A Computational Model of Memory, Attention, and Word Learning

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

Abstract

There is considerable evidence that people generally learn items better when the presentation of items is distributed over a period of time (the spacing effect). We hypothesize that both forgetting and attention to novelty play a role in the spacing effect in word learning. We build an incremental probabilistic computational model of word learning that incorporates a forgetting and attentional mechanism. Our model accounts for experimental results on children as well as several patterns observed in adults.

1 Memory, Attention, and Word Learning

Learning the meaning of words is an important component of language acquisition, and an extremely challenging task faced by young children (*e.g.*, Carey, 1978; Bloom, 2000). Much psycholinguistic research has investigated the mechanisms underlying early word learning, and the factors that may facilitate or hinder this process (*e.g.*, Quine, 1960; Markman and Wachtel, 1988; Golinkoff et al., 1992; Carpenter et al., 1998). Computational modeling has been critical in this endeavor, by giving precise accounts of the possible processes and influences involved (*e.g.*, Siskind, 1996; Regier, 2005; Yu, 2005; Fazly et al., 2010). However, computational models of word learning have generally not given sufficient attention to the broader interactions of language acquisition with other aspects of cognition and cognitive development.

Memory limitations and attentional mechanisms are of particular interest, with recent computational studies reconfirming their important role in aspects of word learning. For example, Frank et al. (2010) show that memory limitations are key to matching human performance in a model of word segmentation, while Smith et al. (2010) further demonstrate how attention plays a role in word learning by forming the basis for abstracting over the input. But much potential remains for computational modeling to contribute to a better understanding of the role of memory and attention in word learning.

One area where there is much experimental evidence relevant to these interactions is in the investigation of the *spacing effect* in learning (Ebbinghaus, 1885; Glenberg, 1979; Dempster, 1996; Cepeda et al., 2006). 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 such studies have looked at the performance of adults as well as children (Glenberg, 1976; Pavlik and Anderson, 2005; Vlach et al., 2008). While this 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.

Hypothesized explanations for the spacing effect have included both memory limitations and attention. For example, many researchers assume that the process of forgetting is responsible for the improved performance in the spaced presentation: Because participants forget more of what they have learned in the longer interval, they learn more from subsequent presentations (Melton, 1967; Jacoby, 1978;

Cuddy and Jacoby, 1982). However, the precise relation between forgetting and improved learning has not been made clear. It has also been proposed that subjects attend more to items in the spaced presentation because accessing less recent (more novel) items in memory requires more effort or attention (Hintzman, 1974). However, the precise attentional mechanism at work in the spacing experiments is not completely understood.

While such proposals have been discussed for many years, to our knowledge, there is as yet no detailed computational model of the precise manner in which forgetting and attention to novelty play a role in the spacing effect. Moreover, while mathematical models of the effect help to clarify its properties (Pavlik and Anderson, 2005), it is very important to situate these general cognitive mechanisms within a model of word learning in order to understand clearly how these various processes might interact in the natural word learning setting.

We address this gap by considering memory constraints and attentional mechanisms in the context of a computational model of word-meaning acquisition. Specifically, we change an existing probabilistic incremental model of word learning (Fazly et al., 2010) by integrating two new factors: (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. 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 can account for the effect.

2 The Computational Model

We extend the model of Fazly et al. (2010) — henceforth referred to as FAS10 — by integrating new functionality to capture forgetting and attention to novelty. The model of FAS10 is an appropriate starting point for our study because it is an incremental model of word learning that learns probabilistic associations between words and their semantic properties from naturalistic data. Nonetheless, the model assumes equal attention to all words and objects present in the input, and, although incremental, it has a perfect memory for the internal representation of each processed input. Hence, as we will show, it is incapable of simulating the spacing effects observed in humans.

