The Interaction of Memory and Attention in Novel Word Generalization: A Computational Investigation

Erin Grant, **Aida Nematzadeh**, and **Suzanne Stevenson**
Department of Computer Science
University of Toronto
{eringrant, aida, suzanne}@cs.toronto.edu

Abstract

People exhibit a tendency to generalize a novel noun to the basic-level of a hierarchical taxonomy – a cognitively salient category such as "dog" – with the degree of generalization depending on the number and type of exemplars. Recently, a change in the presentation timing of exemplars has also been shown to have an effect, surprisingly reversing the prior observed pattern of basic-level generalization. We explore the precise mechanisms that could lead to such behavior by extending a computational model of word learning and word generalization to integrate cognitive processes of memory and attention. Our results show that the interaction of forgetting and attention to novelty, as well as sensitivity to both type and token frequencies of exemplars, enables the model to replicate the empirical results from different presentation timings. Our results reinforce the need to incorporate general cognitive processes within word learning models to better understand the range of observed behaviors in vocabulary acquisition.

Keywords: novel word generalization; word learning; computational modeling

Introduction

A number of computational models have successfully mimicked child behaviors in learning the meaning of words from ambiguous input (e.g., Siskind, 1996; Yu & Ballard, 2007; Frank et al., 2007; Fazly, Alishahi, & Stevenson, 2010). However, one challenge in word-meaning acquisition that has received less attention is that of *novel word generalization*: i.e., correctly identifying the level of a hierarchical taxonomy that a word refers to. After hearing it only a few times, how does the child determine, for example, that the word *dog* refers to Dalmatians, all dogs of different breeds, or any kind of animal? This issue poses difficulties to the learner because the accumulated evidence can be compatible with more than one of these choices. In this example, all dogs are also animals, and thus the meaning "animal" might also be consistent with all the usages of the word *dog*.

Xu and Tenenbaum (2007) (henceforth XT07) studied novel word generalization in both children and adults by observing decisions about category membership for novel objects in various experimental settings. One of their important findings concerned how people responded having seen 1 vs. 3 labeled exemplars of a certain kind of entity within a taxonomy. For example, having seen a single Dalmatian labeled as a *fep*, people assumed that the novel word *fep* could refer to the general category of dogs. However, if people saw several Dalmatians called *fep*, they apparently recognized that it would be a *suspicious coincidence* if *fep* meant "dog", but only one breed of dog was observed. In such cases, people had a lesser tendency to generalize to the higher level category than after seeing a single exemplar.

Spencer, Perone, Smith, and Samuelson (2011) (henceforth SPSS11) investigated the effect of presentation timing in the same task. XT07 had presented multiple exemplars of a novel word simultaneously. SPSS11 found that instead presenting exemplars in sequence reverses the suspicious coincidence effect. That is, after sequentially viewing three exemplars consistent with a more specific level of the taxonomy (e.g., three dogs of a single breed), people have a **greater** tendency to generalize to the higher category than after seeing one exemplar. SPSS11 explained this reversal as an interaction of word learning with the more general cognitive processes of attention and memory, which differ in their operation across the presentation types: People attend to and remember finer-grained similarities among objects when viewed simultaneously (e.g., that they are all Dalmatians), while the sequential presentation leads people to focus on the general commonalities of the objects (e.g., that they are all dogs).

Our goal in this paper is to provide a computational model that accounts for both the XT07 and SPSS11 findings in a well-motivated manner, by incorporating memory and attentional constraints into an incremental model of word learning and word generalization. It is desirable to integrate together all these pieces – novel word generalization, incremental word learning, and memory and attention – because: (i) word generalization is part and parcel of learning the meaning of words, since it allows the abstraction of meaning from a sequence of specific experiences, and (ii) many word-learning behaviors are influenced by the general cognitive processes of memory and attention (*e.g.*, Vlach et al., 2008; Samuelson & Smith, 2000). Importantly, by explicitly specifying such mechanisms within a computational model, we contribute to the precise understanding of the interactions between them that are required to account for empirical data.

