Précis of "Computational Modeling of Word Learning: The Role of Cognitive Processes"

Aida Nematzadeh

Introduction

Word learning is a significant part of child language learning: comprehending the meaning of individual words is the first step in understanding larger linguistic units such as phrases and sentences. This knowledge of word meanings also helps a child understand the relations among the words in a sentence and thus facilitates the acquisition of rules of the language (syntax). Although word learning seems effortless and simple, it is a complex process that we do not fully understand: Children start with no prior knowledge of word meanings, are not explicitly being taught, and receive highly noisy and ambiguous input. Moreover, child word learning happens simultaneously with and depends on the development of other cognitive processes such as memory, attention, and categorization.

This thesis takes a multidisciplinary approach to shed light on the mechanisms underlying child word learning – acquisition of word meanings, their semantic relations, and the organization of this information to enable fast access. More specifically, I developed a computational model of word learning that is subject to the same kind of constraints on input data and processing as people. The model starts with no prior knowledge and very limited assumptions about the word learning problem: I assume that linguistic knowledge is not innate, and children learn word meanings by processing the input they receive using general cognitive mechanisms (such as memory, attention, and statistical learning). Thus, this thesis is in line with and supports the view that word learning is a result of applying *domain-general* cognitive abilities to the linguistic input with no need for a "special cognitive system" (*e.g.*, Tomasello, 2005). Moreover, this assumption enables us to examine the power of the learning algorithms – what is learnable and not learnable given the data.

I showed that semantic acquisition in the model resembles learning in children by evaluating the model against a wide range of empirical data from the cognitive and developmental literature. As a result, compared to models that only account for some specific data, the model's novel predictions are more reliable and can guide the design of further behavioral experiments. Moreover, the computational basis of the model provides a testable implementation of the proposed hypotheses on human semantic acquisition. It also enables full control over experimental settings, making it possible to examine a vast number of conditions difficult or impossible to achieve with human subjects.

I have used the model to explain or examine the possible factors behind several phenomena observed in child and adult learning. In the following sections, I describe the most important components of this research in more detail. As its key contributions, my thesis establishes that:

- No supervision or built-in knowledge about language is needed to learn word meanings and their semantic connections.
- To account for child semantic acquisition, we need to integrate other cognitive mechanisms (*e.g.*, forgetting and attention) into the model's statistical learning (which is often ignored in word learning models). Specifically, Chapters 3 to 5 demonstrate that three important phenomena observed in child vocabulary development (individual differences, spacing effects in learning, and learning semantic relations among words respectively) can only be explained when these cognitive mechanisms are integrated with word learning.

Beyond the scientific advances that result from understanding word learning mechanisms, there are also practical applications, for example in identification, prevention, and treatment of various language deficits, and in devising educational methods to improve students' learning. Moreover, a better understanding of human language acquisition and organization of semantic knowledge can lead to building computer systems that better interact with people, because these systems often need to address the same challenges as those faced by children.

Chapter 2: Word Learning in Children and Computational Models

Chapter 2 discusses the previous behavioral experiments and computational modeling research on child semantic acquisition. It explains what makes word learning a challenging problem, summarize the key theories on child word learning including predominant patterns observed in word learning, and mechanisms and constraints involved in it. Moreover, it provides an overview of the major computational models of word learning, as well as the word learning framework that my word learning model builds on. In this section, I briefly explain how our model represents and learns word meanings.

Given a word learning scenario, there are potentially many possible mappings between words in a sentence and their meanings, from which only some mappings are correct. One of the most dominant mechanisms proposed for vocabulary acquisition is *cross-situational learning*: people learn word meanings by recognizing and tracking statistical regularities among the contexts of a word's usage across various situations, enabling them to narrow in on the meaning of a word that holds across its usages (*e.g.*, Siskind, 1996; Yu and Smith, 2007).

Our computational model is a probabilistic cross-situational word learner; its learning algorithm is incremental and involves limited calculations, thus satisfying basic cognitive plausibility requirements. Our work modifies and extends the model of Fazly et al. (2010) to incorporate various cognitive mechanisms (such as attention, memory, and categorization). The input to our model approximates child word learning data and consists of pairs of *utterances* (the words a child hears) and *scenes* (the semantic features representing the meaning of those words), as shown in Table 1.

