Desirable Difficulty in Learning: A Computational Investigation

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

Abstract

Certain difficulties of a word learning situation can promote long-term learning, and thus are referred to as "desirable difficulties". We use a computational modelling approach to examine the possible explanatory factors of the observed patterns in a cross-situational word learning experiment. Our results suggest that the within-trial ambiguity and the presentation duration of each trial in addition to other distributional characteristics of the input (experimental stimuli) may explain these results. Our findings also emphasize the role of computational modelling in understanding empirical results.

Introduction

One of the important questions in language acquisition is how people learn the mappings between words and their meanings (Quine, 1960). A number of mechanisms and approaches have been proposed in an attempt to address this question (*e.g.*, Tomasello, 1992; Pinker, 1989). A widely-discussed mechanism is *cross-situational learning*, in which people learn word meanings by gathering evidence from various exposures of words in different situations. Recent word learning experiments also confirm that both adults and children keep track of cross–situational statistics across individually ambiguous learning trials, and infer the correct word–meaning mappings even in highly ambiguous conditions (Yu & Smith, 2007; Smith & Yu, 2008). These experiments have gained popularity in recent years (*e.g.*, Yurovsky & Yu, 2008; Vlach, Sandhofer, & Kornell, 2008), and provide opportunities for further investigating the observed patterns in naturalistic word learning.

One interesting aspect of word learning that can be studied in such experiments, is its interaction with other cognitive processes such as memory and attention. An example is the experiments of Vlach et al. (2008) on children, which examine the *spacing effect*, *i.e.*, the observation that people generally learn better when the presentations of the items to be learned are distributed (*spaced*) over a period of time. This and other similar patterns in human learning are referred to as "desirable difficulties": Although a more difficult learning situation may hinder short-term recall of learned material, it may promote long-term retention.

In this work, we use a computational model to shed light on one such case of an observed "desirable difficulty" in cross-situational word learning, studied by Vlach and Sandhofer (2010). Notably, Vlach and Sandhofer (2010) attribute their findings to desirable difficulties in learning, but do not provide an explanation of why and how the sort of difficulty they focus on facilitates long-term retention of the learned words. Computational modelling enables us to investigate the precise learning mechanisms, and the variations in the input conditions, that might explain these findings. We first introduce our computational model of cross-situational word learning, and then explain and analyze the experimental data and results of Vlach and Sandhofer (2010) in the context of our model. Finally, we describe the way we simulate these experiments using our model, and how this enables us to examine the role of several different factors in the observed pattern of word learning.

The Computational Model

In this section, we present our computational model of word learning that was first published in Nematzadeh, Fazly, and Stevenson (2012a). Our model builds on the word learning model of Fazly, Alishahi, and Stevenson (2010), which takes an incremental approach in learning probabilistic associations between words and their meanings. In Nematzadeh et al., we integrated new functionality into this model to capture forgetting (*i.e.*, an effect of memory) and attention to novelty. Our proposed model accounts for several observed patterns of the spacing effect in children and adults, in which experimental subjects learn presented items better when they are spaced apart in time, than when they are shown in immediate succession. We provide a brief overview of the model before turning to modelling of other kinds of "desirable difficulties."

Learning from an Input Pair

Our model learns about the meaning of words by incrementally processing a corpus that contains a sequence of utterances paired with a semantic representation of a scene, which is the hypothetical perception of a learner upon hearing the utterance. Each input to the model pairs a set of words (the representation of the utterance) with a set of semantic features (the representation of the scene), as in:

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

We create corpora drawn from child-directed speech, in which lemmatized, transcribed utterances are paired with artificially generated semantics, based on WordNet or other semantic featural representations of the entities and actions corresponding to the words. In the experiments here on novel word learning, nonce words are paired with these naturalistic semantic representations, in which features corresponding to meaning properties are probabilistically associated with a word.

When processing an input pair, the model bootstraps its current knowledge of word meanings to hypothesize the strength of association between the words in the current input and the meaning features in the current scene. These probabilistic alignments between the words and features of the current input are then used to update the model's knowledge of word meanings.

