A Computational Cognitive Model of Novel Word Generalization

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

Abstract

A key challenge in vocabulary acquisition is learning which of the many possible meanings is appropriate for a word. The word generalization problem refers to how children associate a word such as *dog* with a meaning at the appropriate category level in a taxonomy of objects, such as Dalmatians, dogs, or animals. We present the first computational study of word generalization integrated within a word-learning model. The model simulates child and adult patterns of word generalization in a word-learning task. These patterns arise due to the interaction of type and token frequencies in the input data, an influence often observed in people's generalization of linguistic categories.

1 Introduction

Learning word meanings is a challenging early step in child language acquisition. Imagine a child hears the word *dax* for the first time while observing a white rabbit jumping around – *dax* might mean WHITE RABBIT, RABBIT, ANIMAL, CUTE, LOOK, etc. (Quine, 1960). How does the child learn the correct meaning of a word from a large pool of potential meanings? A possible explanation is that children infer a word's meaning by identifying the commonalities across the situations in which the word occurs (Pinker, 1989). One mechanism for achieving this is *cross-situational learning* (*e.g.*, Siskind, 1996; Frank et al., 2007; Fazly et al., 2010; Kachergis et al., 2012). Recent word learning experiments confirm that both adults and children infer the correct word-meaning mappings by keeping track of cross-situational statistics across individually ambiguous learning trials (Yu and Smith, 2007; Smith and Yu, 2008; Yurovsky et al., 2014).

Although cross-situational learning is a general mechanism for narrowing down the meaning of a word, it does not explain how children overcome an interesting challenge in word learning: determining the correct level of a hierarchical taxonomy that a word refers to. For example, children learn that the word *dog* refers to all kinds of dogs, and not to a specific breed, such as Dalmatians, or to a more general category, such as animals – even though some of these choices (*e.g.*, animals) are compatible with all the cross-situational evidence available for *dog* (because all dogs are also animals). We use the term "word generalization" to refer to this problem of associating a word with the meaning at an appropriate category level, given some sample of experiences with the word.

Previous research has argued that children use a specific bias or constraint – the *basic-level assumption* – to focus their word generalizations appropriately (Markman, 1991; Golinkoff et al., 1994). According to this bias, children prefer to associate a word to a set of objects that form a *basic-level* category, such as dogs or trucks, and that share a significant number of attributes. It is less preferred to associate a new word to much more specific *subordinate* categories, such as Dalmatians or bulldozers, or to more general *superordinate* ones, like animals or vehicles, whose members share fewer attributes (Rosch, 1973; Rosch et al., 1976). It remains an important open question of whether a word learner requires such a bias to acquire appropriate mappings.

Xu and Tenenbaum (2007) (X&T henceforth) studied the word generalization problem in a set of experiments in which children and adults were asked to determine which level of a taxonomy a novel word referred to. X&T further examined this behavioral data through computational modelling. They proposed a Bayesian model that, given a few exemplars of a novel word, matches human behaviour in how it maps the word to its

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1795–1804,
Lisbon, Portugal, 17-21 September 2015. ©2015 Association for Computational Linguistics.

meanings in a taxonomic category.

The Bayesian model of X&T is important in providing insight into how people might reason about samples of data that exemplify categories. However, it relies on having complete, built-in knowledge about the taxonomic hierarchy, including both the detailed composition of categories and the values for between-object similarities, drawn from adult similarity judgments. Furthermore, the X&T model does not address the issue of word generalization in the broader context of word learning: While their model reasons over samples of data associated with a word label, it does not develop a meaning representation of the word over time, as a child must do. It is important to understand how word generalization occurs when embedded in the natural process of learning a word meaning and in the context of more limited category knowledge.

We address these issues by providing a unified account of word learning and word generalization within a computational model of cross-situational learning. Unlike the X&T model, our model is an incremental learner that gradually acquires the meaning of words, and uses these developing meanings in determining the appropriate extension of a word to elements of a taxonomy. Our model has general knowledge of category structure without having an elaborated taxonomy encoding known object similarities. Moreover, in the absence of any bias toward generalization to particular kinds of categories, the model exhibits the observed "basic-level bias" due to general mechanisms of productivity that have been proposed to apply to many aspects of linguistic knowledge (*e.g.*, Bybee, 1985; Croft and Cruse, 2004).[1]

In what follows, we first describe the human experiments of X&T, and then present our computational model and the experiments that simulate the X&T data.