Utterance: {*let*, *find*, *a*, *picture*, *to*, *color* }
Scene: {LET, PRONOUN, HAS_POSSESSION, CAUSE, ARTIFACT, WHOLE, CHANGE, ...}

Table 1: A sample utterance-scene pair.

Given such an utterance-scene input pair, for each word w and semantic feature f, the model incrementally learns $P(f|w)$ from co-occurrences of w and f across all the utterance-scene pairs. For each word, the probability distribution over all semantic features, $P(.|w)$, represents the word's *meaning*. The estimation of $P(.|w)$ is made possible by introducing a set of latent variables, *alignments*, that correspond to the possible mappings between words and features in a given utterance–scene pair. The learning problem is then to find the mappings that best explain the data, which is solved using an incremental version of the expectation–maximization (EM) algorithm (Neal and Hinton, 1998).

Chapter 3: Individual Differences in Word Learning

Chapter 3 examines the individual differences in word learning through computational modeling. Although most children are successful word learners, *late talkers* exhibit substantial delay in word learning. Since these children are at risk for *specific language impairment* – may never reach the normal level of language efficacy – identifying factors involved in late talking is a significant research problem. Previous research has identified different environmental and cognitive factors that might contribute to late talking (*e.g.*, Paul and Elwood, 1991; Jones and Smith, 2005). In particular, late-talking children exhibit difficulty in using communicative cues and initiating joint attention with their partner (Paul and Shiffer, 1991; Rescorla and Merrin, 1998). There is also evidence for individual differences in the development of the ability of a learner to respond to joint attention (Morales et al., 2000). However, it is not clear how these factors contribute to the patterns observed in word learning of late talkers and normally developing learners.

We propose a computational model that enables us to thoroughly examine the possible factors behind late talking, specifically, individual differences in attentional mechanisms and categorization. To our knowledge, no previous computational model of word learning in context demonstrates the effects of factors that could contribute to late talking.

Our model incorporates an attentional mechanism that gradually improves over time, enabling it to focus (more or less) on the features relevant to a word. We simulate a continuum of learners by parameterizing the rate of development of this mechanism, mimicking normally-developing, temporarily delayed, and language-impaired children such that the normally-developing learner has the fastest development rate. Because the attentional mechanism impacts the learning algorithm of the model, the normally-developing and late-talking learners differ in the quality of their learned meanings. We extend our model with a categorization mechanism to further study how individual differences between learners give rise to the differences in abstract knowledge of categories emerging from learned words, and how this affects their subsequent word learning.

Our simulated late-talking learners, similar to late-talking children, exhibit a delayed and slower vocabulary growth in addition to a less semantically-connected vocabulary compared to normally-developing children. Our model also successfully produces the differences observed in subgroups of late talkers, that is, temporarily delayed and language-impaired children (Section 3.3). Our results (Section 3.5.2) suggest that the vocabulary