Suspicious Coincidence: Data and Models

XT07 and SPSS11 explored how people generalized a novel word like *fep* to various levels of a taxonomy of objects (including animals, vehicles, and vegetables). Basic-level categories (e.g., dogs or trucks) are those whose members share a significant number of salient attributes; subordinate categories (e.g., Dalmatians or bulldozers) occur lower in the hierarchy, and their members share many fine-grained attributes; superordinate categories (e.g., animals and vehicles) are higher than the basic-level, and their members have fewer attributes in common (Rosch, 1973).

For the sake of space, we focus only on two of the train-

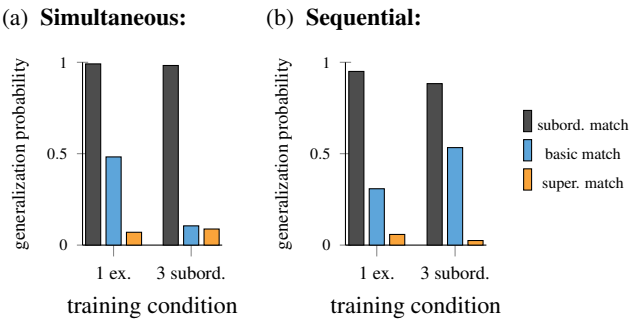
ing conditions in XT07 and SPSS11– the "1-example" and "3-subordinate" conditions – in which the suspicious coincidence effect and its reversal are seen.[1] The 1-example condition has one training trial in which participants observe a single object (*e.g.*, a Dalmatian) that is labeled with a novel word (such as *fep*). In the 3-subordinate condition, participants observe three instances from the same subordinate category (*e.g.*, three different Dalmatians) labeled with the novel word. In XT07's experiment, all three instances were presented *simultaneously*. SPSS11 included a condition in which the three instances are shown and labeled *sequentially*. (Simultaneous and sequential are the same for one example.)

After training, participants select all and only objects that they think are *feps* from a set of test items. Each test object is assessed as exactly one of the following types of match:

- a **subordinate match** has the same subordinate category as a training object (*e.g.*, a Dalmatian).
- a **basic-level match** has the same basic-level category as a training object (*e.g.*, a dog, but not a Dalmatian).
- a **superordinate match** has the same superordinate category as training objects (*e.g.*, an animal other than a dog).

Since SPSS11 replicated the pattern found by XT07 in their simultaneous presentation condition, we report only the results of SPSS11, as shown in Fig. 1.

Figure 1: **SPSS11 Behavioural Data**



SPSS11 data for (1a) simultaneous and (1b) sequential presentations. Each bar is the percent of chosen test objects of each type of match: subord(inate), basic(-level), or super(ordinate). Differences in 1-ex. across the two experiments were not statistically significant.

In the 1-example condition people generalized the novel word to refer to both subordinate matches (*e.g.*, Dalmatians) and (to a lesser extent) basic-level matches (*e.g.*, other kinds of dogs), but not to the superordinate matches (*e.g.*, other animals). This is in line with the idea that people tend to generalize a novel word to a basic-level category such as "dog" because of the perceptual salience of this level of categorization (*e.g.*, Markman, 1991).

In the 3-subordinate condition, when objects are presented simultaneously (Figure 1a), the generalization to the basic level is attenuated compared to the 1-example condition. XT07 explained this behavior as the suspicious coincidence effect. However, when objects are presented sequentially (Figure 1b), there was a surprising reversal of this effect.

