A Computational Study of Late Talking in Word-Meaning Acquisition

Aida Nematzadeh, **Afsaneh Fazly**, and **Suzanne Stevenson**
Department of Computer Science
University of Toronto
{aida,afsaneh,suzanne}@cs.toronto.edu

Abstract

Late talkers (LTs)—children who show a marked delay in vocabulary learning—are at risk for Specific Language Impairment (SLI), and much research has focused on identifying factors contributing to this phenomenon. We use a computational model of word learning to further shed light on these factors. In particular, we show that variations in the attentional abilities of the computational learner can be used to model various identified differences in LTs compared to normally-developing children: delayed and slower vocabulary growth, greater difficulty in novel word learning, and decreased semantic connectedness among learned words.

Introduction

Learning word meanings is a key component of the language acquisition process. While most children are very efficient word learners, some show substantial delay. Late talkers (LTs) are children at an early stage who are on a markedly slower path of vocabulary learning, without evidence of any specific cognitive deficits. Although many LTs eventually catch up to their age-matched peers, some continue on a slower path of learning, and at some point in development are considered as exhibiting specific language impairment (SLI) (Thal et al., 1997; Desmarais et al., 2008).

Early identification of children at risk for SLI is very important, since early intervention is key to alleviating its effects. Because late talking can be an early sign of SLI, many psycholinguistic studies have attempted both to understand its properties and to identify the factors that contribute to it. Research has shown that LTs exhibit not only a *delay* in vocabulary learning, but a slower *learning rate* as well (e.g., Weismer & Evans, 2002). Moreover, the vocabulary of LTs appears to exhibit less semantic connectivity than that of normally-developing children (Beckage et al., 2010; Sheng & McGregor, 2010). Numerous factors may contribute to late talking, including environmental conditions, such as the quantity or quality of the linguistic input (Paul & Elwood, 1991; Rowe, 2008), as well as cognitive properties of the learner, such as differences in categorization skills, working memory, or attentional abilities (Jones & Smith, 2005; Stokes & Klee, 2009; Rescorla & Merrin, 1998).

Computational modeling is necessary for investigating precise proposals of how such a variety of complex environmental and/or cognitive factors can interact in the process of vocabulary learning. One key mechanism believed to help children hone in on the appropriate meaning of a word (given an infinitely large number of possibilities) is cross-situational learning (Quine, 1960). Children gradually glean the meaning of a word by attending to the common elements of the meaning across its various usages, each occurring in a noisy and ambiguous context. Computational models of cross-situational learning have helped shed light on how various factors affect the timecourse of word learning (e.g., Frank et al., 2007; Yu & Ballard, 2008; Fazly et al., 2010b). However, to our knowledge, there are no computational models of word learning in context demonstrating the effects of possible factors that contribute to late talking.

We address this gap here by exploring the relation between an attentional factor and the phenomenon of late talking within a computational model of cross-situational word learning. It has been observed that children's joint attention skills—which underlie their ability to focus on the intended meaning for a word—develop over time (Mundy et al., 2007). However, our computational model as previously formulated (Fazly et al., 2010a) failed to capture the developmental increase in ability to appropriately attend to what is being talked about. Here, we extend the model with an attentional mechanism that improves over time, and show how it can be varied in computational experiments, corresponding to simulations of normally-developing children and LTs. We examine the impact of the model's differing attentional abilities, both on the timecourse of vocabulary acquisition, and on the properties of the learned knowledge. In comparing the different instantiations of the model, we find that a model with weaker attentional abilities, like LTs, shows a delayed and slower vocabulary growth, as well as less semantic connectivity among the words it has encountered. We also investigate whether the attentional factor we explore may underlie behaviour relevant to the observed subgroups of late talkers: those who eventually catch up, and those who are more likely to permanently stay on a slower path of learning.