2.1 The FAS10 Model

The input to the model is a sequence of utterances (a set of words), each paired with a scene representation (a set of semantic features, representing what is perceived when the words are heard), as in:

Utterance: { *she*, *drinks*, *milk* }
Scene: { ANIMATE, PERSON, FEMALE, CONSUME, DRINK, SUBSTANCE, FOOD, DAIRY-PRODUCT }

For each word, the model of FAS10 learns a probability distribution over all possible features, $p(.|w)$, called the *meaning probability* of the word. Before processing any input, all features are equally likely for a word, and the word's meaning probability is uniform over all features. At each time step t, an input utterance–scene pair (similar to the above example) is processed. For each word w and semantic feature f in the input pair, an alignment score, $a_t(w|f)$, is calculated that specifies how strongly the w–f pair are associated at time t. The alignment score in FAS10 uses the meaning probabilities of all the words in the utterance, which reflect the knowledge of the model of word meanings up to that point, as in:

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

where $p_{t-1}(f|w)$ is the probability of f being part of the meaning of word w at time $t-1$.

In the FAS10 model, $p_t(.|w)$ is then updated for all the words in the utterance, using the accumulated evidence from all prior and current co-occurrences of w–f pairs. Specifically, an association score is defined between a word and a feature, $\text{assoc}_t(w,f)$, which is a summation of all the alignments for that w and f up to time t.[1] This association score is then normalized using a smoothed version of the follow-

[1] In FAS10, $\text{assoc}_t(w,f) = \text{assoc}_{t-1}(w,f) + a_t(w|f)$.

ing to yield $p_t(f|w)$:

$$p_t(f|w) = \frac{\text{assoc}_t(f, w)}{\sum\limits_{f' \in \mathcal{M}} \text{assoc}_t(f', w)} \tag{2}$$

where $\mathcal{M}$ is the set of all observed features.

There are two observations to make about the FAS10 model in the context of our desire to explore attention and forgetting mechanisms in word learning. First, the calculation of alignments $a_t(w|f)$ in Eqn. (1) treats all words equally, without special attention to any particular item(s) in the input. Second, the $\text{assoc}_t(f, w)$ term in Eqn. (2) encodes perfect memory of all calculated alignments since it is a simple accumulated sum. These properties motivate the changes to the formulation of the model that we describe next.

2.2 Adding Attention to Novelty to the Model

As noted just above, the FAS10 model lacks any mechanism to focus attention on certain words, as is suggested by theories on the spacing effect (Hintzman, 1974). One robust observation in studies on attention is that people attend to new items in a learning scenario more than other items, leading to improved learning of the novel items (*e.g.*, Snyder et al., 2008; MacPherson and Moore, 2010; Horst et al., 2011). We thus model the effect of attention to novelty when calculating alignments in our new model: attention to a more novel word increases the strength of its alignment with a feature — and consequently the learned word–feature association — compared to the alignment of a less novel word.

We modify the original alignment formulation of FAS10 to incorporate a multiplicative novelty term as follows (cf. Eqn. (1)):

$$a_t(w,f) = \frac{p_t(f|w)}{\sum\limits_{w' \in \text{U}_t} p_t(f|w')} * \text{novelty}_t(w) \tag{3}$$

where $\text{novelty}_t(w)$ specifies the degree of novelty of a word as a simple inverse function of recency. That is, we assume that the more recently a word has been observed by the model, the less novel it appears to the model. Given a word w at time t that was last observed at time t_{last_w}, we calculate $\text{novelty}_t(w)$ as:

$$\text{novelty}_t(w) = 1 - \text{recency}(t, t_{last_w}) \tag{4}$$

where $\text{recency}(t, t_{last_w})$ is inversely proportional to the difference between t and t_{last_w}. We set $\text{novelty}(w)$ to be 1 for the first exposure of the word.

2.3 Adding a Forgetting Mechanism to the Model

Given the observation above (see end of Section 2.1) that $\text{assoc}_t(w, f)$ embeds perfect memory in the FAS10 model, we add a forgetting mechanism by reformulating $\text{assoc}_t(w, f)$ to incorporate a decay over time of the component alignments $a_t(w|f)$. In order to take a cognitively plausible approach to calculating this function, we observe that $\text{assoc}_t(w, f)$ in FAS10 serves a similar function to *activation* in the ACT-R model of memory (Anderson and Lebiere, 1998). In ACT-R, activation of an item is the sum of individual memory strengthenings for that item, just as $\text{assoc}_t(w, f)$ is a sum of individual alignment strengths for the pair (w, f). A crucial difference is that memory strengthenings in ACT-R undergo decay. Specifically, activation of an item m after t presentations is calculated as: $act(m)_t = \ln(\sum_{t'=1}^{t} 1/(t-t')^d)$, where t' is the time of each presentation, and d is a constant decay parameter.