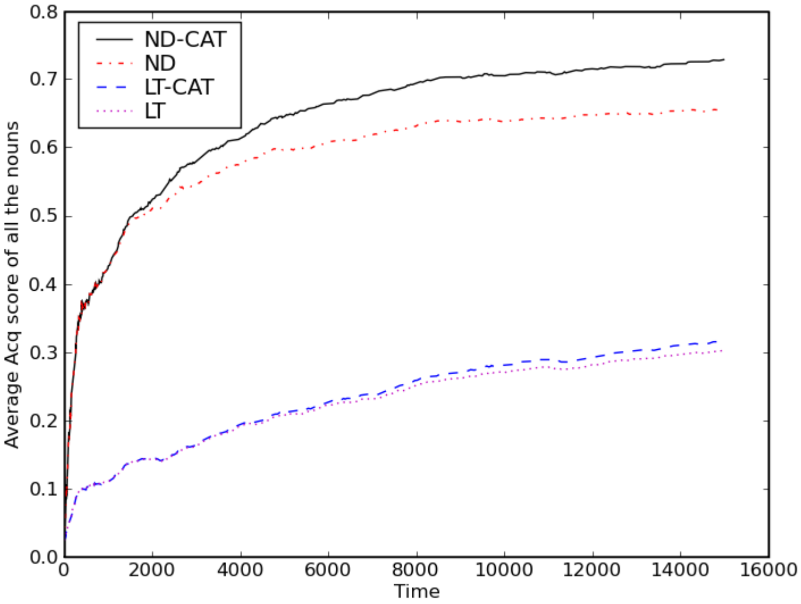


Figure 1: Change in the average acquisition score (Acq) of all nouns over time (measured in number of processed utterances) for normally-developing (ND) and late-talking (LT) learners; ND-CAT and LT-CAT use category information during learning. The normally-developing learner benefits from using the learned categories in word learning (compare ND and ND-CAT); the gap between ND and ND-CAT increases over time, because the quality of the learned meaning and consequently categories improve as the model processes more input. There is no difference between LT and LT-CAT because the LT learner does not form informative categories.

composition of late-talking and normally-developing learners differ, at least partially, due to a deficit in the attentional abilities of late-talking learners, which also results in the learning of weaker abstract knowledge (semantic categories). As a result, the late-talking learner, as opposed to the normally-developing learner, does not benefit from the learned categories in identifying the correct meaning of the words (Section 3.5.3): Figure 1 shows the overall pattern of word learning (measured by the average acquisition score) for both learners with and without category knowledge.

Moreover, we use our model to examine the structure of each learner's semantic network (which represents words and the relations among them). The structure of this network is significant as it might reveal aspects of the developmental process that leads to the network. We find that the learned semantic knowledge of a learner that simulates a normally-developing child reflects the structural properties found in adult semantic networks of words (see Figure 2a). More specifically, the semantic network of the normally-developing network exhibits a small-world structure – a sparse network with highly-connected local sub-networks, where the sub-networks are connected through high-degree hubs (nodes with many neighbours). In contrast, the network of a late-talking learner does not exhibit these properties: As shown in Figure 2b, most words are connected with no clear grouping of the semantically-similar words.

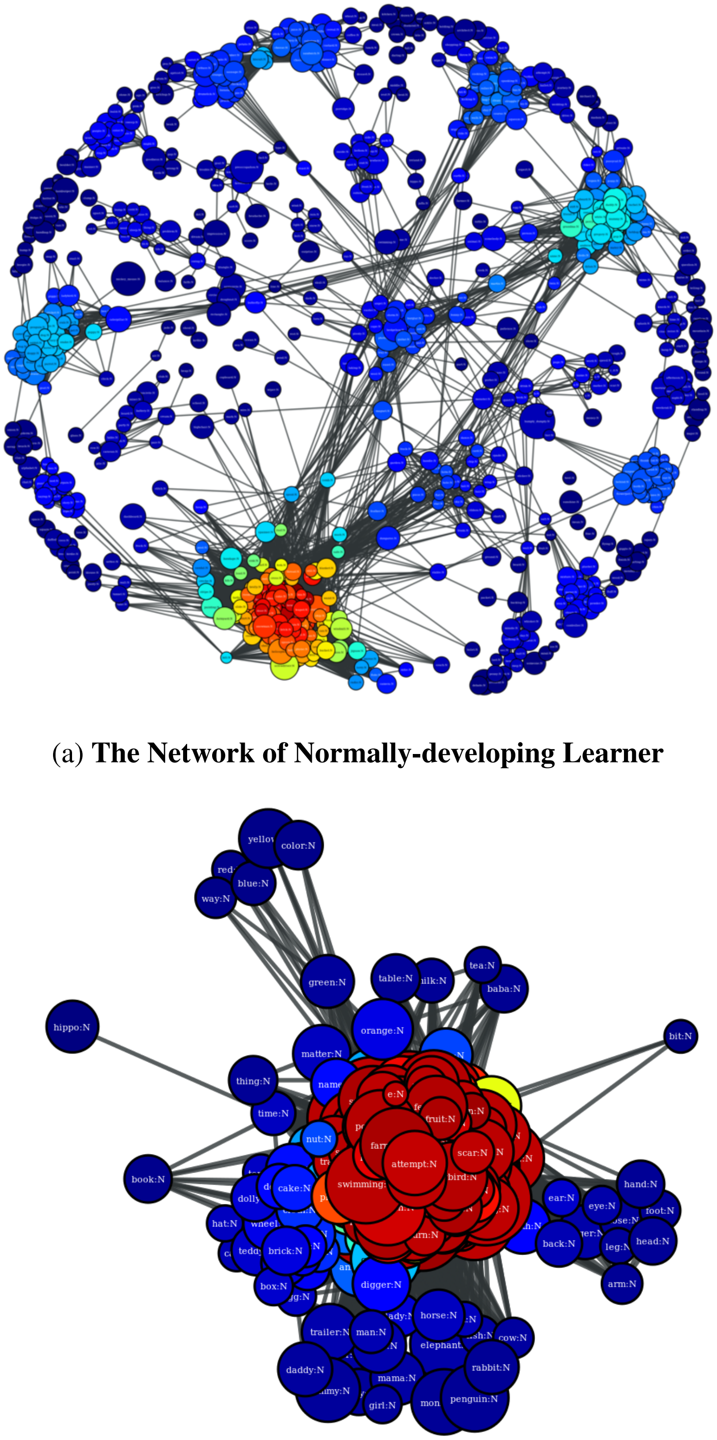
To summarize, our results show that both the quality and structure of the semantic knowledge differ in normally-developing and late-talking learners.

