDI vocabulary checklist. The CDI form includes the individual words a child produces, as reported by a caregiver. In addition, there are features capturing developmental information of the child, such as the child's age and the sex of the child. We also consider measures of vocabulary ability as assessed by the CDI percentile and the child's vocabulary size.

We construct three types of input representations based on the CDI and the other related attributes measured as part of the CDI checklist. The first includes information related to the child only. In the **CDI child** feature model, vocabulary predictions are made using the age of the child (both at the time the CDI was completed and the age of prediction), language ability (as approximated by the CDI percentile), the child's sex, the number of sessions the child has attended in the lab, and the child's full vocabulary size. This child-feature model tests whether differences in acquisition can be accounted for simply by considering high-level features related to the learner. This model assumes that all children learn words in a similar order, just at different ages. This is a

reliable baseline model, as many vocabulary assessments use the child's age and vocabulary size as a proxy for cognitive language skills.

We extend the use of the CDI by building a predictive model based on the individual words the child currently produces. In this **CDI word** feature model, the input features to the neural network are a 677 unit vector with binary values indicating if the child reportedly produces the word or not. We use this input representation to understand how useful the individual child's vocabulary is when predicting acquisition trends through the course of development. It is not clear if this model will outperform the model which has access only to the child features, as there is much more individual variability in the words a specific child knows, and this may wash out learning signals. In fact, previous work using logistic regression models found that the child-features outperformed a model based on the individual words the child knows (Beckage, Mozer, & Colunga, 2015). 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 data 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 *wolf* is usually learned in relation to *dog*). Further, the neural network model 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 with less accuracy than a model that clusters words based on semantic categories. Note that in a strict machine learning context we could avoid this issue by training a deep neural network with an auto encoder. Here we instead trade interpretability and simplicity for such a highly flexible model.

We finally consider both word and child features in our **CDI combined model**. By first considering these features independently and then combining the features, we can quantify the

influence of child specific features as compared to features related to the child's current vocabulary. Previous work finds that an ensemble of independent CDI child features and CDI word features outperform a jointly trained model when logistic regression is used (Beckage, Mozer, & Colunga, 2015). By using a more highly parameterized neural network, we can again ask whether word features contribute information beyond the information of the child features alone, without the need of ensemble methods.

3.5 Vocabulary feature models

We believe that the individual word forms themselves may capture some features that are relevant. However, it seems unlikely that knowing or not knowing a specific word dramatically affects learning. Instead we believe that individual words are the effect of learning important aspects of language such as the relational, categorical, and semantic structure underlying specific words. We explore different language-informed representations of semantic knowledge to summarize the child's current linguistic and vocabulary knowledge and explore predictive accuracy using these various representations.. In this project we limit ourselves to the set of words in the child's productive vocabulary, but the child is learning in a very rich and structured physical and linguistic environment. To approximate the structure of language and the environment that may be relevant and accessible to early language learning, we construct word representations that emphasize aspects of the child's vocabulary we think may be useful in predicting future acquisition.

In practice, The child's productive vocabulary may relate to the latent features that are relevant to learning and language. To explore this idea, we represent each word of the CDI as a vector of features based on a specific feature representation of language. We then aggregate the specific words the child knows to approximate the linguistic knowledge a child has at that point in development. Consider concrete nouns; these words are some of the earliest words learned by young children. Semantic features are often the most salient descriptors of these nouns when adults describe them. It is less clear, however, that these descriptive semantic features are accessible to, and used by, young children to build vocabulary acumen or to make sense of the world around them.

By comparing model accuracy across various representations we may be able to approximate what features summarize the vocabulary growth of a specific child. This might provide insight and, a means to assess, whether a child currently has knowledge of and access to a particular aspect of language at a specific point in development.

If by emphasizing relevant semantic features to young children in our input representations we can improve accuracy of our neural network then we can use performance to assess which features are most relevant to young children. We consider four different semantic feature representations. One representation uses the sensory motor feature norms collected by Howell and colleagues (2005). In this study, undergraduate students were asked to rate, from the perspective of a child, whether a particular noun possessed a particular feature (Howell et al., 2005). In total, 97 features were considered for a total of 355 nouns. For each noun on the CDI, we have a vector indicating how much a specific noun possesses each feature. Features capture properties of the objects as well as object descriptors and include features such as "is large," "can fly," and "is made of metal." The Howell sensory motor features focus on aspects of words that may be salient and relevant to young children, including color, shape, and material. For example, alligator has a scariness rating of 5.5 whereas applesauce has a rating of 1.3. These features may provide a more informed input representation than simply the individual words on the CDI, allowing for better neural network generalization. We refer to this feature representation as **Howell**.

In addition to the Howell feature norms, we also considered the McRae feature norms (McRae et al., 2005). These norms, like the above norms, were collected based on adult participant response. The McRae feature-norming study (McRae et al., 2005) asked individuals to list features of concrete nouns in an open-ended response. Features were aggregated to capture general concepts such as taxonomic and encyclopedic features (e.g. taste, animacy, fact, description) (Barsalou, Simmons, Barbey, & Wilson, 2003). We use the McRae features (e.g. planes have wings) and the number of each type of feature (e.g. 4 taxonomic features) as a vector representation of each word. These individual word representations are aggregated and become the input to a neural network. This particular representation only overlaps with about 200 of the 677 CDI words, but we still evaluate

the model on the prediction to the whole set of CDI words. We call this representation **McRae**. Previous work has found this representation has minimal predictability in accounting for acquisition of young children using network analysis (Hills et al., 2009b, 2009a, 2010). Here we test the usefulness of this representation in training a neural network model.

The Nelson free association norms are the final semantic representation based on adult response that we consider. These norms are based on free response in an association task. The norms aim to capture the first semantically related word which comes to mind in response to specific cue words. Aggregated over more than 6000 participants and responding to more than 5000 cue words, we consider only free associations between CDI words. For each particular word, we consider the strength (e.g. normalized frequency of response) of that word to all other words (on the CDI) in the free association task. For example, the word *dog* is connected to the words *cat*, *bone* and *house* but the highest connection strength is between dog and cat. The connectivity and strength of free associations between words is related to age of acquisition (Nelson et al., 2004; Steyvers & Tenenbaum, 2005; Griffiths et al., 2007; De Deyne & Storms, 2008). These same free association norms may also provide systematic information relevant to early language learning in our neural network models. We call this representation **Nelson** [1] .

Word2Vec is a language representation that has been found useful in applications aimed at modeling adult language use and generation (Mikolov et al., 2013). We extend this representation to child language. Using the Word2Vec algorithm, which considers frequency of co-occurrence between words, we constructed a 200 dimensional vector representation of nearly all words in the CDI. We use a large GoogleNews corpus of more than 6 billion words to inform our representation. Vector representations of compound words, like *peanutbutter*, are the average of 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 (Mikolov et al., 2013). We consider this representation as input to the neural network under the assumption that it captures the complexity and relationships of the language

[1] Note, this is the same Nelson representation used in chapter 2

children must eventually learn. We call this input representation **Word2Vec**. Collectively, we refer to the Howell, McRae, Nelson and Word2Vec models as **semantic representations**, given that the primary source of information in these models comes from adult semantic information.

We also consider the phonemic composition of individual words. Past work shows the sounds of words play a significant role in learning (Stokes, Klee, Carson, & Carson, 2005; Storkel, 2004) and that computational models can capture this effect (Vitevitch & Storkel, 2013). 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. The joint vocabulary representation includes the distribution of phonemes in a child's productive speech. Research related to phonological importance in early learning suggests there is a strong effect of word onset and word rhyme (Goswami & Bryant, 1990) but other work has instead suggested phonemic awareness is a better predictor (Hulme et al., 2002). 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 **Phonology**.

On the CDI form itself, words are classified into 22 different linguistically informed groups, capturing semantic themes such as "animal", "body", and "people", and grammatical classes like "action words", "helping verbs", and "pronouns". 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. Success of this representation would indicate that the words a

child learns next is related to the collective categories of words the child knows. 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 (DeLoache et al., 2007), parent directed speech (Waxman & Leddon, 2002) or other features of the environment. We remain agnostic as to what aspects of learning might motivate accuracy of this model, testing instead whether or not this type of vocabulary representation can capture future language learning. We consider this to be the **CDI labels** representation.