We adapt this formulation for $\text{assoc}_t(w, f)$ with the following changes: First, in the *act* formula, the constant 1 in the numerator is the basic strength of each presentation to memory. In our model, this is not a constant but rather the strength of alignment, $a_t(w|f)$. Second, we assume that stronger alignments should be more entrenched in memory and thus decay more slowly than weaker alignments. Thus, each alignment undergoes a decay which is dependent on the strength of the alignment rather than a constant decay d. We thus define $\text{assoc}_t(w, f)$ to be:

$$assoc_t(f, w) = \ln(\sum_{t'=1}^{t} \frac{a_{t'(w|f)}}{(t-t')^{d_{a_{t'}}}}) \tag{5}$$

where the decay for each alignment $d_{a_{t'}}$ is:

$$d_{a_{t'}} = \frac{d}{a_{t'}(w|f)} \tag{6}$$

where d is a constant parameter. Note that the $d_{a_{t'}}$ decreases as $\text{a}_{t'}(w|f)$ increases.

apple: { FOOD:1, SOLID:.72, PRODUCE:.63,
EDIBLE-FRUIT:.32, PLANT-PART:.22,
PHYSICAL-ENTITY:.17, WHOLE:.06, $\cdots$ }

Figure 1: True meaning features & scores for *apple*.

3 Input Generation

The input data consists of a set of utterances paired with their corresponding scene representations. The utterances are taken from the child-directed speech (CDS) portion of the Manchester corpus (Theakston et al., 2001), from CHILDES (MacWhinney, 2000), which includes transcripts of conversations with 12 British children, ages 1;8 to 3;0. Every utterance is considered as a bag of lemmatized words. Half of the data is used as the development set, and the other half in the final experiments.

Because no manually-annotated semantic representation is available for any such large corpus of CDS, we use the approach of Nematzadeh et al. (2012) to generate scene representations. For each utterance a scene representation is generated artificially, by first creating an input-generation lexicon that contains the *true meaning* ($t(w)$) of all the words (w) in our corpus. The true meaning is a vector of semantic features and their assigned scores (Figure 1). The semantic features for a word, depending on its part of speech, are chosen from different sources such as WordNet.[2] The score of each feature is calculated automatically to give a higher value to the more specific features (such as FRUIT for *apple*), rather than more general features (like PHYSICAL-ENTITY for *apple*).

To generate the scene representation S of an utterance U, we probabilistically sample a subset of features from the features in $t(w)$ for each word $w \in$ U. Thus, in each occurrence of w some of its features are missing from the scene, resulting in an imperfect sampling. This imperfect sampling allows us to simulate noise and uncertainty in the input, as well as the uncertainty of a child in determining the relevant meaning elements in a scene. The scene S is the union of all the features sampled for all the words in the utterance. We note that the input-generation lexicon is only used in creating input corpora that are naturalistic (based on child-directed speech), and not in the learning of the model.

4 Experiments

First, we examine the overall word learning behaviour in our new model. Then we look at spacing effects in the learning of novel words. In both these experiments, we compare the behavior of our model with the model of FAS10 to clearly illustrate the effects of forgetting and attention to novelty in the new model. Next we turn to further experiments exploring in more detail the interaction of forgetting and attention to novelty in producing spacing effects.

4.1 Word Learning over Time

Generally, the model of FAS10 has increasing comprehension of words as it is exposed to more input over time. In our model, we expect attention to novelty to facilitate word learning, by focusing more on newly observed words, whereas forgetting is expected to hinder learning. We need to see if the new model is able to learn words effectively when subject to the combined effects of these two influences.