While SPSS11 outline possible memory and attentional processes to explain their results, we know of no computational model that can account for both sets of data. XT07's Bayesian model formed hypotheses over a detailed hierarchical taxonomy to account for their own data, but it cannot model the difference between presentation timings, as SPSS11 note. The computational word learner of Nematzadeh, Grant, and Stevenson (2015) (henceforth NGS15) can model the XT07 results without the need for elaborated knowledge of the hierarchy or a built-in basic-level bias. Instead, the results of the model arise from a general type-token frequency interaction of the sort that commonly arises in explanations of linguistic phenomena (e.g., Bybee, 1985; Croft & Cruse, 2004). However, the timing of presentations also has no effect on the NGS15 model, and so the reversal of the suspicious coincidence effect is not achieved. In the next section, we explain how the NGS15 model can be naturally extended to integrate memory and attention, and therefore sensitivity to presentation timing.

Our Computational Model

We start with the NGS15 model because it uses an incremental word learning framework that mimics a range of behaviors in vocabulary acquisition (*e.g.*, Fazly, Alishahi, & Stevenson, 2010; Fazly, Ahmadi-Fakhr, et al., 2010). This framework has recently been extended to incorporate the effects of memory and attention on word learning (Nematzadeh, Fazly, & Stevenson, 2012), presenting a natural opportunity for integrating these processes within word generalization. We describe the NGS15 model, then the novel extensions that enable our model to replicate the SPSS11 data.

Learned Meanings in the NGS15 Model

The NGS15 model is a cross-situational learner that tracks weighted co-occurrences of words and semantic features across its input as in Fazly, Alishahi, and Stevenson (2010). The input to the model is intended to reflect the naturalistic input a child is exposed to, which consists of linguistic input (the words a child hears) paired with nonlinguistic data (the things a child perceives). An input pair is the set of words U_t and the set of semantic features S_t observed at time t:

U_t: { *look*, *a*, *fep* }
S_t: { PERCEPTION, LOOK, ..., DALMATIAN, DOG, ANIMAL }

The output of the model at each time t is a set of meaning probabilities, $P_t(f_i|w_j)$, for each feature f_i and each word w_j observed up through time t. The set of all conditional probabilities $P_t(f_i|w_j)$ for w_j represents the meaning of w_j.

The representation of meaning in NGS15 reflects the structure of taxonomic knowledge. Meaning features are arranged into *feature groups*, each corresponding to a level of the taxonomic hierarchy, as shown in Figure 2.For each word

[1] Our model replicates the results of XT07 and SPSS11 for all training conditions, but we only report the results for these two here.

w_j, a meaning probability distribution, $P_t(.|w_j)$, is calculated for each feature group; that is, $P_t(.|w_j)$ is normalized over the features in a group, rather than over all meaning features. The result is that features at the same level of the hierarchy, such as DALMATIAN and POODLE, or DOG and CAT, compete for probability mass; this ensures that such features, which are mutually incompatible given their taxonomic relationship, cannot simultaneously have high probability. Features at different levels of the hierarchy are in different feature groups and thus do not compete for probability mass; this ensures that meaning probabilities such as $P_t(\text{DALMATIAN}|fep)$, $P_t(\text{DOG}|fep)$, and $P_t(\text{ANIMAL}|fep)$ can all be highly activated if *fep* is intended to refer to a Dalmatian (which is also both a dog and an animal). In this approach, the meaning of a word is the set of n distributions, $P_t(.|w_j)$, one per feature group in a taxonomy with n levels.

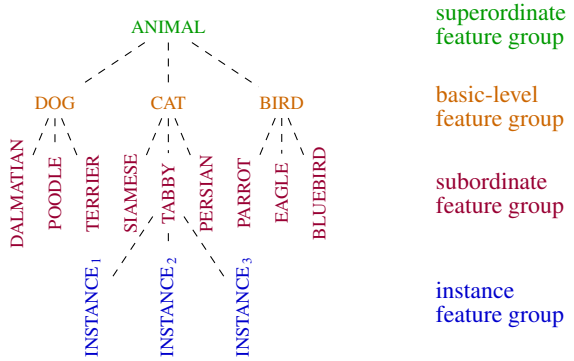


Figure 2: A portion of the taxonomy used in this paper.