Overview of the Computational Model

Model Input and Output

The input to our word learning model consists of a sequence of utterance–scene pairs that link an observed scene (what the child perceives) to the utterance that describes it (what the child hears). We represent each utterance as a set of words (with no order information), and the corresponding scene as a set of semantic features, e.g.:

Utterance: { *anne*, *broke*, *the*, *box* }
Scene: { PERSON, ANNE, TOUCH, CHANGE, SUDDENNESS, DETERMINER, IS-SOLID, MADE-OF-WOOD, $\cdots$ }

Given a corpus of such utterance–scene pairs, our model learns the meaning of each word w as a probability distribution, $p(.|w)$, over all possible semantic features: $p(f|w)$ is the probability of feature f being part of the meaning of word w.

Initially, since all features are equally likely for each word, the model assumes a uniform distribution for $p(.|w)$. Over time, this probability is adjusted in response to the cross-situational evidence in the corpus.

Learning Algorithm

Our model gradually learns the meanings of words through a bootstrapping interaction between two types of probabilistic knowledge. Given an utterance–scene input received at time t, I_t=(U_t, S_t), the model first calculates an alignment probability $a_t(w|f)$ for each $w \in \mathrm{U}_t$ and each $f \in \mathrm{S}_t$, that captures how likely w and f are associated in I_t. This calculation uses the meaning probabilities learned up to time $t-1$, i.e., $p^{(t-1)}(f|w)$, as described in Step 1 below. The model then revises the meaning of the words in U_t by incorporating evidence from the alignment probabilities a_t, as in Step 2 below. This process is repeated for all input pairs I_t, one at a time.

Step 1: Calculating the alignment probabilities. We exploit the cross-situational learning assumption that words and features that have been associated in prior observations are more likely to be associated in the current input pair. Since the meaning probability, $p^{(t-1)}(f|w)$ (the probability of f being a meaning element of w), captures this prior strength of association, the higher this probability, the more likely it is that w is aligned with f in I_t. In other words, $a_t(w|f)$ is proportional to $p^{(t-1)}(f|w)$. We normalize this probability over all word–feature pairs for that feature f in the current input in order to capture the *relative* strength of association of w with f among the current possible alignments. Specifically, we use a smoothed version of the following formula:

$$a_t(w|f) = \frac{p^{(t-1)}(f|w)}{\sum\limits_{w' \in \mathrm{U}_t} p^{(t-1)}(f|w')} \qquad (1)$$

Step 2: Updating the word meanings. We next need to update the probabilities $p^{(t)}(f|w)$ based on the evidence from the current alignment probabilities. For each $w \in \mathrm{U}_t$ and $f \in \mathrm{S}_t$, we add the current alignment probability for w and f to the accumulated evidence from prior co-occurrences of w and f. We summarize this cross-situational evidence in the form of an association score, which is updated incrementally:

$$\mathrm{assoc}^{(t)}(w,f) = \mathrm{assoc}^{(t-1)}(w,f) + a_t(w|f) \qquad (2)$$

where $\mathrm{assoc}^{(t-1)}(w,m)$ is zero if w and f have not co-occurred prior to t. The association score of w and f is basically a weighted sum of their co-occurrence counts.

The model then uses these association scores to update the meaning of the words in the current input:

$$p^{(t)}(f|w) = \frac{\mathrm{assoc}^{(t)}(f, w) + \lambda(t)}{\sum\limits_{f_j \in \mathcal{M}} \mathrm{assoc}^{(t)}(f_j, w) + \beta \times \lambda(t)} \qquad (3)$$

where $\mathcal{M}$ is the set of all features encountered prior to or at time t, β is the expected number of distinct features, and $\lambda(t)$ is a smoothing factor, discussed in the next section.