The work presented in this chapter has potential clinical applications: The predictions of our model can help speech-language pathologists design experiments for identifying early signs of late talking. The work presented in this chapter has been published in Nematzadeh et al. (2011), Nematzadeh et al. (2012b), and Nematzadeh et al. (2014a).

Chapter 4: Memory, Attention, and Word Learning

Chapter 4 investigates the interaction of memory and attention in word learning. One area with a wealth of relevant experimental evidence is the *spacing effect* in learning (*e.g.*, Ebbinghaus, 1885). The observation is that people generally show better learning when the presentations of the target items to be learned are "spaced" — i.e., distributed over a period of time — instead of being "massed" — i.e., presented together one after the other. Investigations of the spacing effect often use a word learning task as the target learning event, and look at the performance of adults as well as children (*e.g.*, Glenberg, 1976; Pavlik and Anderson, 2005; Vlach et al., 2008). While such work involves controlled laboratory conditions, the spacing effect is very robust across domains and tasks (Dempster, 1989), suggesting that the underlying cognitive processes likely play a role in natural conditions of word learning as well.

Hypotheses for the spacing effect have included both memory limitations and attention. Many researchers assume that the process of forgetting is responsible for the improved performance in the spaced presentation (*e.g.*, Melton, 1967; Jacoby, 1978), while others



(a) **The Network of Normally-developing Learner**

(b) **The Network of Late-talking Learner**

Figure 2: The network of (a) normally-developing (ND), and (b) late-talking (LT) learners with all words connected by learned meanings. In ND's network, there are sub-networks of semantically-related words that are connected to each other; however, in LT's network all words are connected without forming meaningful grouping of words.

propose that subjects attend more to items in the spaced presentation because accessing less recent (more novel) items in memory requires more effort or attention (*e.g.*, Hintzman, 1974). However, the precise relation between forgetting and improved learning as well as the precise attentional mechanism at work in the spacing experiments are not completely understood.

We hypothesize that both forgetting and attention to novelty play a role in the spacing effect in word learning. We examine this hypothesis by considering memory constraints and attentional mechanisms in the context of our computational model of word-meaning acquisition. More specifically, we extend our word learning model with two mechanisms: (i) a forgetting mechanism that causes the learned associations between words and meanings to decay over time; and (ii) a mechanism that simulates the effects of attention to novelty on in-the-moment learning. The result is a more cognitively plausible word learning model that includes a precise formulation of both forgetting and attention to novelty. We also use our model to investigate the interaction of these mechanisms, which is extremely difficult to achieve in experiments with human subjects.