To measure how well a word w is learned in each model, we compare its learned meaning $l(w)$ (a vector holding the values of the meaning probability $p(.|w)$) to its true meaning $t(w)$ (see Section 3):

$$\text{acq}(w) \quad = \quad \text{sim}(l(w), t(w)) \qquad (7)$$

where sim is the cosine similarity between the two meaning vectors, $t(w)$ and $l(w)$. The better the model learns the meaning of w, the closer $l(w)$ would get to $t(w)$, and the higher the value of sim would become. To evaluate the overall behaviour of a model, at each point in time, we average the acq score of all the words that the model has seen.

We train each model on $10,000$ input utterance–scene pairs and compare their patterns of word learning over time (Figure 2).[3] We can see that in the original model, the average acq score is mostly increasing over time before leveling off. Our model, starts at a higher average acq score compared to FAS10's model, since the effect of attention to novelty is stronger than the effect of forgetting in early stages of training. There is a sharp decrease in the acq scores after the early training stage, which then levels off. The early decrease in acq scores occurs because many of the words the model is ex-

[2] http://wordnet.princeton.edu

[3] The constant decay parameter d in Eqn. (6) is set to 0.03 in this experiment.

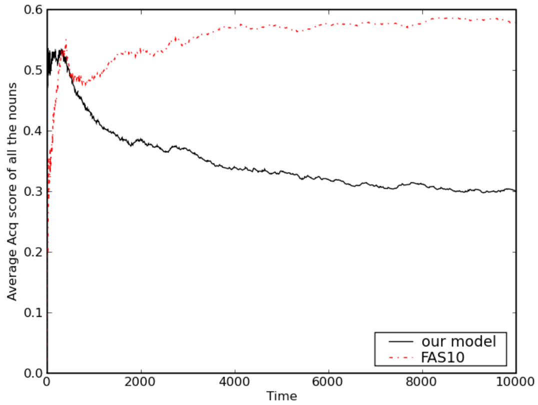

Figure 2: Average acq score of the words over time, for our model and FAS10's model.

posed to early on are not learned very well initially, and so forgetting occurs at a higher rate during that stage. The model subsequently stabilizes, and the acq scores level off although at a lower absolute level than the FAS10 model. Note that when comparing these two models, we are interested in the pattern of learning; in particular, we need to ensure that our new word learning model will eventually stabilize as expected. Our model stabilizes at a lower average acq score since unlike FAS10's model, it does not implement a perfect memory.

4.2 The Spacing Effect in Novel Word Learning

Vlach et al. (2008) performed an experiment to investigate the effect of presentation spacing in learning novel word–object pairs in three-year-old children. Each pair was presented 3 times in each of two settings, either consecutively (massed presentation), or with a short play interval between each presentation (spaced presentation). Children were then asked to identify the correct object corresponding to the novel word. The number of correct responses was significantly higher when the pairs were in the spaced presentation compared to the massed presentation. This result clearly demonstrates the spacing effect in novel word learning in children.

Experiments on the spacing effect in adults have typically examined and compared different amounts of time between the spaced presentations, which we refer to as the spacing interval. Another important parameter in such studies is the time period between the last training trial and the test trial(s), which we refer to as the retention interval (Glenberg, 1976; Bahrick and Phelps, 1987; Pavlik and Anderson, 2005). Since the experiment of Vlach et al. (2008) was designed for very young children, the procedures were kept simple and did not vary these two parameters. We design an experiment similar to that of Vlach et al. (2008) to examine the effect of spacing in our model, but extend it to also study the role of various spacing and retention intervals, for comparison to earlier adult studies.

4.2.1 Experimental Setup

First, the model is trained on 100 utterance–scene pairs to simulate the operation of normal word learning prior to the experiment.[4] Then a randomly picked novel word (*nw*) that did not appear in the training trials is introduced to the model in 3 teaching trials, similar to Vlach et al.'s (2008) experiment. For each teaching trial, *nw* is added to a different utterance, and its probabilistically-generated meaning representation (see Section 3) is added to the corresponding scene. We add *nw* to an utterance–scene pair from our corpus to simulate the presentation of the novel word during the natural interaction with the child in the experimental setting.