The NGS15 Learning Algorithm

The input to the model is processed in an incremental two-step bootstrapping framework: Words and features that co-occur are **aligned** (associated) in proportion to the current meaning probabilities, which are then **updated** with the new evidence regarding strength of association, as follows.

The Alignment Step. For an input pair at time t, the model calculates a probabilistic strength of aligning (associating) each word $w_j \in U_t$ with each feature $f_i \in S_t$:

$$a_t(f_i, w_j) = \frac{P_{t-1}(f_i|w_j)}{\sum_{w' \in U_t} P_{t-1}(f_i|w')} \tag{1}$$

These alignment strengths are incrementally accumulated as:

$$\begin{aligned} \text{assoc}_t(f_i, w_j) &= \text{assoc}_{t-1}(f_i, w_j) + a_t(f_i, w_j) \\ &= \sum_{t' \in \mathcal{T}} a_{t'}(f_i, w_j) \end{aligned} \tag{2}$$

where $\mathcal{T}$ is all times at which f_i and w_j have co-occurred.

In our model, if more than one instance of feature f_i occurs with word w_j at time t, multiple instances of $a_t(f_i, w_j)$ are recorded. For example, in the simultaneous presentation of three exemplars with the word *fep*, the alignment strength $a_t(f, fep)$ will be added three times to the association score for each feature f in the input.

Update of Meaning Probabilities The model next uses the association scores to update the meaning probabilities. Each meaning probability $P_t(f_i|w_j)$ represents the magnitude of the f_i–w_j association *relative to* the association strength between w_j and other features within the same feature group $\mathcal{G}$ as f_i:

$$P_t(f_i|w_j) = \frac{\text{assoc}_t(f_i, w_j) + \gamma_{\mathcal{G}}^t}{\sum_{f_m \in \mathcal{G}} \text{assoc}_t(f_m, w_j) + k_{\mathcal{G}} \gamma_{\mathcal{G}}^t} \tag{3}$$

Here $k_{\mathcal{G}}$ and $\gamma_{\mathcal{G}}^t$ are smoothing terms: $k_{\mathcal{G}}$ reflects the expected number of features in $\mathcal{G}$ and $\gamma_{\mathcal{G}}^t$ represents the *a priori* tendency to observe a feature in $\mathcal{G}$. While $k_{\mathcal{G}}$ is a fixed parameter, $\gamma_{\mathcal{G}}^t$ is a function of the number of observed types within the feature group $\mathcal{G}$, and thus changes over time (see NGS15).

The $\gamma_{\mathcal{G}}^t$ parameters are key to the generalization behavior of the NGS15 model because they influence how much probability mass is allocated to a feature previously unseen with a word (cf. Eqn. 3 when the assoc score is 0). A higher value for $\gamma_{\mathcal{G}}^t$ leads to more probability mass allocated to previously unseen features in group $\mathcal{G}$, allowing for more generalization to new features in that group. Because $\gamma_{\mathcal{G}}^t$ increases with the number of types, it captures the oft-observed tendency in language that people more readily generalize categories for which a greater variety of types of items has been observed. The model matches the child data from XT07 by equating $\gamma_{\mathcal{G}}^0$ across feature groups. But to match the adults, who show a stronger basic-level bias, the model required that the $\gamma_{\mathcal{G}}^0$ parameters be initialized to successively higher values for feature groups successively lower in the hierarchy, entailing that, e.g., it is easier to generalize a novel word to a new breed of dog not seen in training (basic-level generalization), than to a new kind of animal not seen in training (superordinate generalization).

Our Extensions to Integrate Memory and Attention

To render the model sensitive to presentation timing, we adopt the general approach of Nematzadeh et al. (2012), which integrates memory and attention seamlessly into the cross-situational word-learning mechanism. The approach was shown to account for spacing effects in word learning, which are closely related to the presentation timing factors considered by SPSS11. However, the methods must be extended to adequately meet the needs of word generalization in the NGS15 model; we describe those extensions here.