Modeling Changes in Attention over Time

The model as presented above does not address the findings that children's attentional skills develop over time (e.g., Mundy et al., 2007). In particular, we assume that a child at earlier stages of cross-situational learning will consider that a word may be associated with some irrelevant semantic features, and that gradually, she will attend more and more to only the relevant features for the word. However, the input to our model consists of the words of an utterance paired with only semantic features that are relevant to those words. Thus to reflect a less-developed attentional mechanism, our model must be made to give some weight to unobserved word–feature pairs.

In fact, the model does provide for such a mechanism. The function $\lambda(t)$ in Eqn. (3) determines how much of the probability mass of $p(f|w)$ is allocated to unseen word–feature co-occurrences, and thus conversely, reflects the degree to which the model attends to the (relevant) observed co-occurrences. In the original model of Fazly et al. (2010a), however, λ was a very small constant, assuming a highly competent (and unchanging) attentional mechanism in place even in early stages of word learning. Here we have modified the model so that λ is a function of time, in order to simulate a learner whose ability to attend to relevant word–feature co-occurrences improves with age. Specifically, early on the model should give significant weight to unobserved word–feature pairs, reflecting immature attentional skills, but over time this weight should decrease, reflecting improved attentional processes that can appropriately focus on the observed word–feature pairs. This type of development can be achieved by devising λ as an inverse function of time: it starts reasonably large (allocating more probability mass to unseen word–feature pairs), and gradually decreases (increasing the probability mass assigned to observed pairs).

Modeling Normal and Late-talking Learners

The literature provides evidence for individual differences in the development of the ability of a learner to respond to joint attention (Morales et al., 2000). In particular, late-talking children exhibit difficulty in using communicative cues and in initiating joint attention with their partner (Paul & Shiffer, 1991; Rescorla & Merrin, 1998). Varying the λ function provides a way for our model to simulate such individual differences, by manipulating the rate of decrease in λ as a function of t. We assume that a "normal" learner's attentional abilities develop fairly quickly over time, modeled by a $\lambda(t)$ that decreases relatively rapidly (while still providing some allowance for unseen word–feature pairs). In contrast, for a late-talking learner, $\lambda(t)$ should decrease less rapidly. Thus we adopt this simple formulation:

$$\lambda(t) = \frac{1}{1+t^c}, \quad 0 < c \leq 1 \qquad (4)$$

where the value of c determines the rate at which λ decreases over time, and hence determines the type of the learner.

box: { IS-SQUARE:0.82, IS-SOLID:0.77, MADE-OF-WOOD:0.62, SIZE:0.4, MADE-OF-CHINA:0.18, HAS-LEGS:0.13, HAS-LEAVES:0.08, FLIES:0.03, ··· }

Figure 1: Sample sensory-motor features & their ratings for *box*.

Experimental Setup

Input Utterance–Scene Pairs

The training data for our model consists of a sequence of utterances, each paired with a set of semantic features as the scene representation. The utterances are extracted from the Manchester corpus (Theakston et al., 2001, from CHILDES MacWhinney, 2000), transcripts of conversations with 12 British children between the ages of 1;8 and 3;0. We use the child-directed speech (CDS) only, and lemmatize the words. The data from half of the children is used as development data, and the rest for our final experiments.

Because a manually-annotated semantic representation is not available for any such large corpus of CDS, we automatically generate a scene representation for each utterance. To do so, we create an input-generation lexicon which contains the "true" meaning $t(w)$ for each word w in our two semantic resources.[1] Each $t(w)$ is a vector over all possible semantic features. For adjectives and closed class words, each feature (taken from Harm, 2002) has value 1 in $t(w)$ if it is part of the meaning of the word, and 0 otherwise. For nouns and verbs, each feature (taken from Howell et al., 2005) has a value (between 0 and 1) derived from the relevancy ratings of 98 sensory-motor features for 352 nouns, and of 85 features for 91 verbs; see Figure 1 for an example. We then use $t(w)$ to probabilistically generate the set of observed semantic features for each word w in an utterance U. The scene representation is the union of this set of features for all w in U. For each word, we probabilistically sample the features in proportion to their value—i.e., features rated as more relevant to a word are more likely to appear in the scene representation when that word is used. We take this probabilistic approach to more realistically reflect the noise and uncertainty in the input, as well as the uncertainty of a child in determining the relevant meaning elements in a scene.