2 Novel Word Generalization in People

X&T perform a set of empirical studies to investigate how children and adults generalize novel words learned from a few examples to the appropriate level of meaning in a taxonomy. In each training trial of an experiment, participants hear a novel word (such as *fep*) and observe one or more instances exemplifying the word (in the form of pictures for adults and toy objects for children). The conditions vary in that the make-up of the set of training instances is representative of different levels of a taxonomy (*e.g.*, all Dalmatians vs. various kinds of dogs vs. various kinds of animals). In the testing phase, participants are asked to select all objects that they think are *feps* from a set of test items. Both children and adults make various inferences about what a *fep* is depending on the levels of the taxonomy from which the training instances are drawn.

Specifically, X&T use a taxonomy with animals, vehicles, and vegetables, from which instances are drawn to produce the training conditions in Fig. 1(a). For example, in one training condition, participants are shown a Dalmatian, a poodle, and a beagle in three consecutive trials, hearing the word *fep* to refer to each object. After training, participants are asked to select all *feps* from the set of test objects, which includes items from all 3 superordinate categories. As illustrated in Fig. 1(b), each test object is assessed as one of the following types of match to the training data:

- a **subordinate match**: an object of the same subordinate category as a training object (*e.g.*, Dalmatians in Fig. 1)
- a **basic-level match**: an object of the same basic-level category as a training object (*e.g.*, a dog, but not the same breed as one in training [which would be a subordinate match])
- a **superordinate match**: an object of the same superordinate category as the training objects (*e.g.*, another kind of animal, but not one seen in training [which would be a subordinate or basic-level match])

X&T report the percentage of test objects of each type of match that are selected by participants within each training condition; see Fig. 2. (For example, the reported value for "super. match" would be 75% if participants on average chose 3 of the 4 superordinate matches in the test set.)

Consider first the data from adults. After seeing a single object (1-example condition – *e.g.*, a Dalmatian), adults show a strong basic-level bias – i.e., they tend to generalize the word *fep* to refer to *both* Dalmatians (subordinate matches) and to

[1] Computational cognitive models are often categorized with respect to Marr's levels of analysis, *i.e.*, their degree of abstraction (Marr, 1982). The model of X&T is at the computational level, providing a Bayesian framework for the problem of word generalization. In contrast, our model investigates more detailed mechanisms and thus lies between the algorithmic and computational levels of analysis.

Training condition:	Example Trials: Trial 1	Trial 2	Trial 3
1 example	Dalmatian	∅	∅
3 subord.	Dalmatian	Dalmatian	Dalmatian
3 basic	Dalmatian	poodle	beagle
3 super.	Dalmatian	penguin	sheep

(a) An example of the training conditions.

Example test object selected:	Type of test match:
Dalmatian	subordinate match
bulldog	basic match
cat	superordinate match

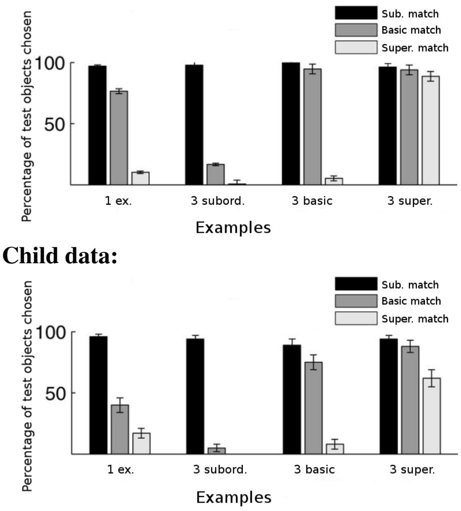
(b) An example of the test responses.