More formally, for each word, the model learns a *meaning probability*, which is a probability distribution over all possible semantic features. The model starts with uniform meaning probabilities for all words; *i.e.*, before processing any input, all features are equally likely for every word. At each time step t, the model processes an input pair and calculates an *alignment score*, $a_t(w,f)$, between each word w and semantic feature f in the input pair. This alignment score reflects how strongly the w–f pair are associated at time t, by considering two sources of information: (1) the meaning probabilities of all the words in the utterance U_t (representing the knowledge of the model of word meanings up to that point), and (2) the *novelty* of words, capturing the attention a learner might pay to the novel words compared to the familiar words (explained below). The alignment score is formulated as:

$$a_t(w,f) = \frac{p_t(f|w)}{\sum\limits_{w' \in \mathrm{U}_t} p_t(f|w')} * \mathrm{novelty}_t(w) \tag{1}$$

where $p_t(f|w)$ is the probability of f being part of the meaning of word w at time t, right before processing the input pair, and $\mathrm{novelty}_t(w)$ is a multiplicative attentional factor.

This factor, $\mathrm{novelty}_t(w)$, taps into empirical studies on attention showing that people attend to novel items in a learning scenario more than other items, leading to improved learning of those items (*e.g.*, Snyder, Blank, & Marsolek, 2008; MacPherson & Moore, 2010; Horst, Samuelson, Kucker, & McMurray, 2011). In the word learning scenario, this corresponds to a person focusing on determining the meaning of novel words. We model this observation by incorporating the multiplicative $\mathrm{novelty}_t(w)$ in the above formula, providing an increase in word–feature association for a more novel word. The $\mathrm{novelty}_t(w)$ measures the degree of novelty of a word as a simple inverse function of recency: The more recently a word w has been observed by the model (t_{last_w}), the less novel it appears to the model at the current time t:

$$\mathrm{novelty}_t(w) = 1 - \mathrm{recency}(t, t_{last_w}) \tag{2}$$

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

Accumulating Evidence over Time

The model keeps track of the accumulation of all the alignment scores of all word–feature pairs, and uses these scores to update the meaning probabilities of the words. These alignment scores reflect the model's knowledge of the associations between words and various potential meanings. To simulate the effect of forgetting in memory, these alignments undergo a decay over time. At each time t, the strength of association of a word and a feature is formulated as:

$$\mathrm{assoc}_t(f,w) = \sum_{t'} \frac{a_{t'(w,f)}}{(t-t')^{d_{a_{t'}}}} \tag{3}$$

where t' is the time at which the alignment $a_{t'}$ is calculated, and $d_{a_{t'}}$ is the decay rate associated with this alignment. We note that our formulation of assoc is inspired by the ACT-R model of memory (Anderson & Lebiere, 1998), in which the sum of individual memory strengthenings for an item determines the item's *activation*. 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:

$$d_{a_{t'}} = \frac{d}{a_{t'}(w,f)} \tag{4}$$

where d is a constant parameter. Note that the alignments between a word and different features may be forgotten at different rates.

This association score is then normalized using a smoothed version of the following to yield the meaning probability of that feature f for that word w at time t:

$$p_t(f|w) = \frac{\mathrm{assoc}_t(f,\, w)}{\sum\limits_{f' \in \mathcal{M}} \mathrm{assoc}_t(f',\, w)} \tag{5}$$

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

Desirable Difficulties in Word Learning

Vlach and Sandhofer (2010) — henceforth V&S — explore the factors involved in "desirable difficulty" through a set of (now standard) cross-situational word learning experiments on adults, varying the presentation and testing conditions. In each $N \times N$ trial, subjects see some number N of novel objects on a computer screen, while hearing N novel words (in arbitrary order) that label the displayed objects; see Figure 1. In testing, subjects hear a single word, and are asked to select the corresponding object from a display of 4 objects. Across three presentation conditions, the total number of word–object pairs, and the number of times each is seen, are held constant, while there is increasing within-trial ambiguity — *i.e.*, the number of possible pairings between the words and the objects within a single presentation: 2×2, 3×3, and 4×4. Furthermore, participants were tested at each of three times: immediately after training, 30 minutes after, and one week after.