The final feature representation is constructed from age of acquisition (AoA) norms. Dale and Fenson (1996) collected word-level acquisition norms from over 1000 parent reports indicating the productive vocabulary of children between the ages of 16 and 30 months. These norms are aggregated such that, for each word, the proportion of children who produced that word at a given age, binned at monthly intervals, is known. We take each word as a vector indicating the proportion of children across the different ages who produce a specific word. This input representation may allow for the model to capture language learning differences between an individual child and normative learning. Success of this model would indicate that, in addition to knowledge of the child's specific vocabulary, difference from normative acquisition may be useful in predicting future language learning. We call this representation our **AoA norms**. Collectively, we call the Phonology, CDI labels and AoA norms **language representations** as they capture features of language that contain non-semantic information..

We believe that by extending the representation of each word to a vector representation, rather than a single binary value indicating knowledge of a word, our modeling approach may more accurately capture the language learning of individual children. We believe that the more salient and meaningful a set of features are to young children, the more predictive our neural network models will be. We also suspect that some of these representations will fail to account for language acquisition. This failure can be used to suggest features which are not readily available to young children, either due to complexity of theme, variability in the environment, too little example data or other reasons. While our modeling cannot directly answer these questions, it can highlight features

and representations that aid in our ability to explain language learning trajectories.

One final consideration of these various representations is how to aggregate the word specific vectors to most 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 regardless of the point in development. However, with summing, information regarding the child's age and vocabulary size is indirectly measurable through the activation level. In addition to the method of aggregation, we also consider whether we see an improvement in predictability when we additionally consider child specific features such as age and language ability within the vocabulary representation.

All in all, we construct four different variants for each feature representation. One averages the individual word representations and one sums the word representations. For each of these two cases we also consider the effect of adding in the CDI child features to each of the input representations. In practice, these models contain different types of information. Figure 3.2 visually represents the vocabulary of two children under the seven input variants discussed above as well as the child feature representation. The top two rows assume individual word representations are summed, while the bottom two rows 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 aggregated based on parts of speech (e.g. noun, adjectives, verbs).

For each of the seven vocabulary representations discussed above, and for the three CDI variations, we choose one of the four model aggregation methods to investigate in detail. All models are assessed on their ability to predict the full (677 word) vocabulary of a specific child based. Models only differ in the architecture and input features. Input features are defined by a particular vocabulary representation and the child's productive vocabulary report at the session previous to the predicted session (or initial CDI in the snapshot). The model architecture is optimized for each specific feature set, using 4-fold cross validation. The architecture is fixed before testing. Once the model architecture is fixed, then the combined training and validation set are used to learn the final model weights. Evaluation is performed on the withheld test set.

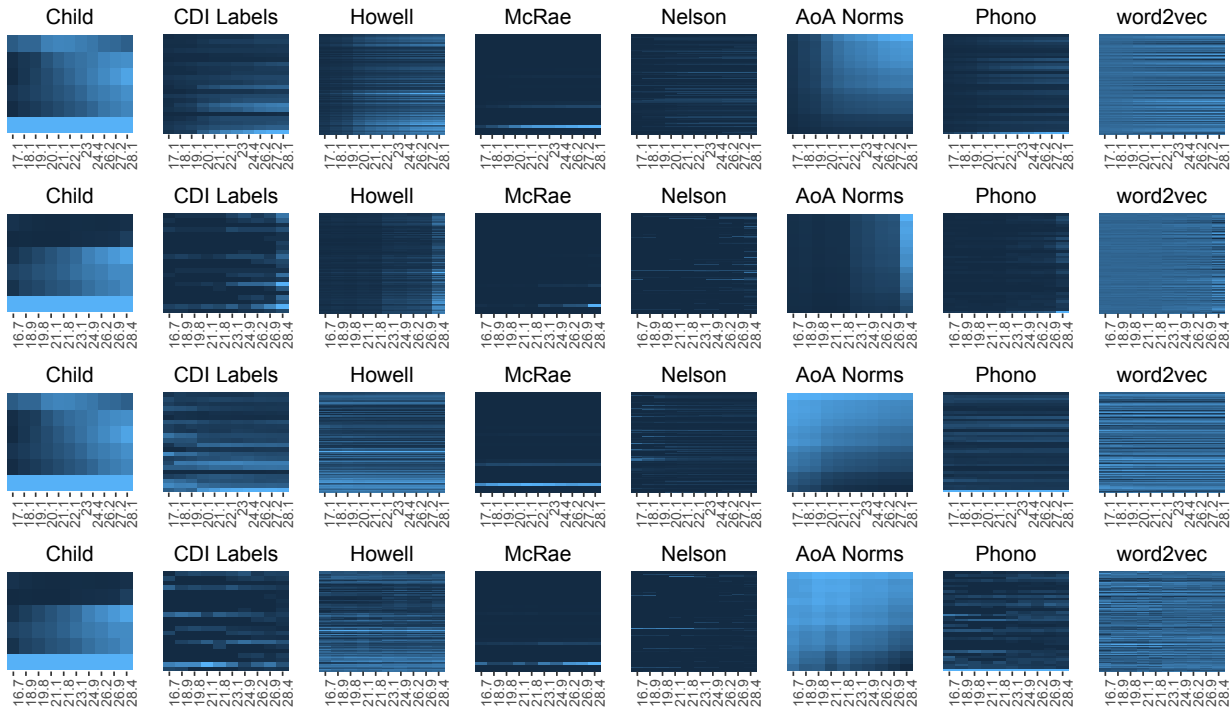

Figure 3.2: Input representations to the neural networks based on different methods of aggregation for two children. The top rows represent adding the individual word specific features, the second two rows capture the averaging of individual words to aggregate a child's current vocabulary knowledge. The neural network model must predict the words to be learned next given the particular input representation (a column vector in each plot). Along the x-axis are the CDI age time points. Along the y-axis are the features. Lighter color indicates higher activation.

3.6 Evaluation

We first select model architectures based on minimizing negative log-likelihood (nLLK) error on a rotating validation set. Once network architecture is optimized for each representation, we select a single model for investigation and comparison. We narrow the exploration of models further by restricting exploration to models whose validation error is always less than that of the CDI Child model. Only after narrowing the set of models do we consider the test set. We assess performance on the test data based on nLLK. We additionally consider the measures of percent overlap, ROC discriminability, which indicates the trade-off of true positives and true negatives as thresholds vary, and accuracy measures. We also report the t-statistic based on assuming there is no differences between the model and the CDI Child model.

Accurately predicting individual word learning has many applications. But simple predictive assessment may mask relevant developmental changes. To explore development, we consider model performance as related to the i) language ability, ii) age, and iii) vocabulary size of the children in our test set. With this approach we hope to uncover developmental changes as related to the role of feature representations. These results can be used to summarize developmental phases. 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, whereas semantic networks could capture changes in productive vocabularies of older children.

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 Howell Sensory Motor Norms (Howell et al., 2005) will be extremely accurate at predicting the acquisition of early words, since concrete nouns are the words included in the original norming study, but generalize less well to later learned words which are more likely to be 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 according toe the CDI age of acquisition norms. Because earliest learned words are often concrete nouns (Gentner, 1982), we might expect the Howell feature model 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 (Mayor & Plunkett, 2011).

Finally, we conclude by considering **ensemble** approaches that aggregate individual model predictions for each word. Using a subset of the representations, selected based on performance, we construct an linear additive model that combines the predictions each individual prediction in order to more accurately predict the learning of individual words for individual unseen children. Optimizing the influence of each model using the a sample validation set, we can assess the importance and contribution of each representation. We also consider **voting models** based on word features and child features. We explain these ensemble models in more detail below.

3.7 Baseline performance

Negative log-likelihood (nLLK) is useful and efficient in training neural networks; but as a metric, it can be difficult to interpret. To understand performance of these neural network models, we orient the readers by introducing a few nLLK scores for comparison. If the model always returned .5 as the probability of learning a word, the average nLLK score of predictions would be 0.63. 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 nLLK 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 which indicate the proportion of children at a given age who reportedly produce a specific word, called the CDI age of acquisition norms above. We can also estimate the learning rate of individual words directly from the training data. Using the published CDI norms or the training data, we get a nLLK of 0.46 and 0.45 respectively.

Our final (informed) baseline nLLK 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 0.44 when using ridge regression regularization. See Beckage, Mozer and Colunga for more detail on the modeling framework and results (2015). Any model that contains useful information to word prediction must clearly outperform this logistic regression model by attaining a score smaller than 0.35.

All neural network models outperform the age and word specific baseline models discussed above. This is true regardless of the particular model architecture chosen. All optimized neural networks also outperform the logistic regression models. Of the models tested, the model with the worst performance still has a negative log-likelihood error of 0.32. We thus 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 models and then we turn to the word feature models.