The spacing interval between each of these 3 teaching trials is varied from 0 to 29 utterances, resulting in 30 different simulations for each *nw*. For example, when the spacing interval is 5, there are 5 utterances between each presentation of *nw*. A spacing of 0 utterances yields the massed presentation. We run the experiment for 20 randomly-chosen novel words to ensure that the pattern of the results is not related to the meaning representation of a specific word.

For each spacing interval, we look at the acq score of the novel word at two points in time, to simulate two retention intervals: One immediately after the last presentation of the novel word (*imm* condition) and one at a later point in time (*lat* condition). By looking at these two conditions, we can further observe the effect of forgetting in our model, since the decay in the model's memory would be more severe in the *lat* condition, compared to the *imm* condition.[5] The results reported here for each spacing

[4] In the experiments of Section 4.2.2 and Section 4.3, the constant decay parameter d is equal to 0.04.

[5] Recall that each point of time in our model corresponds to

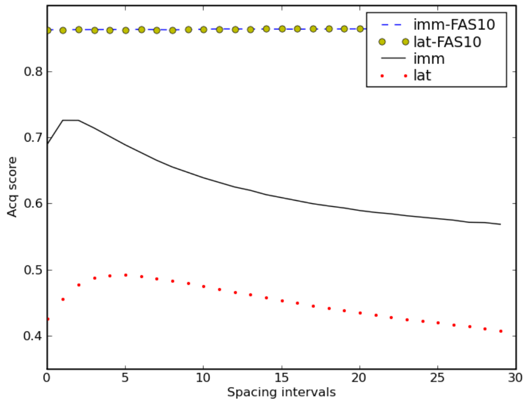

Figure 3: Average acq score of novel words over spacing intervals, in our model and FAS10's model.

interval average the acq scores of all the novel words at the corresponding points in time.

4.2.2 The Basic Spacing Effect Results

Figure 3 shows the results of the simulations in our model and the FAS10 model. We assume that very small spacing intervals (but greater than 0) correspond to the spaced presentation in the Vlach et al. (2008) experiments, while a spacing of 0 corresponds to the massed presentation. In the FAS10 model, the average acq score of words does not change with spacing, and there is no difference between the *imm* and *lat* conditions, confirming that this model fails to mimic the observed spacing effects. By contrast, in our model the average acq score is greater in the small spacing intervals (1-3) than in the massed presentation, mimicking the Vlach et al. (2008) results on children. This happens because a *nw* appears more novel with larger spacing intervals between each of its presentations resulting in stronger alignments.

We can see two other interesting patterns in our model: First, the average acq score of words for all spacing intervals is greater in the *imm* condition than in the *lat* condition. This occurs because there is more forgetting in the model over the longer retention interval of *lat*. Second, in both conditions the average acq score initially increases from a massed presentation to the smaller spacing intervals. However, at spacing intervals between about 3 and 5, the acq score begins to decrease as spacing intervals grow larger. As explained earlier, the initial increase in acq scores for small spacing intervals results from novelty of the words in a spaced presentation. However, for bigger spacing intervals the effect of novelty is swamped by the much greater degree of forgetting after a bigger spacing interval.

Although Vlach et al. (2008) did not vary their spacing and retention intervals, other spacing effect studies on adults have done so. For example, Glenberg (1976) presented adults with word pairs to learn under varying spacing intervals, and tested them after several different retention intervals (his experiment 1). Our pattern of results in Figure 3 is in line with his results. In particular, he found a nonmonotonic pattern of spacing similar to the pattern in our model: learning of pairs was improved with increasing spacing intervals up to a point, but there was a decrease in performance for larger spacing intervals. Also, the proportion of recalled pairs decreased for longer retention intervals, similar to our lower performance in the *lat* condition.

4.3 The Role of Forgetting and Attention

To fully understand the role as well as the necessity of, both forgetting and attention to novelty in our results, we test two other models under the same conditions as the previous spacing experiment: (a) a model with our mechanism for attention to novelty but not forgetting, and (b) a model with our forgetting mechanism but no attention to novelty; see Figure 4 and Figure 5, respectively.