V&S find that in the immediate testing condition, as expected, the number of correctly learned pairs decreases as the within-trial ambiguity increases. That is, the participants performed the best in the 2×2 condition and the worst in 4×4 (Figure 2). However, when tested after 30 minutes of delay, there was no significant difference between the performance

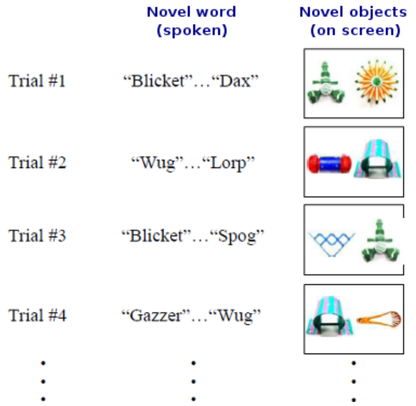

Figure 1: Example stimuli from 2 × 2 condition taken from V&S.

of the participants in the 2 × 2 and the 3 × 3 conditions, while 4 × 4 still had the worst performance. Interestingly, in testing after one week, the participants performed better in the 3 × 3 than the 2 × 2 condition. (Again, 4 × 4 still had the worst performance.) In summary, what should be the "easiest" condition (2 × 2) has the best performance in immediate testing, but a more difficult condition (3 × 3) has better performance one week later.

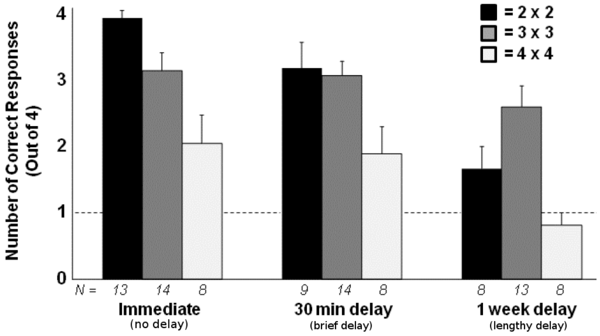

Figure 2: The results of V&S's experiment.

V&S relate their findings to "desirable difficulties" in learning: they argue that the difficulty of a learning situation might hinder immediate performance, but promote longer term performance. However, they do not discuss why the performance of the 4 × 4 condition is the worst compared to the other conditions for all testing intervals. That is, why is the level of difficulty in 3 × 3 desired, but is not so for 4 × 4. Moreover, they do not explain why and how difficulty can boost learning in the long term in this learning scenario.

We observe that, in the V&S experiments, the 2 × 2 condition has more learning trials, each of which is seen for less time, than in the 3 × 3 condition (and similarly for 3 × 3 compared to 4 × 4). This occurs because the total number of word–object pairs, the number of times each is seen, and the total presentation time of the full set of items, are all held constant across the three presentation conditions. We can thus identify three factors that differ across the V&S conditions, each of which may contribute to the observed pattern: (1) the within-trial ambiguity, (2) the presentation duration of each trial, and (3) the average spacing interval (where spacing is the number of trials between the two presentations of a word–object pair).

Computational modelling can be used as a tool to study the necessity and the interaction of these three factors (the within-trial ambiguity, the presentation time of each trial, and the average spacing interval) in a cross-situational learning scenario. In our model, the increase in within-trial ambiguity results in more competition among the possible alignments since there are more words and meanings to potentially align; this results in lower association scores and therefore decreased performance in word learning. We argue that the second factor, the presentation duration, is related to forgetting. In the following section (Methodology), we will explain how we incorporate differences in the presentation duration into our model. The third factor (the spacing interval) relates to the interaction of forgetting and attention to novelty in the model: As the spacing interval becomes larger, the amount of forgetting increases, resulting in lower association scores between words and features; however, the novelty of words and consequently their association scores increases as the spacing interval gets larger. Thus, varying the spacing interval affects the performance of the model (see Nematzadeh et al., 2012a for more details). We use our model to study the interaction of these three factors, with the goal of providing a more precise explanation for the desirable difficulty observed in the experiments of V&S. Next, we explain our methodology, including our input generation, and the simulation of the V&S experiments.

Methodology

Input Generation

To generate the input stimuli for our model, we need to pair words with a meaning representation that corresponds to the depiction of the corresponding object in the experimental situation of Figure 1. To do so, we draw on the input-generation lexicon of Nematzadeh, Fazly, and Stevenson (2012b), which was previously used to automatically annotate corpora of child-directed utterances with meaning features corresponding to the words in those utterances. Here, we use the lexicon to provide a source of naturalistic meaning representations ("novel object descriptions") for a set of "novel" words (*i.e.*, the words in the input stimuli are unknown to the model, as in the experiments we are modeling).