Evaluating the Learned Meanings

To measure how well the model has learned the meaning of a word w, we compare its learned meaning, $l(w)$ (a vector corresponding to the probability distribution $p(.|w)$), to its true meaning, $t(w)$ (a vector as described above). We calculate their similarity, $\text{sim}(l(w), t(w))$, using a simple vector distance measure, cosine. The higher the value of sim, the closer the learned meaning $l(w)$ is to the true meaning $t(w)$, and the better the meaning of w is learned.

Model Parameters

Recall that c in Eqn. (4) determines the level of learner's attentional abilities. In our experiments, we compare three different values for c: $c = 1$ yields a model, ND, corresponding to a normally-developing child; $c = 0.5$ yields a model, $\text{LT}_{.5}$, corresponding to a late talker with less severe difficulties; and $c = 0.25$ yields a model, $\text{LT}_{.25}$, corresponding to a late talker with more severe difficulties. (These values were chosen based on behaviour on development data; all models with $c < 1$ showed some degradation in learning performance.) We experiment with two versions of the LT settings to explore whether we can model two different types of LTs—those that eventually catch up to their normally-developing peers, and those that fail to do so.

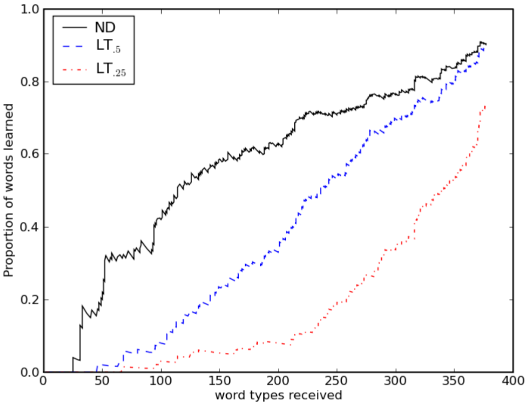

Figure 2: Proportion of noun/verb word types learned.

Experimental Results

As mentioned in the Introduction, several key behaviours have been observed regarding the learning of word meanings by LTs in comparison with their age-matched peers. First, LTs have both delayed vocabulary learning and a slower learning rate; while some LTs catch up to their peers, others do not. Second, LTs have more difficulty in learning novel words in an experimental setting. Third, the learned words of LTs seem to have less strong semantic connectedness among them. Here, we present three corresponding sets of experiments demonstrating that variation in the attention parameter in our model, reflected in the ND, $\text{LT}_{.5}$, and $\text{LT}_{.25}$ learners, can lead to each of these behaviours observed in children.

Patterns of Learning in the Models

LTs have a vocabulary size substantially below typical children at the same age. LTs not only show delayed development, but a different rate of vocabulary learning—i.e., they do not just start later, but learn more slowly (e.g., see Beckage et al., 2010, Figure 2). To see whether our LT learners differ from our ND learner in a similar way, we train each learner on 76K utterances, and look at how the proportion of learned words, out of all words the model has been exposed to, changes over time. We restrict our attention here to nouns and verbs, since we believe their semantic representation is more elaborated (and thus more realistic).

The vocabulary growth plots of the three learners, depicted in Figure 2, show interesting differences in accord with the

[1] We also add about 50 high-frequency words, mostly pronouns and proper nouns, with simple semantic features. Utterances containing words not found in either of the two resources, or our additional word list, are removed from the input.

patterns seen in children. First, the two LT models not only lag behind the ND model with respect to the onset of word learning, but also show a different rate and pattern of vocabulary learning (a very marked difference in the $\text{LT}_{.25}$ case). Whereas ND shows a sharp increase in the rate of vocabulary learning early on — 60% of words are learned by the time the model has received about 150 words — the two LT learners exhibit a slower and more gradual growth rate. In addition, the two LT models differ from each other. As is observed in children, some learners (as with $\text{LT}_{.5}$) who start off slow catch up in vocabulary learning, while others (as with $\text{LT}_{.25}$) continue indefinitely to lag behind their age-matched peers. This distinction is important to understand more fully, since the latter are at risk for SLI.