In the model that attends to novelty but does not incorporate a memory decay mechanism (Figure 4), the average acq score consistently increases as spacing intervals grow bigger. This occurs because the novel words appear more novel following bigger spacing intervals, and thus attract more alignment strength. Since the model does not forget, there is no difference between the immediate (*imm*) and later (*lat*) retention intervals. This pattern does not match the spacing effect patterns of people, suggesting that forgetting is a necessary aspect of our model's ability to do so in the previous section.

On the other hand, in the model with forgetting but no attentional mechanism (Figure 5), we see two different behaviors in the *imm* and *lat* conditions. In the *imm* condition, the average acq score decreases

processing an input pair. The acq score in the *imm* condition is calculated at time t, which is immediately after the last presentation of *nw*. The *lat* condition corresponds to $t+20$.

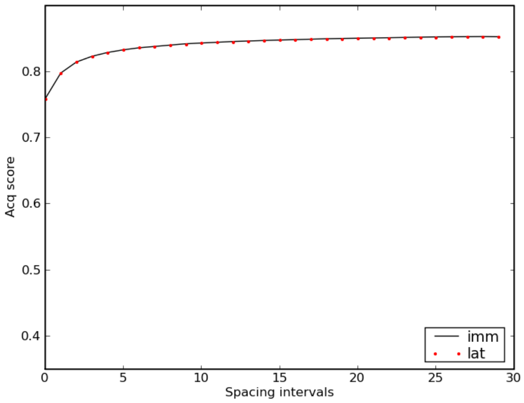

Figure 4: Average acq score of the novel words over spacing intervals, for the model with novelty but without forgetting.

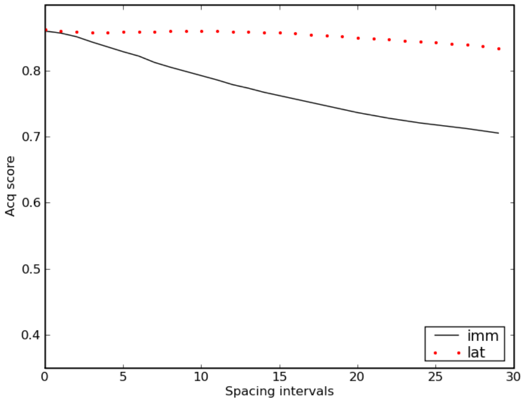

Figure 5: Average acq score of the novel words over spacing intervals, for the model with forgetting but without novelty.

consistently over spacing intervals. This is as expected, because the greater time between presentations means a greater degree of forgetting. Specifically, the alignment scores decay more between presentations of the word to be learned, given the greater passage of time in larger spacing intervals. The weaker alignments then lead to lower acq scores in this condition.

Paradoxically, although this effect on learning also holds in the *lat* condition, another factor is at play, leading to better performance than in the *imm* condition at all spacing intervals. Here the greater retention interval — the time between the last learning presentation and the test time — leads to greater forgetting in a manner that instead improves the acq scores. Consider that the meaning representation for a word includes some probability mass assigned to irrelevant features — i.e., those features that occurred in an utterance–scene pair with the word but are not part of its true meaning. Because such features generally have lower probability than relevant features (which are observed more consistently with a word), a longer retention interval leads to them decaying more than the relevant features. Thus the *lat* condition enables the model to better focus on the features relevant to a word.

In conclusion, neither attention to novelty nor forgetting alone achieves the pattern typical of the spacing effects in people that our model shows in the lower two plots in Figure 3. Hence we conclude that both factors are necessary to our account, suggesting that it is an interaction between the two that accounts for people's behaviour.

4.4 The "Spacing Crossover Interaction"

In our model with attention to novelty and forgetting (see Section 4.2), the average acq score was always better in the *imm* condition than the *lat* condition. However, researchers have observed other patterns in spacing experiments. A particularly interesting pattern found in some studies is that the plots of the results for earlier and later retention intervals cross as the spacing intervals are increased. That is, with smaller spacing intervals, a shorter retention interval (such as our *imm* condition) leads to better results, but with larger spacing intervals, a longer retention interval (such as our *lat* condition) leads to better results (Bahrick, 1979; Pavlik and Anderson, 2005). This interaction of spacing and retention intervals results in a pattern referred to as the spacing crossover interaction (Pavlik and Anderson, 2005). This pattern is different from Glenberg's (1976) experiment and from the pattern of results shown earlier for our model (Figure 3).