3.8 CDI models

Following the approach of modeling vocabulary growth with logistic regressions (Beckage, Mozer, & Colunga, 2015), we use a) global **CDI Child**-level features (n=6), b) local **CDI Word**-level features (n=677), and c) a **combination** of both (n=683) as input to the neural network models. To evaluate performance, we consider the resulting average negative log-likelihood (nLLK) error of all predictions. Note, this more heavily penalizes the vocabulary snapshots of children with smaller vocabularies as we only predict unknown words and thus make more predictions for children with fewer words in their productive vocabulary. We find that the child feature model performs with less accuracy than the word feature and combined models, confirming our expectations that the specific word knowledge contains information useful in predictions of future acquisition.

In Table 3.1, we report summary performance of these CDI representations by average nLLK for all snapshots in the validation data, averaging across all folds. Also included are evaluations on the predictive accuracy as computed based on percent overlap and ROC measures. Percent

overlap measures the overlap between the k words reported as learned by the child and the k' top words learned by the model. The percent overlap between k and k' tells us approximately how accurate the model is at correctly discriminating which words are learned. It does not consider correct predictions of words that are not learned. We report the median percent overlap as there is a large amount of variability across children. Receiver Operating Characteristic (ROC) curves compute the trade-off between true positives and true negatives as the cutoff varies for converting probabilities into binary classes. Summary measures of accuracy and discriminability (d-prime) are also 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 in the data. AUC is the area under the ROC curve. The AUC summarizes the ROC performance 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 an unpaired t-test on average nLLK across runs.. We do not present these standard deviations in the results table. We find that the CDI word representation has the lowest negative log-likelihood error. But the CDI combination model and the CDI Child model perform with high accuracy. The combined model even performs better in terms of overlap and accuracy than the word model.

model	nLLK	% overlap	AUC	acc	d-prime	t-stat.
CDI Child	.317	40.3	.827	.838	.173	—
CDI Word	.313	41.0	.831	.842	.176	24.2
CDI Comb.	.314	41.1	.831	.843	.175	17.4

Table 3.1: Neural network performance on validation data using representations aimed at capturing information on the CDI.

Performance of these models suggests that neural networks are capable of predicting future word learning for individual children. These models may capture specific populations of learners with more accuracy as certain groups may learn more systematically than others. To examine this idea, we consider model performance, grouping snapshots by child and word specific features. If the **CDI Child** model performs well for younger children but not as well for older children, we would expect to see a relative gain in predictive accuracy for this model when compared to a

different model on the same group of children. In Figure 3.3, we order snapshots by the child's age, vocabulary size, or percentile. We consider a smoothed average of the nLLK score. We also consider the average nLLK for individual words when sorting by the earliest to latest learned words according to the CDI norms (Figure 3.3; lower right, Dale & Fenson, 1996). We zero relative performance based on the **CDI Child** feature model predictions in order to compare across models. The graph indicates the smoothed nLLK, averaging over the previous 20 snapshots predicted. The colored bar below indicates which model is the best performing model at that age, vocabulary size, percentile, or age of acquisition word index.

In figure 3.3, we find we are able to more accurately predict younger children, children with small vocabulary sizes, and children with low CDI percentiles. This is because learning at this point in development is generally at a slower rate with learning moments being more isolated and possibly more systematic. For example if a child is a slower learner than their peers, the words the learn are likely words that are often used and also important to communication, allowing the model to learn a meaningful relationship to specific unknown words. Looking at the age of acquisition ordering, words learned early are also the words that are hardest to accurately predict acquisition of. Again this is because learning is happening at a slower rate but the individual learning moments are more difficult to capture with high certainty for a specific word since their are many candidate words. If the child knows nearly all words, the model may give a high estimate of learning unknown words but if the child knows very few words, it may be difficult for the model to predict what specific words will be learned next.

Beyond general developmental effects, we see periods in the course of development where one model outperforms the other models. Children between 16 and 24 months are more accurately modeled by the **CDI Child** feature model than the other models. Also, the learning of the earliest words are best captured by the **CDI Combined** model. We will use these trends found when extending model estimates to validation snapshots of unseen children to inform an ensemble which considers developmental trends.

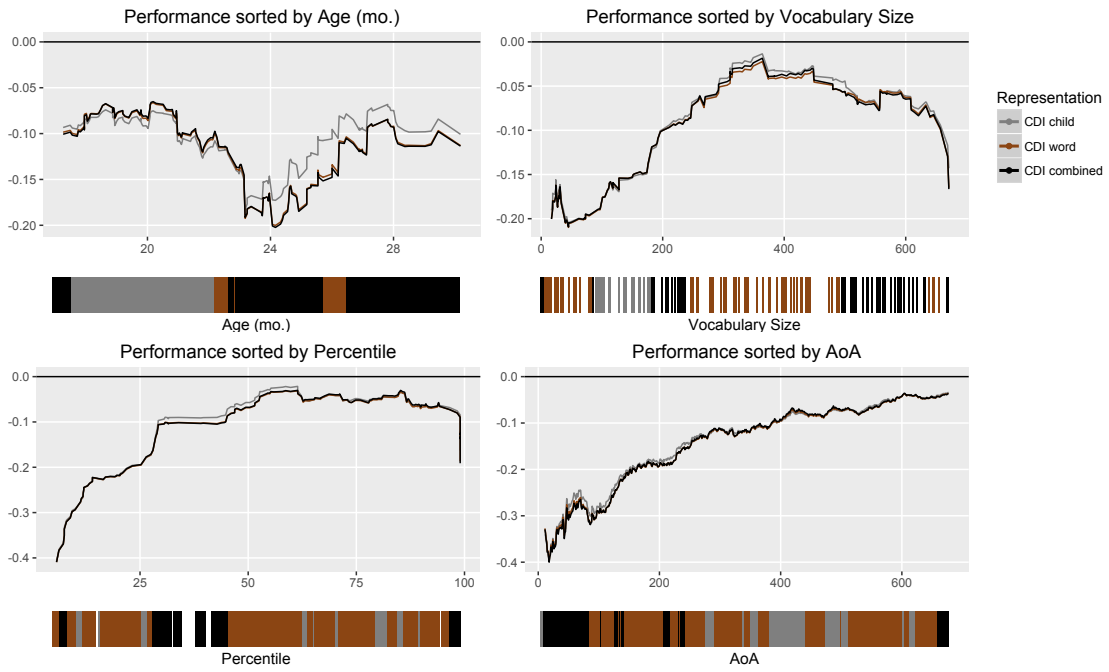

Figure 3.3: We consider performance on the CDI neural network models as a function of age (top left), vocabulary size (top right), CDI percentile (bottom left) or average age of acquisition (AoA, bottom right). The line plots indicate smoothed negative log-likelihood, the colored bars indicate the best performing model at that moment in time.

3.9 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 semantic, phonological, and other language representations as input to the neural network. As mentioned above, we also consider whether averaging or summing the individual word features produces the best predictions. We also consider 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 a large effect on the ability for a model to predict future lexical acquisition. However, across all representations, there is no best aggregation method, suggesting there is different information available in each representation. Interestingly, only two models (the **CDI** Labels and the **McRae** 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 most representations. The other five models, including the Phonology model above, saw no reliable improvement when the child features were included. Additionally, three of the seven representations were best fit by summing the individual word representations and the remaining four showed increased accuracy when the individual word features were averaged. This suggests that not only is different information available and useful in neural network training, but also that the granularity and role of this information differs. Some representations benefit from a relationship between total activation level and vocabulary size while other representations do not. For the rest of our analysis we choose the best aggregation model for each feature representation.

We now turn to the performance of the neural network architectures we classified as being predominantly based on semantic similarity. This includes the **Howell**, **McRae**, and **Nelson**

model	nLLK	% overlap	AUC	acc	*d*-prime	*t*-stat.
Howell	.317	41.3	.826	.841	.174	-2.7
McRae	.318	40.5	.826	.839	.174	-7.8
Nelson	.313	40.9	.830	.842	.175	23.7
Word2Vec	.315	40.3	.828	.843	.175	22.4
CDI Child	.317	40.3	.827	.838	.173	—
CDI Word	.313	41.0	.831	.842	.176	24.2

Table 3.2: Neural network performance of validation data assuming semantic representations.

feature norms as well as the **Word2Vec** representation. Table 3.2 suggests the Howell feature representation and the McRae features do not contain information beyond what is available when considering only the **CDI Child** features when trained networks are extended to validation data. We find the Nelson network reaches comparable performance with the word feature model even though this model has no direct information about the words in the child's vocabulary. Instead, the Nelson representation contains information about the connectivity, in terms of cue and response in free associations, of the words in the child's vocabulary to other words on the CDI. The success of the Nelson model may be because of the strong relationship between free associates and self-reports of age of acquisition (Nelson et al., 2004; Steyvers & Tenenbaum, 2005; De Deyne & Storms, 2008). We summarize performance of these models using various metrics in Table 3.2.