Novel Word Learning Experiments

To understand how the vocabulary learning process of LTs differs from that of typical children, psycholinguists test the performance of the two groups in a contrived novel word learning situation: An experimenter first introduces a novel word and its novel referent to the child, and then examines the child's knowledge of the target (novel) word through explicit tests of comprehension and/or production.

Here, we simulate a simplified version of the novel word learning experiment of Weismer and Evans (2002). First, we train the model on some number of corpus inputs, simulating a child's normal word learning experience. We then introduce a novel noun to the model in several teaching trials as follows: As our novel noun, we randomly pick a noun that has not occurred in the training utterances. To simulate use of the novel noun in natural utterances, we add the noun to an actual (as yet unseen) utterance from the corpus, and add its probabilistically-generated meaning to the corresponding scene. We train our ND and LT learners on N such teaching utterance–scene pairs as usual.

To examine the novel word learning ability of each learner, we repeat the above process for 106 novel nouns, for N = 3 teaching trials, and for different amounts of prior training utterances (here, 10K, 30K, or 60K), and test as follows.

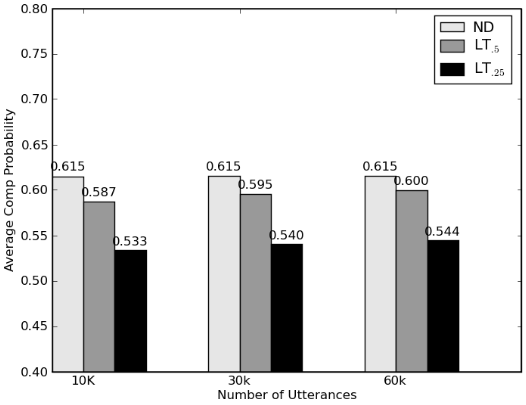

Figure 3: Average Comp probabilities of learners over time.

Comprehension. To test comprehension of a recently-taught novel word, the experimenter asks the child to find the referent of the novel word, when presented with the novel object along with one or more familiar objects. Note that in our computational experimental setting, the "object" corresponding to a word is its true meaning, $t(w)$ (i.e., there is no distinction between the true meaning of a word and a referent corresponding to that meaning). We pair each novel object $t(w_{\text{N}})$ with one familiar object $t(w_{\text{F}})$, and calculate the likelihood of selecting each of these in response to w_{N} as the stimulus. Specifically, we test whether the model's learned representation of the meaning of the novel noun, $l(w_{\text{N}})$, is closer to the true meaning of the novel noun, $t(w_{\text{N}})$, or that of the familiar noun, $t(w_{\text{F}})$. We use the Shepard-Luce rule (Shepard, 1957; Luce, 1959), to calculate the probability of choosing the novel object in response to the novel word in this forced-choice task:

$$\begin{aligned}\text{Comp}(w_{\text{N}}) &= \text{P}(t(w_{\text{N}})|w_{\text{N}})) \\ &= \frac{\text{sim}(l(w_{\text{N}}),t(w_{\text{N}}))}{\sum_{w' \in \{w_{\text{N}},w_{\text{F}}\}} \text{sim}(l(w_{\text{N}}),t(w'))} \end{aligned} \qquad (5)$$

To ensure that w_{F} is familiar to the model, we select it from nouns with a minimum frequency of 5.