Our model accounts for experimental results on children as well as several patterns observed in adults (Section 4.3). Moreover, in simulations using this new model, we show that a possible explanation for the spacing effect is the interplay of these two mechanisms neither of which on its own accounts for the effect (Section 4.3.3). Figure 3 demonstrates one of the observed patterns in spacing experiments, the spacing crossover interaction, that our model successfully replicates (Section 4.3.4). In these experiments, with smaller spacing intervals, a shorter retention interval (such as our "immediate" condition) leads to better results, but with larger spacing intervals, a longer retention interval (such as our "later" condition) leads to better results (Bahrick, 1979; Pavlik and Anderson, 2005). The spacing-crossover experiment we modeled differs from other spacing effect experiments in that it uses a longer presentation duration for the learning events. We hypothesize that this longer presentation results in better learning, and consequently a decreased level of forgetting (which we model with a smaller decay rate). Our model suggests an explanation for the observed crossover: in tasks which strengthen the learning of the target item — and thus lessen the effect of forgetting — we expect to see a benefit of later retention trials in experiments with people.

In Section 4.5, we use our model to examine the possible explanatory factors behind *desirable difficulties* in a cross-situational word learning experiment where – paradoxi-

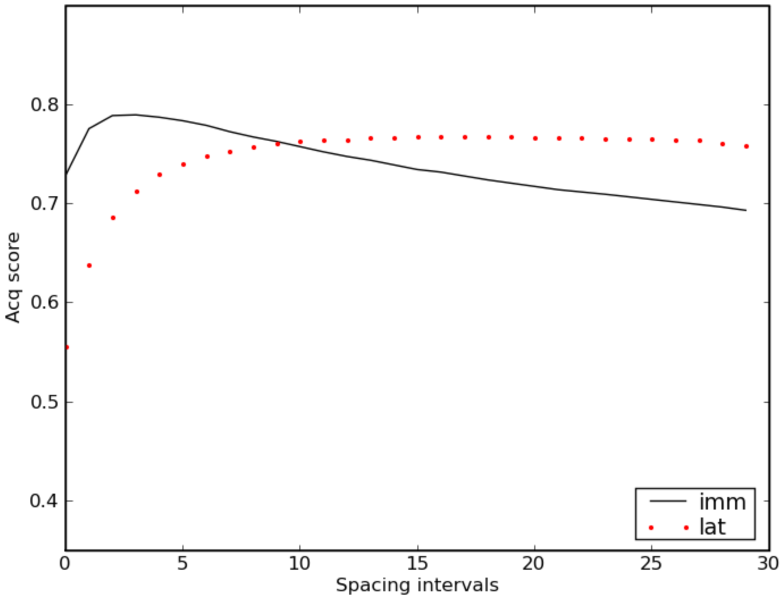


Figure 3: Average acquisition (Acq) score of the novel words over spacing intervals for two test intervals, (imm)ediately after training and in a (lat)er condition. When tested after a delay (the lat condition), the performance of the model is better for larger spacing intervals that interestingly are more difficult learning conditions.

cally – difficulties of the word learning situation promote long-term learning (Vlach and Sandhofer, 2010). Notably, the experimental results were not clearly pointing to the factors causing the patterns observed in the performance of the human participants. Using our model, we have suggested that an interaction between two factors (the within-trial ambiguity of the learning trials, and the presentation duration of each trial) might explain the observed patterns. In addition, our results point to other distributional characteristics of the input (experimental stimuli) that might have an impact on the performance of the learner. These findings illustrate the role of computational modeling, not only in explaining observed human behaviour, but also in fully understanding the factors involved in a phenomenon. There are several factors involved in a cross-situational word learning experiment, such as the contextual familiarity of words, and the average spacing interval of words. Our findings signify the importance of controlling for these factors in order to understand the reasons behind the observed patterns. But it is difficult do so in human experiments because the factors can interact in complex ways.