The *true meaning* of each word in the lexicon, $tm(w)$, is a vector of semantic features and their assigned scores or weights (Figure 3).[1] When a word is used in an input trial, its meaning features are probabilistically sampled from $tm(w)$ according to the weight of each feature in the lexical entry of the word. This probabilistic sampling captures our intuition that a participant, when faced with a trial in the cross-situational experiment of Figure 1, will grasp some features of the novel objects but not necessarily all. Each trial of the input is then composed of a set of N words (2, 3, or 4 words,

[1] We note that this lexicon is only used in input generation and evaluation, and not in the learning of the model.

depending on the condition), paired with a set of features which is the union of the N sets of meaning features sampled for each of the words in that trial.

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

Figure 3: True meaning features & probabilities for *apple*.

To produce a full set of experimental trials, we first convert the exact stimuli of V&S to the format of our input. That is, in their stimuli, we replace each word with a specific word from our lexicon, and each object with the probabilistically-generated meaning representation for its corresponding word (as explained above). The precise combination of corresponding word/object pairs in each trial, and the order of the trials, are exactly the same as in the V&S stimuli. We refer to this data as the input of V&S.

The V&S input includes 18 novel word–object pairs, each of which occurs 6 times, resulting in 54, 36, and 27 trials in the 2×2, 3×3, and 4×4 conditions, respectively. We note that the V&S input, as a specific set of stimuli, might have particular spacing properties that contribute to their results. Thus we also randomly generate input stimuli in order to evaluate the effect of arbitrary variation in the precise presentation order of the word/object pairs. We randomly generate 20 sets of input stimuli for each condition, keeping the number of pairs, their frequency, and the number of trials the same as in the V&S input. We use the same novel words that we used in generating V&S data, and randomly generate their meaning representations as explained. The result is that we can experiment both with the precise data of V&S, as well as 20 randomly generated sets of input stimuli with the same basic properties.

Modeling of the Presentation Duration

One aspect of the V&S experimental conditions that we cannot directly replicate in our model is the presentation duration of each trial in a stimulus set. Recall that because of the various properties of the stimuli, the individual trials in each of the three conditions (2×2, 3×3, and 4×4) have different presentation durations. Our model does not have a notion of "presentation duration" — it simply processes each input as it receives it. Thus to simulate these differences, different degrees of forgetting decays are used in the model (see Eqn. (4)). The intuition is that subjects forget faster in a condition with a shorter presentation duration, since they have less time to absorb the stimuli in each trial. The forgetting decay is thus set to a larger value in the 2×2 condition (where the presentation time is the smallest), and successively smaller in each of the 3×3 and 4×4 conditions.

Simulation of the V&S Experiments

We train our model by presenting the set of inputs for a given condition, where it learns incrementally in response to each trial. Similarly to V&S, we evaluate our model at three points of time after training: immediately after processing the last input (time = t), at $t + 30$, and at $t + 350$. These times were chosen to loosely reflect the three time intervals in V&S's experiments. We will use the labels "no delay", "brief delay", and "lengthy delay", to refer to these timings in describing our results.

To evaluate the performance of the model at each testing point, we measure how well each word is acquired by comparing its learned meaning $lm(w)$ – a vector holding the values of the meaning probability (Eqn. (5)) – to its true meaning $tm(w)$ from the input-generation lexicon (see Figure 3):

$$\text{acq}(w) \quad = \quad \text{sim}(lm(w), tm(w)) \qquad (6)$$

where sim is the cosine similarity between the two meaning vectors, $tm(w)$ and $lm(w)$. The higher acq(w) is, the more similar $lm(w)$ and $tm(w)$ are. We use the average acq score at time t of all the words in the input to reflect the overall learning of the model at that time.

Results

We first examine the behavior of our model when trained on the V&S input, and then compare these with results on our randomly generated stimuli.

The Input of V&S

The results of training and evaluating our model on the V&S input are presented in Figure 4. We see the same interesting pattern as found in V&S (shown in Figure 2) for the 2×2 and the 3×3 conditions. That is, 2×2 is better with no delay, but similar with brief delay and worse with lengthy delay, even though 3×3 is "harder" due to its higher degree of within-trial ambiguity. Unlike the V&S results, 3×3 and 4×4 are similar for all delays.

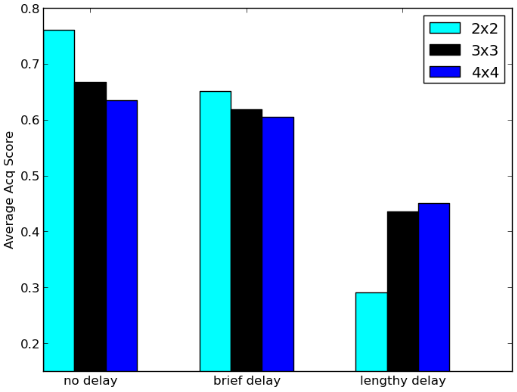

Figure 4: Average acq score of words (from the model) given the three conditions and three time intervals similar to the V&S experiments.