We looked at an experiment in which the spacing crossover pattern was observed: Pavlik and Anderson (2005) taught Japanese–English pairs to subjects, varying the spacing and retention intervals. One difference we noticed between the experiment of Pavlik and Anderson (2005) and Glenberg (1976) was that in the former, the presentation period of the stimulus was 5 seconds, whereas in the latter, it was 3 seconds. We hypothesize that the difference between the amount of time for the presentation peri-

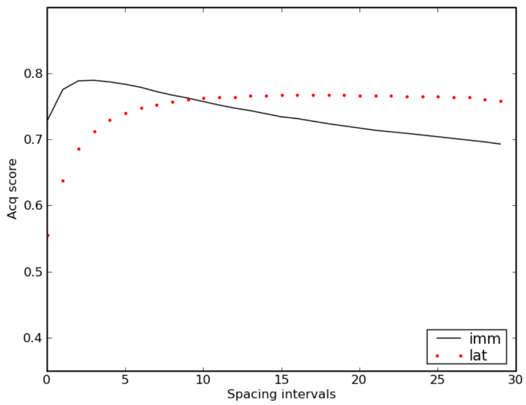

Figure 6: Average acq score of the novel words over spacing intervals

ods might explain the different spacing patterns in these experiments.

We currently cannot model presentation time directly in our model, since having access to an input longer does not change its computation of alignments between words and features. However, we can indirectly model a difference in presentation time by modifying the amount of memory decay: We assume that when an item is presented longer, it is learned better and therefore subject to less forgetting. We run the spacing experiment with a smaller forgetting parameter to model the longer presentation period used in Pavlik and Anderson's (2005) versus Glenberg (1976).[6]

Our results using the decreased level of forgetting, given in Figure 6, show the expected crossover interaction between the retention and spacing intervals: for smaller spacing intervals, the acq scores are better in the *imm* condition, whereas for larger spacing intervals, they are better in the *lat* condition. Thus, 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.

5 General Discussion and Future Work

The spacing effect (where people learn items better when multiple presentations are spread over time) has been studied extensively and is found to be robust over different types of tasks and domains. Many experiments have examined the spacing effect in the context of word learning and other similar tasks. Particularly, in a recent study of Vlach et al. (2008), young children demonstrated a spacing effect in a novel word learning task.

We use computational modeling to show that by changing a probabilistic associative model of word learning to include both a forgetting and attentional mechanism, the new model can account not only for the child data, but for various patterns of spacing effect data in adults. Specifically, our model shows the nonmonotonic pattern of spacing observed in the experimental data, where learning improves in shorter spacing intervals, but worsens in bigger spacing intervals. Our model can also replicate the observed cross-over interaction between spacing and retention intervals: for smaller spacing intervals, performance is better when tested after a shorter retention interval, whereas for bigger spacing intervals, it is better after longer retention intervals. Finally, our results confirm that by modelling word learning as a standalone development process, we cannot account for the spacing effect. Instead, it is important to consider word learning in the context of fundamental cognitive processes of memory and attention.

Much remains to be investigated in our model. For example, most human experiments examine the effect of frequency of presentations of target items. Also, the range of retention intervals that has been studied is greater than what we have considered here. In the future, we plan to study the effect of these two parameters. In addition, with our current model, the amount of time an item is presented to the learner does not play a role. We can also reformulate our alignment mechanism to incorporate a notion of the amount of time to consider an item to be learned. Another interesting future direction, especially in the context of word learning, is to develop a more complete attentional mechanism, that considers different parameters such as social cues and linguistic cues. Finally, we will study the role of forgetting and attention in modelling other relevant experimental data (*e.g.*, Kachergis et al., 2009; Vlach and Sandhofer, 2010).

[6] Here, the decay parameter is set to 0.03.

References

John .R. Anderson and Christian Lebiere. 1998. *The atomic components of thought*. Lawrence Erlbaum Associates.