Production. The production test evaluates the ability of a learner to produce a recently-taught novel word when presented with the corresponding novel object. We calculate the probability that a learner produces the target novel noun w_{N} given its true meaning $t(w_{\text{N}})$, as in:

$$\begin{aligned}\text{Prod}(w_{\text{N}}) &= \text{P}(w_{\text{N}}|t(w_{\text{N}})) \\ &= \frac{\text{sim}(l(w_{\text{N}}),t(w_{\text{N}}))}{\sum_{w' \in \mathcal{W}} \text{sim}(l(w'),t(w_{\text{N}}))} \end{aligned} \qquad (6)$$

where $\mathcal{W}$ is the set of all words that we assume the model *could* produce in response to $t(w_{\text{N}})$. Here $\mathcal{W}$ consists of all words with a minimum frequency of 3.[2] Given the above formulation, the production probability of a word is high if: (i) the learned meaning of the word and its true meaning are sufficiently similar; and (ii) this similarity is much higher than the similarity between the target object and the learned meaning of the other words.

Analysis of the Results. The Comp and Prod probabilities of the three learners, averaged over the 106 novel test words, are given in Figure 3 and Figure 4, respectively. Similar to what Weismer and Evans (2002) reported, here we can see that ND performs significantly better than $\text{LT}_{.25}$ in the comprehension test, at all three stages of learning (t-test: $p \ll 0.01$). In contrast, we observe a significant difference between the comprehension performance of $\text{LT}_{.5}$ and that of ND only at early stages (after processing 10K and 30K utterances; $p < 0.01$), again suggesting that $\text{LT}_{.5}$ may represent a group of learners who start off late, but eventually catch up to their

[2] We use the frequency of the novel word as this threshold.

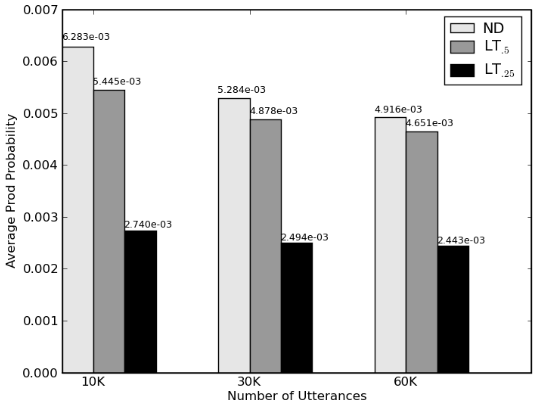

Figure 4: Average Prod probabilities of learners over time.

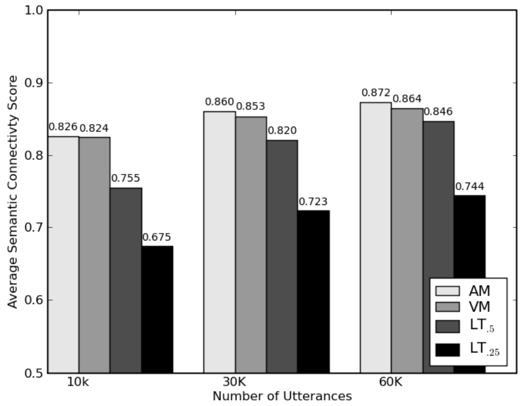

Figure 5: Semantic connectivity scores of learners over time.

normal peers. In the production test, ND performs significantly better than both LTs during all the stages of learning; however, the difference between ND and $\text{LT}_{0.5}$ is decreasing over time.

One issue should be noted here: The production scores of all learners decrease over time. This happens because at later stages the learners know more words, many of which are semantically related (such as *cat*, *dog*, *lion*, etc.). Thus, the denominator in Eqn. (6) increases over time due to encountering more words that are semantically similar to the target word (to be produced), and this results in lower production probabilities. Future work will need to consider alternative probabilistic formulations of production, and explore the degree to which our particular meaning representation contributes to the observed effect.