Figure 1: The training and testing conditions of X&T; see text in Section 2.

other dogs (basic-level matches), but not to other animals (superordinate matches). But with 3 instances of a Dalmatian (3-subordinate condition), this behaviour is attenuated – the number of basic-level matches is much lower. For the 3-basic-level and 3-superordinate conditions, the adults show generalization up to categories consistent with the evidence – i.e., at the basic and superordinate levels, respectively.

Interestingly, children also show a basic-level bias, but differ from adults in that it is less pronounced – e.g., they are less likely than adults to select basic-level matches (other dogs) having seen a single Dalmatian or having seen 3 Dalmatians. In the other conditions, children's behaviour is similar to adults, but shows somewhat less generalization to unseen types of objects (*e.g.*, other kinds of dogs/animals than those in training).

(a) **Adult data:**



(b) **Child data:**

Figure 2: X&T data for (a) adults and (b) children. Each bar is the percentage of chosen test objects of a type of test match: *i.e.*, subord(inate), basic(-level), or super(ordinate).

3 The Word Learning Framework

Our computational model is based on the cross-situational word learner of Fazly et al. (2010) (henceforth, FAS), which accounts for a range of observed patterns in child and adult vocabulary acquisition. Here we give an overview of the FAS model; the next section explains extensions to handle the novel word generalization task.

A naturalistic language learning scenario consists of both linguistic data (what a child a hears) and non-linguistic data (what a child perceives). This input is modeled as a sequence of utterance–scene (U–S) pairs, where an utterance is a group of words and a scene is a set of semantic features representing the meaning of those words:

U: { *look*, *a*, *fep*, ... }
S: { PERCEPTION, LOOK, ..., DALMATIAN, DOG, ... }

Given such input, for each word w, the model of FAS learns a probability distribution over all semantic features, $P_t(.|w)$, which represents the word's meaning at time t. Initially, at time $t = 0$, $P_0(.|w)$ is a uniform distribution. The word meanings are incrementally learned using an algorithm that implements cross-situational learning: for each pair of a word w and a semantic feature f, the model learns $P_t(f|w)$ from co-occurrences of w and f across all the utterance–scene pairs seen up to time t, as follows.

Given an utterance–scene pair U–S at time t, and drawing on its learned knowledge of word meanings up to time $t-1$, the model of FAS calculates an *alignment* probability for each w_j–f_i pair. This probability reflects how strongly the feature f_i is associated with w_j compared to its association with other words in U:

$$P_t(a_{ij}|U, f_i) = \frac{P_{t-1}(f_i|w_j)}{\sum\limits_{w' \in U} P_{t-1}(f_i|w')} \qquad (1)$$

where a_{ij} indicates the mapping between the word w_j and the semantic feature f_i.

These probabilities are incrementally accumulated for each w_j–f_i pair, capturing the overall strength of association of w_j and f_i at time t:

$$\begin{aligned} \text{assoc}_t(f_i, w_j) = \text{assoc}_{t-1}(f_i, w_j) \\ + P_t(a_{ij}|U, f_i) \end{aligned} \qquad (2)$$

The (normalized) association scores then serve as the basis for the incremental adjustment of the meaning probabilities of all features f_i for each word w_j seen in the input at time t:

$$P_t(f_i|w_j) = \frac{\text{assoc}_t(f_i, w_j) + \gamma}{\sum\limits_{f_m \in \mathcal{M}} \text{assoc}_t(f_m, w_j) + k\,\gamma} \qquad (3)$$

Here $\mathcal{M}$ is the group of all features that the model has observed, k is the expected number of such features, and γ is a small smoothing parameter, which determines the prior probability of observing a new feature.

Smoothing entails that features previously unseen with a word (all f_i such that $\text{assoc}_t(f_i, w_j) = 0$) have a small but non-zero probability. That is, when f_i is unseen with w_j, Eqn. (3) reduces to:

$$P_t^u(f_i|w_j) = \frac{\gamma}{\sum\limits_{f_m \in \mathcal{M}} \text{assoc}_t(f_m, w_j) + k\,\gamma} \qquad (4)$$

This unseen probability, P_t^u, reflects the learner's "openness" to the word being associated with new features (Nematzadeh et al., 2011): a higher or lower $P_t^u(f_i|w_j)$ will affect how strongly a previously unseen f_i can be associated with w_j in the alignment process (Eqn. (1)). We return to this property of the model below, as it relates to the behaviour of our model in making generalizations.