We consider these findings in the context of the discussed factors of presentation duration, within-trial ambiguity, and average spacing of items, which we proposed might explain the desirable difficulty in learning. The differences in presentation duration (shortest for 2×2 and longest for 4×4) entails that, generally, the learning in the 2×2 condition should decline most steeply over time, and learning in the 4×4 should decline least steeply: *i.e.*, for each set of same-coloured bars in Figure 4, we expect learning to decrease over time, and more rapidly for lower values of N in the $N \times N$ conditions. We see this predicted behaviour with our model, which results from our modeling of presentation duration with an inversely proportional decay rate (*i.e.*, the shorter the presentation duration, the greater the degree of forgetting).

It is expected that in the absence of other factors, increasing within-trial ambiguity from the 2×2 to the 4×4 conditions results in a decline in average acq score, since greater ambiguity should lead to decreased learning. However, in our model, the presentation duration also plays a role. Similar to results of V&S, we see the decline pattern in the "no delay" condition, and in the "brief delay" condition (albeit with less difference), due to the increased competition for word–meaning alignments that occurs with a higher number of items in a trial (see Figure 4). However, we do not see this pattern in the lengthy delay condition.

To summarize, our results are similar to those of V&S, who found that while the 2×2 condition led to best learning when tested immediately, it led to poorer performance than the 3×3 condition given a lengthy delay before testing — a pattern V&S attribute to the "desirable difficulty". It seems that these factors of presentation duration and within-trial ambiguity may interact, such that the steep decline in performance in subsequent testing in the 2×2 condition more than offsets the advantage it has from the lesser within-trial ambiguity.

In the experiments of V&S, the performance in the 4×4 condition is always worse than the two other conditions. However, our model produces very similar results for the 3×3 and the 4×4 conditions. Also, the role of the spacing interval is not clear in these results. The problem is that by just considering one set of stimuli within each $N \times N$ condition (each of which has a set spacing of items), we do not have a variation of the average spacing interval that is independent of the presentation duration and the within-trial ambiguity. We turn to these issues in the next subsection.

Randomly Generated Input

We observed that the performance of the model in the 3×3 and 4×4 conditions on the V&S input is very similar. We also investigate a condition here with higher within-trial ambiguity to see if such a condition might be "hard" enough for the model (because of the higher within-trial ambiguity) so that it results in a similar patten to the 4×4 condition in V&S. As with the others, we generate 20 sets of input stimuli for this 6×6 condition, using 18 word-object pairs, each of which occurs 6 times, producing 18 trials. Thus the generated input stimuli for the four conditions allows us to examine both the role of average spacing interval, and the impact of a more difficult condition with higher within-trial ambiguity.

We train our model on the randomly-generated inputs (with different average spacing intervals) for all four $N \times N$ conditions. To evaluate the performance of the model, the average *acq* score of words for all 20 sets of inputs within a single $N \times N$ condition are averaged (see Figure 5). We can see that when tested with "no delay", the 2×2, 3×3, and 4×4 conditions have similar scores. Moreover, we can see a pattern similar to V&S's experiments: the 3×3 and 4×4 conditions have the best results after the "lengthy delay". We also observe that by increasing difficulty in the 6×6 condition (due to the high within-trial ambiguity), the model produces a pattern similar to the pattern observed in the 4×4 condition in V&S's experiments. This confirms our hypothesis that for our model, the 4×4 condition is not "hard" enough to result in a steep decline over time intervals as in the V&S's results.

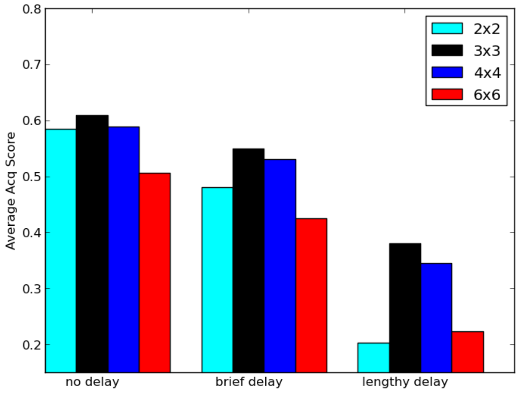

Figure 5: Average acq score of words (from the model) given the four conditions and the three time intervals, averaged over 20 sets of stimuli.