Modeling the Effects of Forgetting. To model the effect of memory, we use the association score formulation of Nematzadeh et al. (2012), which implements "forgetting" by applying a decay factor to each alignment probability (cf. Eqn. 2 above):

$$\text{assoc}_t(f_i, w_j) = \sum_{t' \in \mathcal{T}} \frac{a_{t'}(f_i, w_j)}{(t - t' + 1)^{d_{a_{t'}}}} \tag{4}$$

Each alignment in the sum is scaled by the temporal distance between the current time t and the time t' that the alignment

was made, exponentiated to a decay function $d_{a_{t'}}$ that is inversely proportional to the strength of alignment.

However, we must extend this decay formulation to accommodate our hierarchical knowledge of feature groups.[2] In particular, we find that using the same decay rate across all feature groups is not sufficient. As noted above, appropriate word generalization in the NGS15 model requires that lower levels in the taxonomy be more "open" to generalizing to new features than higher levels in the taxonomy. It is important to note that the decay of alignments also influences "openness" to generalization because it shifts probably mass away from observed word–feature pairs onto unseen events. Thus, to appropriately reflect the nature of the hierarchy – that openness increases with greater depth in the taxonomy – we must parameterize decay by feature group. Just as feature groups lower in the taxonomy must have successively higher γ values to indicate more "openness" to generalization, lower feature groups also require higher decay rate parameters.

We thus use the following formulation of decay:

$$d_{a_t} = \frac{d_{\mathcal{G}}}{a_t(f_i, w_j)} \tag{5}$$

where $d_{\mathcal{G}}$ controls the rate of decay for features in feature group $\mathcal{G}$, and is set successively higher for lower-level feature groups in the taxonomy.

Modeling Attention to Novelty. Building on research showing that people attend more to novel stimuli in learning (e.g., Snyder et al., 2008; MacPherson & Moore, 2010; Horst et al., 2011), we use the general idea of Nematzadeh et al. (2012) in allocating more strength to alignments that are more novel (cf. Eqn. 1):

$$a_t(f_i, w_j) = \frac{P_{t-1}(f_i|w_j)}{\sum_{w' \in U} P_{t-1}(f_i|w')} \cdot \text{novelty}_t(f_i, w_j) \tag{6}$$

In this model, $\text{novelty}_t(f_i, w_j)$ was inversely proportional to how recently w_j had been observed, and thus focused solely on novelty of words; the novelty of the feature f_i was not considered. We must broaden this approach because the experiments here are focused on a single novel word.

Here instead we consider the novelty of the observed *word–feature pairing*, and again draw on considerations of type–token frequencies, as in other aspects of the NGS15 model. Specifically, we scale the alignment strength by the ratio of the token frequency of f_i–w_j observations at time t to the total frequency of all such observations, by formulating $\text{novelty}_t(f_i, w_j)$ as:

$$\text{novelty}_t(f_i, w_j) = \frac{\text{token}_t(f_i, w_j)}{\sum_{t' \in \mathcal{T}} \text{token}_{t'}(f_i, w_j)} \tag{7}$$

where $\text{token}_t(f_i, w_j)$ is the number of tokens of feature f_i that occurred at time t with word w_j.

[2] Nematzadeh et al. (2012) used a single meaning probability distribution over all features – i.e., there are no feature groups.

This formulation achieves attention to novelty as follows. Generally, earlier observations of feature f_i with word w_j will have a stronger alignment than later observations, where the increased number of observations will increase the denominator of $\text{novelty}_t(f_i, w_j)$, and lead to attenuation of the alignment strength. Note that when the co-occurrence of f_i with word w_j is truly novel – *i.e.*, the first time they are observed together – the strength of alignment is undiminished, since the numerator and denominator of the novelty factor are equal in the initial observation of f_i with w_j.