An future direction for this chapter is investigating methods for finding the *optimal spacing* in word learning – the learning arrangement that results in the best performance. Such methods can be used to improve children's word learning. The work presented in this chapter has been published in Nematzadeh et al. (2012a) and Nematzadeh et al. (2013).

Chapter 5: Semantic Network Learning

Children simultaneously learn word meanings and the semantic relations among words, and also efficiently organize this information. A presumed outcome of this development is the formation of a semantic network – a graph of words as nodes and semantic relations as edges – that reflects this semantic knowledge (*e.g.*, Collins and Loftus, 1975). Steyvers and Tenenbaum (2005) show that a semantic network that encodes adult-level knowledge of words exhibits a *small-world* structure. That is, it is a sparse network with highly-connected local sub-networks, where these sub-networks are connected through high-degree hubs (nodes with many neighbours). The structure of semantic knowledge is significant as it impacts how word meanings are stored in, searched for, and retrieved from memory (Steyvers and Tenenbaum, 2005).

An important open question is how such a semantic network can be gradually acquired as word meanings are learned. In this chapter, we extend our model to provide a cognitively-plausible and unified account for both acquiring and representing semantic knowledge, in particular, simultaneously learning words and creating a semantic network structure over them. The requirements for cognitive plausibility enforce some constraints on the semantic network creation process. The first requirement is incrementality, which means that the model gradually builds the network as it processes the input. Also, the number of computations the model performs at each step must be limited.

To satisfy these requirements in our model, as we learn words incrementally, we also structure those words into a semantic network based on the (partially) learned meanings. More specifically, when adding or updating a word's connections in the network, the model only looks at a subset of words rather than comparing the target word to all the nodes in the network. For a given word, this subset of words is selected by using the evolving knowledge of semantic connections among words as well as their usage context. To capture the semantic connections among words, the model incrementally forms semantic clusters as it processes each word.

Our model is successful in creating networks that reflect the semantic connectivity and structure of adult semantic knowledge. To evaluate the semantic connectivity of our learned network, we compare these learned semantic distances to the "gold-standard" similarity scores that are calculated using a similarity measure based on WordNet – a manually created lexical hierarchy (Fellbaum, 1998). The network structure is examined to see whether, similar to adults, it exhibits a small-world structure, *i.e.*, has certain connectivity properties – short paths and highly-connected neighborhoods – that are captured by various graph metrics. To summarize, the model's success stems from incorporation of the knowledge of semantic categories and information inherent in the context of words.

A significant application of this work is *unsupervised ontology learning*, specifically, adding new word senses to existing ontologies without having to compare them to all the existing words in the ontology. The work presented in this chapter has been published in Nematzadeh et al. (2014b).

Chapter 6: Conclusions

This thesis uses computational modeling to investigate the mechanisms responsible for vocabulary development. I have designed and developed a computational model that mimics child vocabulary development – it learns the meaning of words along with the semantic connections among them. Moreover, in the model, word learning is naturally integrated with other cognitive processes such as memory and attention. This thesis demonstrates that the domain-general learning mechanisms are sufficient for modeling word learning. Moreover, it shows that word learning in the context of cognitive processes is needed to predict the patterns observed in child vocabulary development.

The model presented in this thesis has been extended to explain how children learn to generalize a word to the appropriate level of a hierarchical taxonomy, *i.e.*, *dog* refers to all dogs, not just the Dalmatians or any animal. Our model, without incorporating any additional biases and, simply, through learning meanings for words accounts for the empirical word generalization data. More specifically, we find that capturing the interaction of category and instance frequencies in the data (*e.g.*, number of Dalmatians versus breeds of dogs) is the key factor in modeling the observed generalization data in children

(Nematzadeh et al., 2015).

A long-term future direction is to learn meaning representations for sentences. Understanding the meaning of individual words and their semantic relations is not sufficient for comprehending the meaning of sentences. To understand a sentence's meaning, children need to recognize how the meaning of its words relate and interact. Consider the sentences "Sebastian ate the apple" and "The apple was eaten by Sebastian". To recognize that these sentences express similar information, a computational model needs to (a) know the word meanings, and (b) identify the *thematic relations* between verbs and their arguments, *i.e.*, how noun phrases relate to the verbs: in both sentences, despite the difference in word orders, "Sebastian" *performed* the action of eating, and "apple" was the entity *acted on*. Simultaneous learning of word meanings and these thematic relations enables the model to provide rich meaning representations for sentences.