Closer inspection of these models suggest the performance of the McRae feature representation is mostly due to the addition of the child features. Removing the child features from the representation results in a decrease in performance for every measure. For the other semantic representations, the addition of child specific features did not improve performance suggesting that the contribution and effect of these measures may be correlated with vocabulary, making the child features redundant. Overall the Nelson representation is the best semantic representation as it minimizes error across all predictions and according to ROC measures. It is also the most statistically different model from the CDI child feature model. The Word2Vec model is also more accurate than the CDI child feature model and is the best performing model when assessed by percent overlap.

In figure 3.4, we consider the performance of these models in relation to child language

development by plotting smoothed nLLK error as ordered by child- and word-specific measures. We find interesting trends suggesting the Word2Vec representation is most accurate for older children, children with small vocabularies and children with lower percentiles. The Nelson model seems to be the overall best performing model but we see the greatest relative gain for young children and children with a percentile between 25 and 50. While the McRae feature model looks to be performing well for young children, the difference between this model and the Nelson model is not robust. We return to the effect of development later when constructing informed ensemble models.

We now turn to the final three representations used in our neural network model. These representations we collectively call language representations, to distinguish from the CDI and semantic representations above. These representations are based on phonological composition of the child's vocabulary, the number (or proportion) of each CDI category label a child knows, and the aggregated age of acquisition (AoA) norms for each individual word. Each of these language models outperforms the CDI child model. The AoA model also outperforms the CDI word model. This implies some representations of a child's vocabulary can provide additional information, beyond predictions based on the individual words a child knows. This result is confirmed in Table 3.3, where we again summarize model performance with the measures explained above. Our findings suggests that we gain improvement in predictability of individual acquisition when the model also has information about normative acquisition through input of AoA Norms. This may imply that what is most important to word learning is the difference from normative learning, rather than the individual words the child currently knows.

model	nLLK	% overlap	AUC	acc	*d*-prime	*t*-stat.
Phonology	.317	40.5	.825	.839	.172	-8.3
CDI Labels	.314	40.9	.828	.841	.174	20.5
AoA Norms	.310	41.5	.837	.844	.179	85.4

Table 3.3: Neural network performance on validation data using representations aimed at capturing semantic information.

Finally, we again visualize order predictions based on relevant developmental features of the

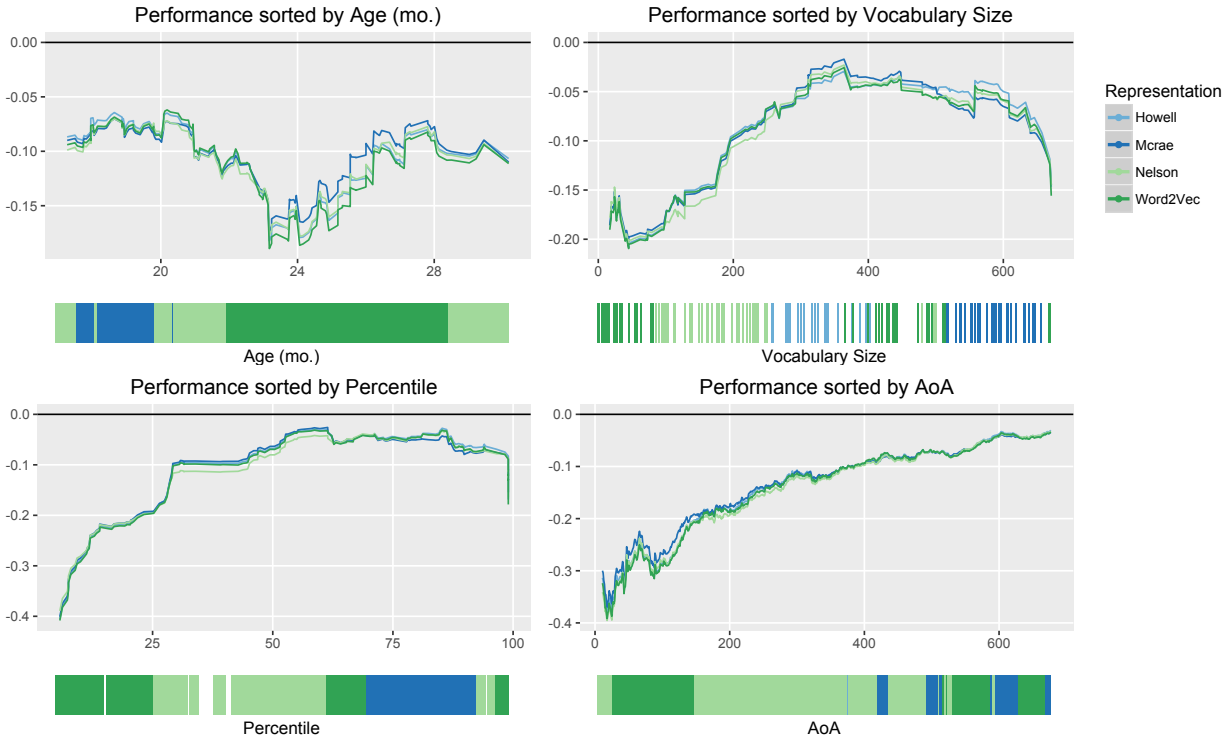

Figure 3.4: We consider performance as a function of age (top left), vocabulary size (top right), CDI percentile (bottom left) or average age of acquisition (AoA, bottom right). The line plots indicate smoothed negative log-likelihood, the colored bars indicate the best performing model at that moment in time.

learner or word. In figure 3.5, we see clearly that the neural network trained on **AoA Norms** performs the best across most percentiles. This AoA model is usually outperformed for early talkers and large vocabularies by the **CDI Labels** model. This Label model may be accurate for children with more language knowledge because the labels may capture categories that are related to interests or environmental categories that are related to directed learning. The **Phonological** model here does not account for individual language acquisition well. This may be due to the coarseness of aggregating based only on the presence of individual phonemes. Collectively, the results suggest the CDI label model and the Age of Acquisition norms increase predictive capabilities of our models in accurately capturing what words are likely to be learned next.

3.10 Ensemble models

Before extending the best performing models to the withheld testing data, we consider ensemble models. Focusing on a subset of the models that outperform the **CDI Child** model and nearly as well as the **CDI Word** model, we use the neural network models trained on the 1) **Nelson** free association norms, 2) **CDI Labels** and 3)**AoA Norms**. We also include the CDI Child and Word models in our ensembles. Below we explain the construction of a variety of ensemble models that combine the predictions of best performing models explored above. We could train neural network models that include multiple representations as input to the neural network, here we instead focus on averaging the final predictions. To avoid biasing our results more than necessary, we use the validation data sets to explore these models.

A basic ensemble model simply considers each of the best performing models equally. In this **Avg. Ensemble** model, we combine prediction across the **CDI Child** and CDI Word model as well as our best performing language representation models – **Nelson**, **CDI Labels** and **AoA Norms**. The performance of this ensemble model, as reported in table 3.4 is comparable to the child feature model but does not outperform many of the neural network models discussed above. The second **Wgt. Ensemble** uses the training data to optimizes the contribution of each model to construct a linear weighted average that minimizes the training error. This weighting is then

Figure 3.5: We consider neural network prediction accuracy sorting by the child's age (top left), vocabulary size (top right), CDI percentile (bottom left), or word age of (average) acquisition (AoA, bottom right). The line plots indicate smoothed negative log-likelihood, the colored bars indicate the best performing model at that moment in time.

extended to the unseen validation fold. We can see in table 3.4 that this model performs better than simply averaging all predictions together but still does not outperform our best single feature model of age of acquisition norms when applied to unseen data. Looking at the weighting of the individual representations, this model suggests that the vocabulary representations that are most useful are the **AoA Norms** and the **CDI Word** representations, accounting for 52% and 27% of the total estimates respectively. The CDI Child model and the CDI Labels are almost completely ignored in this ensemble model.

We also consider voting models. In our **Word Vote** model, using the training data, we select which model performs best for each word. We then generalize to the validation data by combining the best performing model for each word to jointly predict the full vocabulary. Table 3.4 shows performance of this model. This model does not outperform even the basic **CDI child** model suggesting that the variance in predictive ability of our neural network models is not related to specific words that the model is predicting. Instead the variability is likely based on differences across children.

model	nLLK	% overlap	AUC	acc	*d*-prime	*t*-stat. (AoA)
Avg. Ensemble	.309	41.1	.836	.843	.177	78.4
Wgt. Ensemble	.308	41.3	.837	.843	.179	89.8
Word Vote	.314	40.2	.834	.844	.178	55.8

Table 3.4: Average performance of ensemble models on validation data.