Semantic Organization Experiments

Late talkers have been shown to not only learn more *slowly* than their age-matched normally developing children, but also to be learning *differently* (e.g., Beckage et al., 2010; Sheng & McGregor, 2010; Jones & Smith, 2005). In particular, Beckage et al. (2010) examine the vocabulary of several late talking and normally developing children, and show that the learned words of late talkers are less semantically connected than those of normally developing children.

Recall that in our input representation, features are generated probabilistically to reflect the noise and uncertainty in the input and/or the uncertainty of a child's perception of the relevant meanings for a word. Moreover, in our model, the weaker attentional abilities of our LT learners (especially $\text{LT}_{.25}$) requires them to observe a word–feature pair more times in order to learn that association. This can lead to (some) semantic features of the word being less well learned. The more sparsely learned features may then lead to less semantic connectivity among the words. Here, we compare the "semantic organization" of nouns for our two LT learners, with those of two normally-developing learners: an age-matched ND (trained on the same number of utterances as the two LTs), and a vocabulary-matched (younger) ND (trained on a proportion of these utterances to account for the age difference).

For each learner, we build a semantic network as follows: We connect each word to all other words the learner has encountered during training, weighting each connection by the similarity between the learned meanings of the connected words. We expect the semantic networks of the two normal learners (the age-matched, AM, and the vocabulary-matched, VM) to be more connected compared to the two LT learners. We calculate a semantic connectivity score for each learner by comparing the connectivity of the nouns in its network to that of nouns in a gold-standard network formed analogously using the true meanings of words. (As in other experiments, here we focus on nouns because of their more elaborate semantic representation.) We represent the connection weights of each noun in a network as a vector, and measure the similarity of the noun's connections in a learned network and in the gold-standard network using cosine over the two corresponding vectors. The average of these vector similarities over all nouns is taken as the semantic connectivity score of the target learned network.

Figure 5 shows the connectivity scores for the four learners trained on different amounts of input. The results show that, in line with the findings of Beckage et al. (2010), both AM and VM learners have more semantic connectivity in their learned knowledge of nouns compared to both LTs (all differences are statistically significant; $p \ll 0.01$). Once again, $\text{LT}_{.5}$ seems to be catching up to the ND learners: The semantic connectivity of $\text{LT}_{.5}$ is getting closer to that of AM at the latest stage of learning.

Conclusions

There are several possible explanations behind language deficiencies in late talkers, such as inadequacies in their general cognitive abilities (e.g., attention, categorization, and memory skills), or in the quality and quantity of their linguistic input. Here, we have focused on modeling variations in the development of attentional abilities in normal and late-talking children. Specifically, we have incorporated an attention mechanism into an existing model of learning word meanings in context, enabling us to model both a learner's cognitive development over time, as well as some individual differences among learners in lexical development.

Results of our experiments comparing late-talking (LT) and normally-developing (ND) learners are compatible with the psycholinguistic findings: Compared to our ND model, the LT model with severe difficulties ($\text{LT}_{.25}$) exhibits marked delay in the onset of vocabulary learning, performs significantly worse in learning novel words, and has less strong semantic connections among its learned words. In contrast, the $\text{LT}_{.5}$ learner (with less severe difficulties) is significantly different from ND only at earlier stages of development, reflecting some normal degree of variation in vocabulary learning.

The model presented here has the potential for studying many more issues pertaining to normal versus impaired lexical development. One important issue that needs further investigation is the (possibly differential) effect of the linguistic input on lexical development in ND and LT children. In fact, our probabilistic input generation method enables us to vary the input quality, possibly corresponding to the use of social cues or some other attentional mechanism children use to hone in on relevant word–meaning associations.

Another future direction is to further examine the effect of semantic connectedness among words in their acquisition, in both ND and LT children. Late talkers have been shown to do worse in explicit word association tasks (Sheng & McGregor, 2010), as well as in recognizing abstract categories (e.g., Jones & Smith, 2005). By adding explicit categorization abilities to our model (e.g., as in Alishahi & Fazly, 2010) we can further investigate the differences of our various learners, both in capturing the semantic connections among words, and in using these connections to bootstrap word learning.