Summary of Novel Extensions to the NGS15 Model In summary, we have extended both the model of NGS15, and the memory and attention mechanisms of Nematzadeh et al. (2012), by: (i) incorporating a forgetting mechanism that is sensitive to the taxonomic level of a feature group, which reflects the needs of taxonomic structure and the process of novel word generalization; and (ii) formulating a mechanism for attention to novelty of word–feature pairings, rather than just to recency of words, consistent with the key role of word–feature association statistics in the model.

These mechanisms have a direct impact on the processing of stimuli in simultaneous vs. sequential presentations in a novel word generalization task. The forgetting mechanism ensures that more general features, such as the kind of animal observed (*e.g.*, dog or cat), are remembered better than more detailed features, such as particular breeds of dogs. The attention-to-novelty mechanism has the consequence that successive observations of word–feature pairings in a sequential presentation scenario are "discounted" with respect to earlier presentations. We demonstrate in our experiments below that, together, these mechanisms interact to enable the model to account for both the suspicious coincidence effect in a simultaneous presentation as found by XT07, and its reversal in a sequential presentation as found by SPSS11.

Methodology

We follow the methods of NGS15, adapted where needed for our extended model on the SPSS11 data.[3]

Training the Model. We use a taxonomy with three levels, corresponding to the subordinate, basic, and superordinate categories of animals. This yields four feature groups, one per category level plus an "instance" group to distinguish multiple objects of the same subordinate category. See Figure 2. In each U_t–S_t input pair, U_t consists of the novel word, and S_t is a set of four features (one per feature group) representing a unique instance of the same subordinate category across all training trials; for example:[4]

U_t: { *fep* }
S_t: { INSTANCE$_1$, DALMATIAN, DOG, ANIMAL }

[3] Our code and data are available at https://github.com/eringrant/novel_word_generalization.

[4] Each FEATURENAME stands for all features of an object at that level of the hierarchy. Such features could be replaced with an appropriate *set* of features without changing the model results.

In the 1-example condition, training consists of just one such U_t–S_t pair. In the 3-subordinate condition, training has three such U_t–S_t pairs, differing only in the unique instance feature (*i.e.*, INSTANCE$_1$, INSTANCE$_2$, INSTANCE$_3$) in each S_t. In the simultaneous condition, the three U_t–S_t pairs are all presented at the same time t. In the sequential condition, the three U_t–S_t pairs are presented one at a time, at t, $t+1$, and $t+2$.

Testing the Model. After training, the level of generalization of the novel word is assessed against test objects, each of which is a subordinate match, a basic-level match, or a superordinate match; for example:

subord. match:	{ INSTANCE$_4$, DALMATIAN, DOG, ANIMAL }
basic match:	{ INSTANCE$_5$, POODLE, DOG, ANIMAL }
super. match:	{ INSTANCE$_6$, TOUCAN, BIRD, ANIMAL }

We adapt the P_{gen} formula of NGS15 to test whether the model generalizes the learned meaning of the novel word w to the various levels of match at test time T (after training):

$$P_{gen}(m|w) = \frac{\text{avg}_{Y \in m}\, P_T(Y|w)}{\text{avg}_{Y' \in \{\text{sub.}\}}\, P_T(Y'|w)}$$

Here $P_T(Y|w)$ is the probability of a test object Y given w, and m is the set of test objects at a certain level of match. The measure in the numerator of P_{gen} is the average such probability across test matches at that level, $\text{avg}_{Y \in m}\, P_T(Y|w)$. This is not directly comparable to the empirical data, which are the percentages of test objects selected from each type of match. To obtain a comparable measure, we scale each probability (for each level of match) by the probability of the subordinate matches in that condition, $\text{avg}_{Y' \in \{\text{sub.}\}}\, P_T(Y'|w)$ (the denominator of P_{gen}). Thus $P_{gen}(m|w)$ is the relative average preference for test items at level m. This renders the subordinate match probability as 1.0 (reflecting that people generally pick close to 100% of the subordinate test items), and shows the other type of matches relative to that amount.