To test the role of individual differences in children, we consider a voting model based on features of the learner. As suggested by ordering performance by the child's age, percentile or vocabulary size, we construct a voting model that uses similarity on a particular child feature (e.g. age) to select a single model for extension to the unseen snapshot. We extend this model to take into account multiple features by defining similarity, to be the euclidean distance between a vector of child features. Below in table 3.5 we consider the performance of 5 different variants of this child voting model. The first version selects based on similarity in age; we call this model **Child Vote (age)**. Following the age voting model, we also consider CDI percentile **(%)** or vocabulary

size (abbreviated **voc sz**). In the **Child Vote All** we consider age, percentile and vocabulary size computing the euclidean distance between a validation fold and training snapshots. We find that the percentile voting model performs best, in terms of nLLK and ROC measures performing better on unseen data than even the model considering all features jointly.

model	nLLK	% overlap	AUC	acc	*d*-prime	*t*-stat.
Child Vote (age)	.313	40.6	.830	.844	.178	6.4
Child Vote (%)	.309	40.9	.833	.844	.177	43.7
Child Vote (voc sz)	.311	40.7	.833	.845	.179	31.4
Child Vote All	.314	40.6	.835	.843	.178	8.0

Table 3.5: Average validation performance of ensemble child-voting models.

As expected the method of constructing an ensemble model, combining our individual neural network model predictions, affects predictive ability. Surprisingly, many of these models fail to perform as well as the **AoA norms**, suggesting that the information contained in these representations may be redundant or have less information as compared to the AoA norms. We find that the **Wgt. Ensemble** model performs the best of our ensembles. This weighted ensemble model considers the AoA norms most heavily but also weighs the CDI word representation and Nelson representation. Our word voting model did not generalize well to unseen children but the **Child Vote (%)** performed similarly to the weighted ensemble model. The child voting results suggests individual differences can be captured by percentile more than by either vocabulary size or age. We now extend these ensemble models, and our best performing models, to the withheld test set.

3.11 Extension to the test set

Give the results on the validation data, we consider performance of the best performing subset of models on the withheld testing data. We include the CDI feature based models that use 1) CDI Child features 2) CDI Word features as well as vocabulary representations based on the 3) Nelson Norms, 4) CDI category labels and 5) Age of Acquisition Norms. We also consider three ensemble

models: 1) Averaged Ensemble 2) Weighted Ensemble, and 3) the Voting model based on CDI percentiles.

We consider one additional model in an attempt to capture development. In this model, we aggregate predictions of model representations that best predict particular children with the highest accuracy at a specific point in development in the training data. For example, if the initial snapshot of the child indicated the child's age to be 25 months, percentile to be 15 and the vocabulary size of 330 words, we would combine the CDI word predictions (most accurate for children older than 24 months), AoA norms predictions (most accurate for children with low percentile), and the Nelson estimates (most accurate on vocabulary sizes of 300 to 400 words) to predict the probability of learning unknown words for this new child. We call this model our **ordering** ensemble model as it aggregates different models based on their ability to account for certain periods of development in the training data. This model will hopefully generalize well to unseen children, indicating that there is some systematic information in these representations that are useful in predicting future acquisition trajectories.

First considering the highest performing neural networks given the validation results, we report the results in table 3.6. We average across 10 runs, after training using the combined training and validation data. Table 3.6 shows average performance in terms of nLLK, % overlap and ROC measures. nLLK on the test set is generally slightly lower than the validation sets, suggesting these neural network models benefit from the additional training data from the validation sets. We do see, however, a decrease in the percent overlap measure and ROC measures. A significance test on nLLK across runs confirms all models outperform the CDI child model. Performance across runs does not vary much with a standard deviations of less than 2% (not reported in table). Recall that the test data is not only unseen snapshots but also unseen children. In general, the accuracy in terms of standard error and percent overlap is quite high, given other previous models in this area (Beckage, Aguilar, & Colunga, 2015; Beckage, Mozer, & Colunga, 2015; Beckage & Colunga, 2013). We again find that the **AoA Norms** model is the best performing of the single representations though this model is not statistically better than the **CDI Labels** model. The results suggests that

much of the information in predicting future language may be related to differences from normative acquisition rather than the content of the specific items or category classes in the child's current productive vocabulary.

model	nLLK	% overlap	AUC	acc	*d*-prime	*t*-stat.
CDI Child	.312	36.9	.815	.840	.167	–
CDI Word	.311	36.5	.815	.843	.167	6.7
Nelson	.311	37.1	.816	.840	.167	5.8
CDI Labels	.308	37.6	.819	.843	.169	51.3
AoA Norms	.307	36.6	.819	.841	.169	66.1
Avg. Ensemble	.305	37.4	.823	.844	.172	70.7
Wgt. Ensemble	.305	37.3	.821	.843	.171	62.9
Child Vote (%)	.309	36.8	.817	.842	.169	24.5
Ordering Ens.	.307	36.9	.819	.841	.168	49.8

Table 3.6: Performance of neural network models on the test set.

Table 3.6 also includes performance of the ensemble models. In general, we find the ensemble models perform better than expected and in fact better than the neural network models differing from our exploration on the validation data. These results suggest that there may be unique information in the different input representations that aid in predictions of individual word acquisition of unseen children. The similarity between the weighted ensemble model and the average ensemble model suggests that the benefit of these various representations can be accessed by a simple averaging predictions without the deed to approximate a weighted average.

We also see that the ordering model does better than many of the single input representations and performs on par with the single AoA norms representation. This suggests the need for researchers not only to consider the individual words the child is learning but also the differences in learning across the course of development. We discuss these findings and their implications in more detail in the discussion section.

3.12 Conclusions and discussion

Individual words in a child's vocabulary are informative in predicting future vocabulary growth. The **CDI word** model with just the productive word information reliably outperformed the **CDI child** feature model. This confirms our intuition that the individual words a child knows contains information beyond the child's age and vocabulary size in terms of predicting what words are to be learned next. This is an interesting result given that many of the current diagnostic and intervention techniques in developmental psychology rely on information pertaining to the size of the child's vocabulary, with little attention given to the specific words known by the individual learner.

These results also suggest the need to consider differences in learners. The content of the vocabulary improves our ability to predict future acquisition but we still see clear developmental effects. While we remain agnostic as to what the relationship between known words and future learning, we find strong evidence of its presence and strong evidence for change over time. But **Howell** and **McRae** features (capturing semantics) and the phonemic transcription model performed significantly worse than the CDI word models indicating that the modeling framework alone is not powerful enough to model acquisition. Instead success is dependent on the information available in the input representations relevant to word learning. We see that these representations do not aggregate the child's current vocabulary knowledge in a meaningful way for future prediction. In later work it may be interesting to consider why these models fail. For example, the chosen phonological representation may be the incorrect choice for young learners. Instead we may need to consider representations based on phonemic onset, rhyme, sound similarity, or the difficulty of pronouncing individual phonemes (Goswami & Bryant, 1990; Hulme et al., 2002; McMurray, 2007; Gierut, Morrisette, Hughes, & Rowland, 1996).

More interesting than the failure of individual representations is that some of the feature representations perform on par with, or better than, the vocabulary vector representations of each knonw word. The feature aggregation using the **CDI Label** or word class labels reliably out

performs the CDI word model. This suggests knowing something about the class of words the child knows can help in predicting acquisition of individual words beyond knowledge of each individual word. This is possibly because this representation captures important structures and features to young learners. The **AoA norm** model reliably outperforms all other models. This suggests a very interesting idea. The norms when used to directly predict acquisition of words for the children in our study (e.g. CDI age baseline as discussed in section 1.3) do not capture language learning well. But when we aggregate these word-level acquisition norms to represent a child's full vocabulary knowledge, we find the norms hold information beyond the child's current productive vocabulary and are our most accurate predictive model

The success of the AoA norms requires further investigation. The increased performance of this model compared to the other models explored could indicate that aggregation of normative acquisition accentuates different learning styles. These differences in styles of learning would allowing for easy detection by a neural network of a learner's trajectory. This learning trajectory may be the most reliable clue of the trajectory for future growth for an individual child. The AoA norms model may additionally provide insight into important differences in learning trajectories. Classification of trajectory and of different learning styles may also be possible within this neural network framework. This model performs especially well for a particular group of children commonly known as **late talkers**, children who know far fewer words than their age-matched peers. In the future, we plan to extend this approach to diagnosis and possible interventions of children with language learning difficulties. We plan to use these input representations to not only to predict the acquisition of individual words, but to predict a child's language skill over the course of development, modeling the **language trajectory** as opposed to individual word learning.