References

Alishahi, A., & Fazly, A. (2010). Integrating syntactic knowledge into a model of cross-situational word learning. In *Proc. of CogSci'10.*

Beckage, N. M., Smith, L. B., & Hills, T. (2010). Semantic network connectivity is related to vocabulary growth in children. In *Proc. of CogSci'10.*

Desmarais, C., Sylvestre, A., Meyer, F., Bairati, I., & Rouleau, N. (2008). Systematic review of the literature on characteristics of late-talking toddlers. *Int'l J. of Lang. and Communication Disorders*, *43*(4), 361–389.

Fazly, A., Ahmadi-fakhr, F., Alishahi, A., & Stevenson, S. (2010b). Cross-situational learning of low frequency words: The role of context familiarity and age of exposure. In *Proc. of CogSci'10* (pp. 2615–20).

Fazly, A., Alishahi, A., & Stevenson, S. (2010a). A probabilistic computational model of cross-situational word learning. *Cognitive Science*, *34*(6), 1017–1063.

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In *NIPS'07* (Vol. 20).

Harm, M. W. (2002). *Building large scale distributed semantic feature sets with WordNet* (Tech. Rep. No. PDP.CNS.02.1). Carnegie Mellon University.

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

Jones, S. S., & Smith, L. B. (2005). Object name learning and object perception: a deficit in late talkers. *J. of Child Lang.*, *32*, 223–240.

Luce, R. D. (1959). *Individual Choice Behavior: A Theoretical Analysis.* New York: Wiley.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk* (3rd ed., Vol. 2: The Database). Erlbaum.

Morales, M., Mundy, P., Delgado, C. E. F., Yale, M., Messinger, D., Neal, R., et al. (2000). Responding to joint attention across the 6- through 24-month age period and early language acquisition. *Journal of Applied Developmental Psychology*, *21*(3), 283–298.

Mundy, P., Block, J., Delgado, C., Pomares, Y., Hecke, A. V. V., & Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child Development*, *78*(3), 938–954.

Paul, R., & Elwood, T. J. (1991). Maternal linguistic input to toddlers with slow expressive language development. *J. of Speech and Hearing Research*, *34*, 982–988.

Paul, R., & Shiffer, M. E. (1991). Communicative initiations in normal and late-talking toddlers. *Applied Psycholing.*, *12*, 419–431.

Quine, W. (1960). *Word and object.* MIT Press.

Rescorla, L., & Merrin, L. (1998). Communicative intent in late-talking toddlers. *Applied Psycholing.*, *19*, 398–414.

Rowe, M. L. (2008). Child-directed speech: relation to socioeconomic status, knowledge of child development and child vocabulary skill. *J. of Child Lang.*, *35*, 185–205.

Sheng, L., & McGregor, K. K. (2010). Lexical–semantic organization in children with specific language impairment. *J. of Speech, Lang., & Hearing Research*, *53*, 146–159.

Shepard, R. (1957). Stimulus and response generalization: a stochastic model, relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.

Stokes, S. F., & Klee, T. (2009). Factors that influence vocabulary development in two-year-old children. *J. of Child Psychology*, *50*(4), 498–505.

Thal, D. J., Bates, E., Goodman, J., & Jahn-Samilo, J. (1997). Continuity of language abilities: An exploratory study of late- and early-talking toddlers. *Developmental Neuropsychology*, *13*(3), 239–273.

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

Weismer, S. E., & Evans, J. L. (2002). The role of processing limitations in early identification of specific language impairment. *Topics in Language Disorders*, *22*(3), 15–29.

Yu, C., & Ballard, D. H. (2008). A unified model of early word learning: Integrating statistical and social cues. *J. of Neurocomputing*, *70*(13–15), 2149–65.