However, we also see that, in contrast to V&S's results (and our model's performance on the V&S data), the 2×2 condition with no delay fails to show better learning than the other conditions.

To better understand this difference between the two sets of results, we look more closely at the scores of the individual randomly-generated stimuli sets. We find that there is a notable difference in the average *acq* score across the 20 input files for the 2×2 condition, which shows its maximum value of 0.76 for the V&S's data, while the minimum is 0.50. This suggests that the characteristics of the particular input (as a result of varying the average spacing interval) may be responsible for some of the observed patterns in the V&S's results.

We were interested to understand why the V&S data has the maximum score, especially since there was a sizable gap between the score of this input and the next best score among the randomly-generated inputs (of 0.64). In an attempt to identify the factor behind this variation, we measured various statistics for each input set, such as the following: (1) the average spacing interval of words, which has been shown to affect learning both in people (Vlach et al., 2008) and in our model (Nematzadeh et al., 2012a); (2) the average time since the last occurrence of words, that impacts the amount of forgetting that occurs; and (3) the average context familiarity of words (that is, the familiarity of the words that occur with a word in an utterance), a factor that has been noted to affect word learning (see, *e.g.*, Fazly, Ahmadi-Fakhr, Alishahi, & Stevenson, 2010). However, we found that none of these measures explain the variation of the scores in all the inputs. Future research is needed to fully understand the impact of the properties these measures tap into, and whether they may (individually or in combination) contribute to explaining the pattern of the results.

Summary

The "desirable difficulty" of a learning condition can promote the long term retention of the learned items. We have used a computational model to investigate the possible factors behind one such case of a "desirable difficulty" in a cross-situational word learning experiment (Vlach & 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 a computational 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 modelling, 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.

Our work is an initial attempt at shedding light on the interaction of memory, attention and word learning, and understanding "desirable difficulty" in learning. Other factors (*e.g.*, working memory) might play a role in the performance of people as well. For example, because the number of items that people can store in their working memory is limited (Miller, 1956), the participants might store more trials in their working memory in the 2×2 condition, compared with the other conditions. The participants might use this information of the multiple trials (in their working memory) to make inferences about word–object mappings that repeat in successive trials. One future direction would be to incorporate a working memory module into our word learning model, and examine the impact of such inferences in a cross-situational learning scenario.

Acknowledgements

We would like to thank Haley Vlach for valuable discussion, and for providing us with the stimuli of her experiments. We are also grateful for the financial support from NSERC of Canada, and University of Toronto.

References

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought*. Lawrence Erlbaum Associates.

Fazly, A., Ahmadi-Fakhr, F., Alishahi, A., & Stevenson, S. (2010). Cross-situational learning of low frequency words: The role of context familiarity and age of exposure. *Proc. of CogSci*, *10*.

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

Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). Whats new? children prefer novelty in referent selection. *Cognition*, *118*(2), 234 - 244.

MacPherson, A. C., & Moore, C. (2010). Understanding interest in the second year of life. *Infancy*, *15*(3), 324–335.

Miller, G. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological review*, *63*(2), 81.

Nematzadeh, A., Fazly, A., & Stevenson, S. (2012a). A computational model of memory, attention, and word learning. In *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics.* Association for Computational Linguistics.

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

Pinker, S. (1989). *Learnability and cognition: The acquisition of argument structure*. Cambridge, Mass.: MIT Press.

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

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, *106*(3), 1558–1568.

Snyder, K. A., Blank, M. P., & Marsolek, C. J. (2008). What form of memory underlies novelty preferences? *Psychological Bulletin and Review*, *15*(2), 315 - 321.

Tomasello, M. (1992). The social bases of language acquisition. *Social development*, *1*(1), 67–87.

Vlach, H. A., & Sandhofer, C. M. (2010). Desirable difficulties in cross-situational word learning. In *Proceedings of the 32nd annual conference of the cognitive science society.*

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008, October 0). The Spacing Effect in Children's Memory and Category Induction. *Cognition*, *109*(1), 163–167.

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

Yurovsky, D., & Yu, C. (2008). Mutual exclusivity in crosssituational statistical learning. In *Proceedings of the 30th annual conference of the cognitive science society* (pp. 715–720).