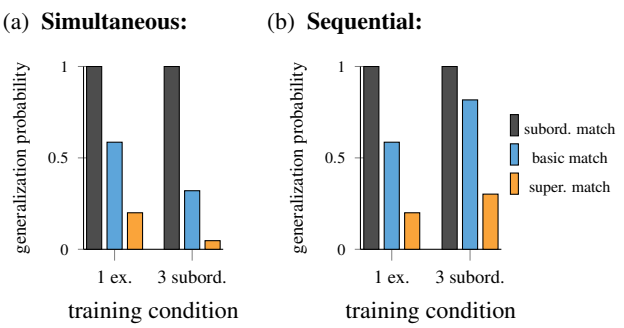
Model Parameters. Since the SPSS11 participants are adults, we use the *adult parameter settings* of NGS15 for the four $\gamma_{\mathcal{G}}^0$ parameters and the four $k_{\mathcal{G}}$ (one each per feature group), which are tuned to achieve a match to adult data of XT07. For our decay parameters, we use:

$$d_{\text{inst}} = 0.8 \quad d_{\text{subord}} = 0.5 \quad d_{\text{basic}} = 0.05 \quad d_{\text{super}} = 0.01$$

Model Results and Discussion

Figure 3 shows the results of our model in the simultaneous and sequential conditions; cf. Figure 1 for the human behavioral data in SPSS11. Following simultaneous presentation of training input (Figure 3a), our model shows the suspicious coincidence effect: Generalization to the basic level is inhibited in the 3-subordinate condition as compared to the 1-example condition. In contrast, sequential presentation reverses the suspicious coincidence effect (Figure 3b): the model exhibits greater basic-level generalization in the 3-subordinate condition. Thus, these results replicate the qualitative pattern evident in the behavioural data of SPSS11.

Figure 3: **Our Model Data**



Our model data for (3a) simultaneous and (3b) sequential presentations. Each bar is the probability of a type of test match: *i.e.*, subord(inate), basic(-level), or super(ordinate), scaled by the subordinate match probability (see text).

The interaction (between presentation type and amount of training) seen in the human data arises as a result of a corresponding interaction in the model. Consider each 3-subordinate condition (simultaneous and sequential) compared to the 1-example condition. **In the simultaneous 3-subordinate case**, the attentional mechanism yields higher alignment strengths between the word and features because their three co-occurrences are all novel at the single presentation time; in addition there is little forgetting because the items are all seen at time t and test is at time $t+1$. This yields stronger subordinate alignments compared to the 1-example case, and therefore somewhat less basic-level generalization.

By contrast, **in the sequential 3-subordinate case**, the word–feature co-occurrences are less salient because they decrease in novelty over the three presentation times. In addition, greater forgetting occurs because there is more time between the (first two) presentation times and test time ($t+4$). In this case, because subordinate features decay faster than basic features, the interaction yields weaker subordinate features compared to the 1-example case, and more basic-level generalization is achieved.

The interaction of memory and attention effects are required to obtain this pattern of results in the model. If the model includes only the decay mechanism, differentiated by taxonomic level, this enables it to focus more on abstract than specific features, and consequently raises the basic generalization closer to the level of the subordinate generalization in *all* conditions. On the other hand, using the attention mechanism alone enables the model to distinguish the sequential and simultaneous conditions, but it cannot on its own raise the basic generalization high enough. Only when the two are used together does the model produce the reversal of the suspicious coincidence effect in the sequential presentation.

The necessity of both memory and attention is suggestive of how word learning occurs in people. In particular, the attention mechanism in the model focuses more probability onto word–feature co-occurrences in their earlier presentations, simulating the general tendency for people to attend more to less-familiar things. In addition, the mechanism that

increases the decay rate for lower-level features in the taxonomy simulates the tendency in people to remember abstract features of objects over very specific features. SPSS11 contended that people are able to attend to specific features in the simultaneous condition due to the close spatial and temporal proximity of the items, and correspondingly attend only to the abstract commonalities of items in sequential presentations. Our model explains this effect as the result of general memory and attention mechanisms that have been shown to play a role in word learning more widely (cf. Nematzadeh et al. (2012); Nematzadeh, Fazly, and Stevenson (2013)). Interestingly, attention in our model is a function of the token frequency of word–feature co-occurrences (as opposed to a fixed parameter) and is therefore a response to the statistics of the data, as are other components of our word generalization formulation. All this further supports that attention, memory and statistical learning interact to produce the suspicious coincidence effect and its reversal across presentations.