Looking at performance differences across development, we see evidence that developmental changes and learning have an impact on what words a child is going to learn next. We measure influence of development by ordering validation performance for a specific fold by certain features we think are relevant to language learning in young children. We also considered child-based ensemble models that weight the similarity of individual children on different features. We also constructed

an ensemble model that used ordering results 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. 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 (Weizman & Snow, 2001; Heilmann et al., 2005; Beckage et al., 2011). An implication of this result is that children who are having more difficulty with language are the children who have widely variable learning strategies. Given this idea, the success of the AoA norms for this population of low percentile children is hopeful. It provides us a model to predict future word learning and may also be extended to 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. In future work, we hope to use this insight to explore the attentional and learning mechanisms that result in learning differences between these groups.

By considering the developmental aspects inherent in this type of modeling, we can make predictions (and evaluate those predictions) on where a specific child's productive vocabulary will be in the future. This type of modeling approach allows us to capture and explain the effect of certain features in language learning as related to development and to the vocabulary knowledge of a young child. This type of approach and assessment may, in turn, not only be useful for predicting word learning of an individual child but could allow us to distinguish late talking children who catch up to their peers from those late talking children who do not. This predictive modeling approach would also suggest targeted interventions for individual learners. Given that the neural networks are able to predict future acquisition one month into the future, we can begin to predict further than one month, assessing language ability throughout the course of development. We can also use these different language representations to further tease apart different types of learners and different acquisition processes, possibly capturing and explaining differences of late and typically developing children.

Chapter 4

Comparing Network Models and Neural Network Models

In the previous two chapters we considered models of lexical acquisition using either network growth models or neural network models. While quite different in the underlying assumptions, these models do, in the end predict the probability that a particular word is learned by a particular child. In this section, we compare performance of these models directly. We previously discussed the conclusions we can draw from these models individually and now look for insight from joint network and neural network models. We explore these joint models by comparing the performance of these models individually to logistic regression models. We then build ensemble models, aggregating the predictions of the network growth models and the neural network models to further increase our predictive accuracy and understanding of acquisition.

In table 4.1 we again report performance of 4 individual and 3 ensemble network-based models. The details of these models and conclusions of their performances are explained and discussed in detail in chapter 2. Table 4.2 reports performance for the best performing individual neural

model	NLLK	% overlap	AUC	acc	*d*-prime
Net Howell	.395	29.6	.702	.807	.110
Net Nelson	.381	23.4	.711	.814	.113
Net Phono Dist	.363	31.1	.740	.823	.122
Net Word2Vec	.375	26.7	.718	.813	.114
Wgt. Ensemble	.363	31.5	.738	.822	.124
Child Vote (%)	.368	30.3	.732	.818	.123
Child Vote Avg	.368	30.1	.733	.818	.122

Table 4.1: Performance of network growth models on the test set.

model	NLLK	% overlap	AUC	acc	*d*-prime
CDI Child	.312	36.9	.815	.840	.167
CDI Word	.311	36.5	.815	.843	.167
Nelson	.311	37.1	.816	.840	.167
CDI Labels	.308	37.6	.819	.843	.169
AoA Norms	.307	36.6	.819	.841	.169
Avg. Ensemble	.305	37.4	.823	.844	.172
Child Vote (%)	.309	36.8	.817	.842	.169
Ordering Ens.	.307	36.9	.819	.841	.168

Table 4.2: Performance of neural network models on the test set.

network representations as well as the best performing ensemble neural networks; see chapter 3 for more information. The tables report performance on the testing data and include negative log-likelihood, percent overlap, area under the ROC curve (AUC), accuracy and *d*-prime values. For further discussion of these measures, please refer back to chapters 2 and 3.

Considering performance on these tables, the neural network models greatly outperform the network growth models. In our neural network analysis, we considered the CDI child feature model as our baseline model. Here it is clear that none of the network growth model are as accurate as even our baseline CDI model. Comparing performance of individual snapshots using the best performing weighted ensemble network growth model, we find that for all except for 2 snapshots, the child feature neural network model outperforms the best performing network growth model. We similarly compare the performance of these two models on the accuracy of individual words. Nearly every word is predicted with higher accuracy by a neural network model than by a network growth model.

While it is clear that neural network models are outperforming the network growth models, it is unclear if that is because the network growth models are inaccurate overall or if the neural network models are just benefiting from the non-linear transformations allowed by the hidden layer. To explore this question, we train a logistic regression variation using the various input representations that were successful in the neural network models. We train the models on 3 out of the 5 folds, validate on the remaining fold and then evaluate our chosen model on the remaining test fold used

model	NLLK	% overlap	AUC	acc	*d*-prime
LR Word	.440	35.6	.803	.839	.167
LR Nelson	.404	35.5	.810	.838	.166
LR CDI Labels	.417	35.6	.811	.840	.167
LR AoA Norms	.418	35.7	.807	.836	.163

Table 4.3: Performance of linear regression with ridge regression normalization. For each word, an independent linear regression model is trained.

to evaluate all models in the dissertation. Due to the size of the data and number of predictors, ordinary least-squared (OLS) models do not generalize well to the validation data. Instead,we use ridge regression for regularization. Table 4.3 indicates performance of these models, evaluated on the same metrics discussed above. We can see here that the logistic regression variant has very comparable performance to some of the single networks but that the network growth aggregate performs better. This suggests that there is some additional information available in the network growth models that is not accessible in the logistic regression model. This is a strong result as the logistic regression model still has many more parameters than the network model as each word requires its own logistic regression function.

The superior performance of the neural network models may not be surprising as neural network model have many more parameters (namely the weights in the neural network architecture). These weights allow for a more flexible model, resulting in a model that can be trained to more accurately predict the observed training data. Additionally, the results of the logistic regression analysis suggest that the hidden layer is crucial to the performance improvement. The relatively accurate performance on unseen snapshots supports the idea that language learning is systematic and thus statistical learning is a useful approach in predicting future acquisition even in the presence of individual variability. These results also suggest that investigating more sophisticated neural network architectures such as deep neural networks or recurrent neural networks may be of great benefit if the goal is accurate prediction of words to be learned next.

Network growth models do outperform logistic regression variants at predicting individual acquisition trends but are still less accurate than the neural network models. Additionally, these

network growth models, more so than neural network models, make predictions about the **mechanisms** of acquisition that influence and alter future lexical learning. The growth models also suggest how the current vocabulary of the child relates to future acquisition allowing for design of experiments and interventions in ways neural network models do not. Clearly, though, the network growth models are not accurate models in terms of capturing at the level of individual learners what words the child is likely to learn next. It is of future work to investigate where exactly the weakness of these network models is. It could be that a network representation is not the right representation of language for young children or the process of proposed growth is incomplete. It is also likely that the mapping of a specific role in the network to a probability of learning is incorrect. Future work will involve expanding the possible mechanisms of growth underlying these network models as well as algorithms that include other relevant features to word learning beyond only considering contributions of relational similarity.

In the future, we hope to apply these model to predicting the trajectory of language acquisition in young learners rather than the specific words that are learned. It is possible that the network growth model may be more useful in this context than the neural network model–it may suggest not only the individual words to be learned but also how to maximize the structure of the child's vocabulary in order to improve future language acquisition. For a task of predicting the trajectory of word learning, knowing the order of learning is not enough. In this case, the structure of the current vocabulary, available to network growth models but not to neural network models, may suggest relevant clues about the learning process that a predictive neural network model could not. Additionally, network growth models may be better able to distinguish and explain different types of young learners.

4.1 Input representations

Besides performance differences, one other difference across these two models is the usefulness of the specific representations in yielding accurate individual word predictions. Both classes of models were trained on similar features including a representation based on the Howell sensory motor

norms (Howell et al., 2005), Nelson free association norms (Nelson et al., 2004), McRae feature norms (McRae et al., 2005), and phonological features (Vitevitch, 2008). For both models, the McRae feature norms performed poorly, suggesting that the information in this representation accessible to the models may not be related to early language growth. These feature norms were collected on adults and included many taxonomic and encyclopedic features which are likely not accessible to young children. Additionally, these features were free response of attributes of individual words. Because of the task, responses may highlight unique, rather than important, features of objects. The responses also cover only a small portion of all meaningful features of an object, limiting their usefulness in modeling early language learning. More generally, attentional mechanisms may be directing children to specific features that may not be the most salient or relevant from the perspective of an adult. While collecting semantic norms on young, toddler age children is an impossible task, it may be useful to consider representations that more accurately capture the environment and perspective of young children. For example, future work may include co-occurrence in child directed speech as input to such models as Word2Vec.