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

Harry P. Bahrick and Elizabeth Phelps. 1987. Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(2):344–349.

Paul Bloom. 2000. *How Children Learn the Meanings of Words*. MIT Press.

Susan Carey. 1978. The child as word learner. In M. Halle, J. Bresnan, and G. A. Miller, editors, *Linguistic Theory and Psychological Reality*. The MIT Press.

Malinda Carpenter, Katherine Nagell, Michael Tomasello, George Butterworth, and Chris Moore. 1998. Social cognition, joint attention, and communicative competence from 9 to 15 months of age. *Monographs of the Society for Research in Child Development, 63(4)*.

Nicholas J. Cepeda, Harold Pashler, Edward Vul, John T. Wixted, and Doug Rohrer. 2006. Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, 132(3):354 – 380.

Lauren J. Cuddy and Larry L. Jacoby. 1982. When forgetting helps memory: an analysis of repetition effects. *Journal of Verbal Learning and Verbal Behavior*, 21(4):451 – 467.

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

Frank N. Dempster. 1996. Distributing and managing the conditions of encoding and practice. *Memory*, pages 317–344.

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

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

Michael C. Frank, Sharon Goldwater, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2010. Modeling human performance in statistical word segmenation. *Cognition*, 117:107–125.

Arthur Glenberg. 1979. Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory and Cognition*, 7:95–112.

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

Roberta M. Golinkoff, Kathy Hirsh-Pasek, Leslie M. Bailey, and Neil R. Wegner. 1992. Young children and adults use lexical principles to learn new nouns. *Developmental Psychology*, 28(1):99–108.

Douglas L. Hintzman. 1974. Theoretical implications of the spacing effect.

Jessica S. Horst, Larissa K. Samuelson, Sarah C. Kucker, and Bob McMurray. 2011. Whats new? children prefer novelty in referent selection. *Cognition*, 118(2):234 – 244.

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

George Kachergis, Chen Yu, and Richard Shiffrin. 2009. Temporal contiguity in cross-situational statistical learning.

Amy C. MacPherson and Chris Moore. 2010. Understanding interest in the second year of life. *Infancy*, 15(3):324–335.

Brian MacWhinney. 2000. *The CHILDES Project: Tools for Analyzing Talk*, volume 2: The Database. Erlbaum, 3rd edition.

Ellen M. Markman and Gwyn F. Wachtel. 1988. Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, 20:121–157.

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

Aida Nematzadeh, Afsaneh Fazly, and Suzanne Stevenson. 2012. Interaction of word learning and semantic category formation in late talking. In *Proc. of CogSci'12*. To appear.

Philip I. Pavlik and John R. Anderson. 2005. Practice and forgetting effects on vocabulary memory: An activationbased model of the spacing effect. *Cognitive Science*, 29:559–586.

W.V.O. Quine. 1960. *Word and Object*. MIT Press.

Terry Regier. 2005. The emergence of words: Attentional learning in form and meaning. *Cognitive Science*, 29:819–865.

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

Linda B Smith, Eliana Colunga, and Hanako Yoshida. 2010. Knowledge as process: Contextually-cued attention and early word learning. *Cogn Sci*, 34(7):1287–314.

Kelly A. Snyder, Michael P. BlanK, and chad J. Marsolek. 2008. What form of memory underlies novelty preferences? *Psychological Bulletin and Review*, 15(2):315 – 321.

Anna L. Theakston, Elena V. Lieven, Julian M. Pine, and Caroline F. Rowland. 2001. The role of performance limitations in the acquisition of verb–argument structure: An alternative account. *J. of Child Language*, 28:127–152.

Haley A Vlach and Catherine M Sandhofer. 2010. Desirable difficulties in cross-situational word learning.

Haley A. Vlach, Catherine M. Sandhofer, and Nate Kornell. 2008. The Spacing Effect in Children's Memory and Category Induction. *Cognition*, 109(1):163–167, October.

Chen Yu. 2005. The emergence of links between lexical acquisition and object categorization: A computational study. *Connection Science*, 17(3–4):381–397.