4 Extensions to the Model

We assume that the representation of meaning can be abstracted to features that correspond to different levels of categorization. For example, a Dalmatian in an input scene is represented as {DALMATIAN, DOG, ANIMAL} and a Bulldog as {BULLDOG, DOG, ANIMAL}, where we use FEATURENAME to refer to all the features that are specific to that level of object category. (Note that we could replace each of these features with the appropriate "true" set of features, but use the more compact representation for simplicity.) To acquire the meaning of the word *Dalmatian*, the model must learn a probability distribution in which $P(f|\textit{Dalmatian})$ is relatively high for the features DALMATIAN, DOG, and ANIMAL, and low for features such as BULLDOG, CAT, and VEGETABLE.

Introducing Feature Groups. In the FAS model, all the features for a word are dependent: increasing the probability of any feature results in decreasing the probability of others. However, this interaction is not always desirable, as many features regularly co-occur in the world. This is especially an issue for features from a category hierarchy, where features of a subordinate category should not compete with features of the parent. That is, while a higher probability of DALMATIAN features (*e.g.*, black spotted coat) may lessen the likelihood of BULLDOG features (*e.g.*, wrinkles), it should *not* decrease the probability of DOG features (*e.g.*, having 4 legs).

To address this, we extend the model by using *feature groups* that collect together sets of features that sensibly compete. Each feature group is comprised of all features at the same level of specificity in the category hierarchy, which are therefore mutually exclusive, such as DOG, CAT, and BIRD (*i.e.*, different kinds of animals). Instead of learning a single probability distribution over all features as the meaning of a word, the extended model learns *a set of* probability distributions for a word, one for each feature group (*i.e.*, one per level of the hierarchy). Features within a group thereby compete for the probability mass associated with a word, but those from across groups (*e.g.*, DALMATIAN and DOG) can freely co-occur without competing.

The model does not know a priori all the features in a group, but when presented with a newly observed feature, it can identify the appropriate group for it. In taking this approach, we assume the learner can distinguish the level of specificity of features perceived in the scene. For example, in the scene representations {DALMATIAN, DOG, ANIMAL} and {SIAMESE, CAT, ANIMAL}, the learner can recognize that DOG and CAT are at the same level of the hierarchy (kinds of animals) and that DALMATIAN and SIAMESE are at the same, *more specific* level in the hierarchy (finer-grained breeds of animals). Our assumption is that children (at this stage in their development) can identify a degree of similarity among concepts that enables them to recognize that Dalmatians and Siamese are distinguished by similar properties (such as fur color), which differ from more distinguishing properties at higher taxonomic levels (such as number of legs). The model has no other prior knowledge of the category structure. For example, it is not built into the model that DALMATIAN is a type of DOG, only that it is more specific than DOG; any association between them would be learned from their pattern of co-occurrence with a word over time.

Note that, in contrast to the model of X&T, our model does not start with a full taxonomy (it does not know, for example, that Dalmatians and poodles are hyponyms of dogs) and it does not have built-in knowledge of similarities among concepts. Still, it encodes some taxonomic knowledge in the feature groups, and an important future direction will be to show that this knowledge is learnable from the input.

Calculating Feature Group Probabilities. To appropriately split the probability mass within a feature group $\mathcal{G}$ (but not across feature groups), we use a new formulation of Eqn. (3) to update the meaning probabilities for $f_i \in \mathcal{G}$ as follows:

$$P_t(f_i|w_j) = \frac{\text{assoc}_t(f_i, w_j) + \gamma_{\mathcal{G}}}{\sum\limits_{f_m \in \mathcal{G}} \text{assoc}_t(f_m, w_j) + k_{\mathcal{G}}\gamma_{\mathcal{G}}} \quad (5)$$

where $k_{\mathcal{G}}$ is the expected number of features in $\mathcal{G}$, and the smoothing factor $\gamma_{\mathcal{G}}$ reflects the prior belief in observing a feature f in $\mathcal{G}$.[2]