The most accurate representations within the network growth models were not the same representations that improved over the CDI word representation in the neural network models. This may suggest that some of the salient features of these representations are relational in nature. The Howell sensory motor feature norms used adult response to capture specific features that may be important to young children by asking individuals to respond from a child's perspective. For our network growth model, this type of information, represented as relational cosine similarity, predicted the word learning of individual children above the baseline network representation and age of acquisition norms, but not above the accuracy of many of the neural network models. Further, when we built an ensemble model using the network measures, the Howell and phonological network representations were the two most important for network growth. This is an interesting result especially when we consider that, for the neural network models, the Howell and phonological representation were not useful in predicting acquisition. This may be because the relevant sensory motor features and the relevant phonological features are relational and challenging for the neural

network model to use.

Future work using both neural network and network growth models may help us uncover the types of relationships that are relevant to early acquisition. These results already suggest that some features seem to be relational from the perspective of capturing word learning trajectories of individual children. Many of the features that were of most use to the neural network representations are clearly not relational however. Information about normative acquisition increases performance of the neural network model, indicating that information about the general progression of learning is useful to predicting the learning of a specific child. Future work could further explore different types of features that might impact word learning. There is no clear network representation of such information. Going beyond simple predictive accuracy, we may be able to use differences in performance across models to understand how specific attentional mechanisms affect language learning. The success of a particular network representation indicates that pairwise relationships may influence learning but the success of a particular representation in neural networks is less clear and needs more investigation. Understanding why certain representations perform better or worse is future work that may provide additional insights into the mechanisms behind early word acquisition.

4.2 Ensemble network growth and neural network models

We explore a set of ensemble network models to better understand the comparative strengths and advantages of the network growth and neural network approach. Because of the superior performance of the neural network models, we begin with the network growth models (**Net**) and incorporate neural network predictions. Specifically we build a network growth and neural network model that includes either the CDI neural network models (**NN CDI**) or our best performing neural network model which uses the Age of Acquisition norms as input (**NN AoA**). We train a weighted ensemble model using the validation folds to approximate performance on the withheld test set. We then construct an additive model from the best performing ensemble model of each the network growth (**Net Ens**) and neural network (**NN Ens**) We consider only a weighted ensemble model to restrict the possible set of ensemble models we can build from these model prediction. We

model	NLLK	% overlap	AUC	acc	*d*-prime	*t*-stat.	net influence
CDI Child	.312	36.9	.815	.840	.167	–	–
AoA Norms	.307	36.6	.819	.841	.169	66.1	–
Net + NN CDI	.311	41.5	.833	.843	.177	32.8	.05
Net + NN AoA	.309	41.6	.838	.845	.180	69.6	.06
Net + NN	.308	41.2	.838	.844	.179	87.9	.02
Net Ens + NN Ens	.305	37.1	.823	.844	.172	721.1	.03

Table 4.4: Performance of ensemble network growth and neural network models on the test set.

discuss possible extensions to ensemble modeling briefly in the discussion section below.

In table 4.4 we see performance of a few ensemble models as compared to the CDI Child feature and AoA Norms models. We find very small contributions of the network growth models (last column) but do see an improvement of these ensembles in terms of negative log-likelihood and discriminability (*d*-prime) and a large boost in percent overlap measures. The network and age of acquisition (Net + NN AoA) combination returns the largest *d*-prime seen so far, suggesting there may be some additional information present in the network models that helps us discriminate learning events. We also see a boost in percent overlap suggesting that the network growth models may be very certain about individual words which would affects the percent overlap measure the most. It is clear though that the neural network model outperforms the network growth model. If the network growth model offers predictive ability beyond that captured in the neural network model, it is for specific situational contexts that are washed out in our current ensemble approach.

4.3 Discussion and future work

Neural network models clearly outperform network growth models in predicting future word learning. However, there is also evidence that relational information may be informative in explaining acquisition. Specifically, for the phonological and Howell feature representations, relational models were comparatively more accurate than non-relational models. This suggests that the neural network models could be improved by having access to some of this relational information. In the future, we hope to build in this relational information by training a neural network model on the word

specific representations collectively with network connectivity measures from the network growth models. For example, we can train a neural network on the individual words the child knows but also give the neural network information pertaining to the network growth importance value of each unknown word. We could also train a neural network model that has access to the network summary statistics of the child's induced subgraph. We can also consider a network growth model that utilizes a neural network model to learn the word's importance. Given our network growth model, we can combine multiple measures of word importance, some that are network based and others that are based on neural networks. In the future we hope to explore this type of joint model.

Because of the increase in percent overlap of ensemble models that include the network growth models, it may be that these network growth models are particularly good at selecting only a few of the most likely words for learning and not the whole set. This suggests a few extensions. One is that we could aggregate the predictions of the neural network and network growth models only for the top k words of the network growth predictions. Another method might be to try to use the neural networks to infer a sequential ordering of words to be learned. The network growth model should be more accurate if we have a sequence for which words are learned since the network structure changes with the addition of each word. Most likely, the neural network could benefit from having the growth value as input to the network directly.

As mentioned above, one major difference between these two models is the number of free parameters of each model. We showed that the accuracy of the neural network model is related to the hidden layer as the individual logistic regression model is unable to perform as well as the network growth model. It is likely the case that the accuracy of the neural network model will improve with more data but it is unclear that this is true for the network growth models. Instead it may be that different network representations improve the network growth model performance. The network growth model may also benefit from more fine-grained data about the order in which words were learned. Because each prediction is based on the individual words the child knows and the specific network representation at the time of learning, the monthly snapshot data may be too coarse-grained. Specifically for the network growth models, we may find accuracy improvement by

inferring the order in which words are learned or having vocabulary reports that are more frequent than every month.

Predictive models of late talking children are particularly important. As discussed in the previous chapters, our models show an improved ability to account for acquisition of this population of learners. Now that we have the ability to predict future acquisition, future work can consider intervention-based experiments. Right now it is unclear if teaching children words with high or low probability of being learned is the most useful or if words should be taught that help bridge the gap between late talkers and typical talkers. It may even be that we need separate predictive models for each group. Though there is much more work to be done, these current models confirm systematicity in language learning that can be used for predicting future language learning.

References

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

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. Science, 286(5439), 509–512.

Baronchelli, A., Ferrer-I-Cancho, R., Pastor-Satorras, R., Chater, N., & Christiansen, M. (2013). Networks in cognitive science. Trends in Cognitive Sciences, 17(7), 348–60.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. Trends in cognitive sciences, 7(2), 84–91.

Baxter, G. J., Blythe, R. A., Croft, W., & McKane, A. J. (2006). Utterance selection model of language change. Physical Review E, 73(4), 046118.

Beckage, N. M., Aguilar, A., & Colunga, E. (2015). Modeling lexical acquisition through networks. Proc. of the 37th Conf. of the Cog. Sci. Society.

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

Beckage, N. M., & Colunga, E. (2015). Towards a theoretical framework of analyzing complex linguistic network. In (pp. 3–30). Springer.

Beckage, N. M., Mozer, M., & Colunga, E. (2015). Predicting a child's trajectory of lexical acquisition. Proc. of the 37th Conf. of the Cog. Sci. Society.

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

Bloom, P. (2002). How children learn the meanings of words.

Borge-Holthoefer, J., & Arenas, A. (2010). Semantic Networks: Structure and Dynamics. Entropy, 12(5), 1264–1302.

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 107–117.

Chan, K. Y., & Vitevitch, M. S. (2009). The Influence of the Phonological Neighborhood Clustering-Coefficient on Spoken Word Recognition. Journal of Experimental Psychology: Human Perception and Performance, 35(6), 1934–1949.

Chan, K. Y., & Vitevitch, M. S. (2010). Network structure influences speech production. Cognitive Science, 34(4), 685–97.

Christiansen, M. H., & Chater, N. (2001). Connectionist psycholinguistics: Capturing the empirical data. Trends in Cognitive Sciences, 5(2), 82–88.

Clark, E. V. (2002). Making use of pragmatic inferences in the acquisition of meaning. In The construction of meaning (pp. 45–58).

Clauset, A., Shalizi, C., & Newman, M. (2009). Power-law distributions in empirical data. SIAM Review, 51(4), 661–703.

Collins, A., & Loftus, E. (1975). A Spreading-Activation Theory of Semantic Processing. Psychological Review, 82(6), 407–428.

Collins, A., & Quillian, M. (1969). Retrieval Time from Semantic Memory. Journal of Verbal Learning and Verbal Behavior, 247, 240–247.

Colunga, E., & Sims, C. E. (2012). Early-talker and late-talker toddlers and networks show different word learning biases. In Proceedings of the 34th annual conference of the cognitive science society (pp. 246–251).