Conclusions and Future Work

Novel word generalization – understanding how a word maps to the appropriate level of a taxonomic hierarchy – is an important aspect of novel word learning, but one that has not received much attention in the word-learning community. We propose a unified model of word learning that accounts for the various observed patterns of novel word generalization – in particular, the suspicious coincidence effect (Xu & Tenenbaum, 2007) and its reversal under differing presentation conditions (Spencer et al., 2011). We extend the model of Nematzadeh et al. (2015) with a novel integration of the general cognitive mechanisms of memory and attention, and show that our model's success is a result of the interaction of forgetting and attention to novelty of word–feature co-occurrences. Our approach builds on the earlier NGS15 model in highlighting the importance of type and token frequency patterns in the input to capturing interesting generalization effects, but here these patterns are manifest in our formulation of memory and attention mechanisms.

In incorporating these cognitive processes into our model, we drew on the approach of Nematzadeh et al. (2012), whose model had been shown to account for various spacing effects in word learning (see also Nematzadeh et al., 2013). Much further work is needed to explore whether our model can explain other such effects. For example, Vong, Perfors, and Navarro (2014) showed that people's categorization of novel object instances depends on the distribution of training examples both that are labelled with a word as well as those that are unlabelled. Currently, our model only takes into account word–feature co-occurrences, and is therefore insensitive to features that occur without a word label. We will need to consider how to integrate learning from unlabelled data in order to better model how statistical word learning interacts with object categorization, as it does in people.

References

Bybee, J. L. (1985). *Morphology: A study of the relation between meaning and form*. Philadelphia: Benjamins.

Croft, W., & Cruse, A. (2004). *Cognitive linguistics*. Cambridge University Press.

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. In *CogSci Proceedings* (Vol. 10, pp. 2362–2367).

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

Frank, M. C., Goodman, N. D., & Tenenbaum, J. B. (2007). A Bayesian framework for cross-situational word-learning. In *NIPS Proceedings* (Vol. 20, pp. 457–464).

Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What's 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.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. MIT Press.

Nematzadeh, A., Fazly, A., & Stevenson, S. (2012). A computational model of memory, attention, and word learning. In *CMCL Proceedings* (pp. 80–89).

Nematzadeh, A., Fazly, A., & Stevenson, S. (2013). Desirable difficulty in learning: A computational investigation. In *CogSci Proceedings* (pp. 1073–1078).

Nematzadeh, A., Grant, E., & Stevenson, S. (2015). A computational cognitive model of novel word generalization. In *EMNLP Proceedings* (pp. 1795–1804).

Rosch, E. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive development and the acquisition of language* (p. 111-144).

Samuelson, L. K., & Smith, L. B. (2000). Grounding development in cognitive processes. *Child Develop.*, 98–106.

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

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.

Spencer, J. P., Perone, S., Smith, L. B., & Samuelson, L. K. (2011). Learning words in space and time: Probing the mechanisms behind the suspicious-coincidence effect. *Psychological science*, *22*(8), 1049–1057.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*(1), 163–167.

Vong, W. K., Perfors, A., & Navarro, D. J. (2014). The relevance of labels in semi-supervised learning depends on category structure. In *CogSci Proceedings* (p. 1718-1723).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psych. Rev.*, *114*(2), 245–272.

Yu, C., & Ballard, D. H. (2007). A unified model of early word learning: Integrating statistical and social cues. *Neurocomputing*, *70*(13-15), 2149–2165.