To conclude, computational modeling is a powerful tool for studying language acquisition, and has gained tremendous popularity in the last decade. I believe that, instead of building small independent models that only explain some specific data, we should design unified models that account for all the "significant" data available for a given phenomenon. Such frameworks, when validated thoroughly, can produce reliable predictions. My thesis is in line with this research philosophy; our computational model provides a general framework for studying vocabulary development: it is used to examine various aspects of word learning, predicts several patterns observed in word learning, and produces novel predictions.

References

H. P. Bahrick. Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, 108(3):296–308, 1979.

A. M. Collins and E. F. Loftus. A spreading-activation theory of semantic processing. *Psychological Review*, 82(6):407, 1975.

F. Dempster. Spacing effects and their implications for theory and practice. *Educational Psychology Review*, 1:309–330, 1989.

H. Ebbinghaus. *Memory: A contribution to experimental psychology*. New York, Teachers College, Columbia University, 1885.

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

C. Fellbaum, editor. *WordNet, An Electronic Lexical Database*. MIT Press, 1998.

A. M. Glenberg. Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior*, 15(1), 1976.

D. L. Hintzman. Theoretical implications of the spacing effect. In R. Solso, editor, *Theories in cognitive psychology: the Loyola symposium*. Lawrence Erlbaum Associates, 1974.

L. L. Jacoby. On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, 17(6):649 – 667, 1978.

S. S. Jones and L. B. Smith. Object name learning and object perception: a deficit in late talkers. *Journal of Child Language*, 32:223–240, 2005.

A. W. Melton. Repetition and retrieval from memory. *Science*, 158:532, 1967.

M. Morales, P. Mundy, C. E. F. Delgado, M. Yale, D. Messinger, R. Neal, and H. K. Schwartz. 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, 2000.

R. M. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

A. Nematzadeh, A. Fazly, and S. Stevenson. A computational study of late talking in word-meaning acquisition. In *Proceedings of the 33th Annual Conference of the Cognitive Science Society*, pages 705–710, 2011.

A. Nematzadeh, A. Fazly, and S. Stevenson. A computational model of memory, attention, and word learning. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pages 80–89. Association for Computational Linguistics, 2012a.

A. Nematzadeh, A. Fazly, and S. Stevenson. Interaction of word learning and semantic category formation in late talking. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society*, pages 2085–2090, 2012b.

A. Nematzadeh, A. Fazly, and S. Stevenson. Desirable difficulty in learning: A computational investigation. In *Proceedings of the 35th Annual Conference of the Cognitive Science Society*, pages 1073–1078, 2013.

A. Nematzadeh, A. Fazly, and S. Stevenson. Structural differences in the semantic networks of simulated word learners. In *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 1072–1077, 2014a.

A. Nematzadeh, A. Fazly, and S. Stevenson. A cognitive model of semantic network learning. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 244–254. Association for Computational Linguistics, 2014b.

A. Nematzadeh, E. Grant, and S. Stevenson. A computational cognitive model of novel word generalization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1795–1804, Lisbon, Portugal, 2015. Association for Computational Linguistics.

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

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

P. I. Pavlik and J. R. Anderson. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29:559–586, 2005.

L. Rescorla and L. Merrin. Communicative intent in late-talking toddlers. *Applied Psycholinguistics*, 19:398–414, 1998.

J. M. Siskind. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition*, 61:39–91, 1996.

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

M. Tomasello. *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, March 2005. ISBN 674017641.

H. A. Vlach and C. M. Sandhofer. Desirable difficulties in cross-situational word learning. In *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, 2010.

H. A. Vlach, C. M. Sandhofer, and N. Kornell. The Spacing Effect in Children's Memory and Category Induction. *Cognition*, 109(1):163–167, Oct. 2008.

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