Colunga, E., & Smith, L. B. (2004). Dumb mechanisms make smart concepts. In Proceedings of the annual conference of the cognitive science society (Vol. 26, pp. 239–244).

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

Dale, P. S., Bates, E., Reznick, J. S., & Morisset, C. (1989). The validity of a parent report instrument of child language at twenty months. Journal of child language, 16(02), 239–249.

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

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

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. Behavior Research Methods, 40(1), 213–231.

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

Dorogovtsev, S., & Mendes, J. (2001). Language as an evolving word web. Proceedings of the Royal Society for Biological Sciences, 268(1485), 2603–6.

Feldman, H. M., Dollaghan, C. A., Campbell, T. F., Kurs-Lasky, M., Janosky, J. E., & Paradise, J. L. (2014). Measurement properties of the macarthur communicative development inventories at ages one and two years. Child development, 71(2), 310–322.

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

Ferrer i Cancho, R., & Solé, R. V. (2001). The small world of human language. Proceedings of the Royal Society of London. Series B: Biological Sciences, 268(1482), 2261–2265.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2009). Using speakers' referential intentions to model early cross-situational word learning. Psychological Science, 20(5), 578–585.

Gentner, D. (1982). Why nouns are learned before verbs: linguistic relativity versus natural partitioning. In Language development: Language cognition and culture.

Gershkoff-Stowe, L., & Smith, L. B. (2004). Shape and the first hundred nouns. Child development, 75(4), 1098–1114.

Gierut, J. A., Morrisette, M. L., Hughes, M. T., & Rowland, S. (1996). Phonological treatment efficacy and developmental norms. Language, Speech, and Hearing Services in Schools, 27(3), 215–230.

Goñi, J., Arrondo, G., Sepulcre, J., Martincorena, I., Vélez de Mendizábal, N., Corominas-Murtra, B., ... Villoslada, P. (2011). The semantic organization of the animal category: evidence from semantic verbal fluency and network theory. Cognitive Processing, 12(2), 183–96.

Goswami, U., & Bryant, P. (1990). Phonological skills and learning to read. Wiley Online Library.

Griffiths, T., Steyvers, M., & Firl, A. (2007). Google and the mind: predicting fluency with PageRank. Psychological Science, 18(12), 1069–76.

Gruenenfelder, T., & Pisoni, D. (2009). The Lexical Restructuring Hypothesis and Graph Theoretic Analyses of Networks Based on Random Lexicons. Journal of Speech, Language and Hearing Research, 52(3), 596–609.

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

Hills, T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The Associative Structure of Language: Contextual Diversity in Early Word Learning. Journal of memory and language, 63(3), 259–273.

Hills, T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. B. (2009a). Categorical structure among shared features in networks of early-learned nouns. Cognition, 112(3), 381–96.

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

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. Journal of Memory and Language, 53(2), 258–276.

Hulme, C., Hatcher, P. J., Nation, K., Brown, A., Adams, J., & Stuart, G. (2002). Phoneme awareness is a better predictor of early reading skill than onset-rime awareness. Journal of experimental child psychology, 82(1), 2–28.

Ke, J., & Yao, Y. (2008). Analyzing language development from a network approach. Journal of Quantitative Linguistics, 1–22.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. Cognitive development, 3(3), 299–321.

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

Liu, H., & Li, W. (2010). Language clusters based on linguistic complex networks. Chinese Science Bulletin, 55(30), 3458–3465.

MacWhinney, B. (2000). The childes project: The database (Vol. 2). Psychology Press.

Marchman, V. A., & Fernald, A. (2008). Speed of word recognition and vocabulary knowledge in infancy predict cognitive and language outcomes in later childhood. Developmental science, 11(3), F9–F16.

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

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

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. Nature Reviews Neuroscience, 4(4), 310–322.

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

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

McRae, K., Cree, G., Seidenberg, M., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. Behavior Research Methods, 37(4), 547–59.

Merriman, W. E., Bowman, L. L., & MacWhinney, B. (1989). The mutual exclusivity bias in children's word learning. Monographs of the society for research in child development, i–129.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Advances in neural information processing systems (pp. 3111–3119).

Miller, G. A. (1995). Wordnet: a lexical database for english. Communications of the ACM, 38(11), 39–41.

Motter, A., de Moura, A., Lai, Y., & Dasgupta, P. (2002). Topology of the conceptual network of language. Physical Review E, 65(102), 1–4.

Munakata, Y., & Stedron, J. M. (2001). Neural network models of cognitive development. In Handbook of developmental cognitive neuroscience (p. 159-172). MIT Press.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The university of south florida free association, rhyme, and word fragment norms. Behavior Research Methods, Instruments, & Computers, 36(3), 402–407.

Quillian, M. (1967). Word concepts: A theory and simulation of some basic semantic capabilities. Behavioral Science, 12, 410–430.

Quillian, M. R. (1969). The teachable language comprehender: a simulation program and theory of language. Communications of the ACM.

Roget, P. M. (1911). Roget's thesaurus of english words and phrases. TY Crowell Company.

Saffran, J. R. (2003). Statistical language learning mechanisms and constraints. Current directions in psychological science, 12(4), 110–114.

Sandhofer, C. M., Smith, L. B., & Luo, J. (2000). Counting nouns and verbs in the input: differential frequencies, different kinds of learning? Journal of Child Language, 27(3), 561–85.

Seidenberg, M. S., & McClelland, J. L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. Psychological Review.

Sims, C. E., Schilling, S. M., & Colunga, E. (2012). Interactions in the development of skilled word learning in neural networks and toddlers. In 2012 ieee international conference on development and learning and epigenetic robotics (icdl) (pp. 1–6).

Sims, C. E., Schilling, S. M., & Colunga, E. (2013a). Beyond modeling abstractions: learning nouns over developmental time in atypical populations and individuals. Frontiers in psychology, 4.

Sims, C. E., Schilling, S. M., & Colunga, E. (2013b). Exploring the developmental feedback loop: word learning in neural networks and toddlers. In Proceedings of the 35th annual conference of the cognitive science society.

Smith, C., Carey, S., & Wiser, M. (1985). On differentiation: A case study of the development of the concepts of size, weight, and density. Cognition, 21(3), 177–237.

Smith, C. B., Adamson, L. B., & Bakeman, R. (1988). Interactional predictors of early language. First Language, 8(23), 143–156.

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

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L., & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. Psychological Science, 13(1), 13–19.

Solé, R. V., Corominas-Murtra, B., Valverde, S., & Steels, L. (2010). Language networks: Their structure, function, and evolution. Complexity, 22, 1–9.

Stamer, M. K., & Vitevitch, M. S. (2012). Phonological similarity influences word learning in adults learning spanish as a foreign language. Bilingualism: Language and Cognition, 15(03), 490–502.

Stella, M., & Brede, M. (2015). Patterns in the English language: Phonological networks, percolation and assembly models. Journal of Statistical Mechanics, 2015, P05006.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. Cognitive Science, 29(1), 41–78.

Stokes, S. F., Klee, T., Carson, C. P., & Carson, D. (2005). A phonemic implicational feature hierarchy of phonological contrasts for english-speaking children. Journal of Speech, Language, and Hearing Research, 48(4), 817–833.

Storkel, H. L. (2004). Do children acquire dense neighborhoods? An investigation of similarity neighborhoods in lexical acquisition. Applied Psycholinguistics, 25(02), 201–221.

Storkel, H. L. (2009). Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. Journal of Child Language, 36, 291–321.

Thal, D. J., O'Hanlon, L., Clemmons, M., & Fralin, L. (1999). Vaildity 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.

Van Veen, R., Evers-Vermeul, J., Sanders, T., & Van den Bergh, H. (2009). Parental input and connective acquisition: A growth curve analysis. First Language, 29(3), 266–288.

Vitevitch, M. S. (2008). What Can Graph Theory Tell Us About Word Learning and Lexical. Journal of Speech, Language, and Hearing Research, 51(2), 408–422.

Vitevitch, M. S., Chan, K. Y., & Roodenrys, S. (2012). Complex network structure influences processing in long-term and short-term memory. Journal of Memory and Language, 67(1), 30–44.

Vitevitch, M. S., Ercal, G., & Adagarla, B. (2011). Simulating retrieval from a highly clustered network: implications for spoken word recognition. Frontiers in psychology, 2(December), 1–10.

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

Watts, D., & Strogatz, S. (1998). Collective dynamics of 'small-world' networks. Nature, 393(6684), 440–442.

Waxman, S. R., & Leddon, E. M. (2002). 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, 102–126.

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

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

Yu, C., & Smith, L. B. (2007). Rapid word l under uncertainty via cross-situational statistics. Psychological Science, 18(5), 414–420.