With this new formulation, the probability of a feature f_i previously unseen with word w_j now reduces to (cf. Eqn. (4)):

$$P_t^u(f_i|w_j) = \frac{\gamma_{\mathcal{G}}}{\sum\limits_{f_m \in \mathcal{G}} \text{assoc}_t(f_m, w_j) + k_{\mathcal{G}}\gamma_{\mathcal{G}}} \quad (6)$$

for $f_i \in \mathcal{G}$. Note that the smoothing factor $\gamma_{\mathcal{G}}$ depends on $\mathcal{G}$, and thus the openness of the word to be associated with new (previously unseen) features can vary depending on the feature group.

This unseen probability is very important to the model's generalization behaviour. Generalization involves the model accepting that a learned word can refer to objects not seen with it before: e.g., in the experiments here, we would expect that the learned meaning for *fep* after seeing three animals such as a dog, a penguin, and a sheep could also accommodate the meaning of a different animal such as a cat. This ability of the model to associate new meaning features with a word depends precisely on the unseen probability formulation: the higher the unseen probability for a feature and a word, the more the feature will be acceptable as a generalization of the word.

Type-Token Effects on Generalization. The unseen probability is sensitive to how many instances of features from a group have already been seen with a word w_j: As the model observes more instances (tokens) of features from $\mathcal{G}$ with w_j, the corresponding $assoc_t$ score(s) increase, thereby increasing the denominator in Eqn. (6) and decreasing P_t^u. Thus the tendency to generalize w_j to more features in $\mathcal{G}$ – i.e., to accept additional features as part of the meaning of w_j – will decrease as the model has more evidence of (observed) features in that group occurring with w_j.

Generalization of a category to include new kinds of items is typically a function of both token *and* type frequency (*e.g.*, Bybee, 1985; Croft and Cruse, 2004): a category with more diverse types is more easily extended to new cases. While the evolving association scores capture the effect of observing more feature *tokens*, our model as given does not distinguish the number of different *types* of features seen within a group (*e.g.*, two DOGs vs. one DOG and one CAT).

We address this issue by having $\gamma_{\mathcal{G}}$ depend on the number of observed types of features in the group:

$$\gamma_{\mathcal{G}}^t = \gamma_{\mathcal{G}}^0 \times \text{type}(\mathcal{G}, t)^2 \quad (7)$$

where $\text{type}(\mathcal{G}, t)$ is the number of different kinds of features seen in that group (e.g., DOG and CAT are two different feature types from the same group) up through time t. In this way, the P_t^u of a feature that occurs in a group with more observed feature types is higher than the P_t^u in a group with fewer observed types.

Thus both the type frequency of features in $\mathcal{G}$ and their token frequency of co-occurrence with word w_j will influence – the first positively and the second negatively – how readily w_j can refer to objects with previously unseen features from $\mathcal{G}$.

5 Experimental Set-up

We model X&T's behavioural experiments with our computational word learner as extended above.[3] Following X&T, we use a three-tiered category hierarchy, and the four training conditions and assessment of three types of test matches as described in Figure 1.

Training the model. In each condition, the model processes a sequence of 3 utterance-scene pairs, and updates $P_t(f_i|w_j)$ after each pair using Eqns. (5) and (6). The utterance-scene pair in each trial consists of the novel word coupled with the scene representation of a training object from the

[2] Each feature group forms a Categorical distribution with $k_{\mathcal{G}}$ categories ($\text{Cat}(\theta_1, ..., \theta_{k_{\mathcal{G}}})$), where the θ_i are drawn from a prior Dirichlet distribution $\text{Dir}(\gamma_1, ..., \gamma_{k_{\mathcal{G}}})$ at time $t = 0$, and the θ_i are updated at time t to be the expected value of the posterior Dirichlet distribution, given in Eqn. (5) or Eqn. (6).

[3] Link to our code/data: `github.com/eringrant/word_learning/tree/hypothesis-space`.

category hierarchy. The object's scene representation is given as a set of four features, each taken from one of four feature groups: one feature corresponding to each of the subordinate, basic, and superordinate levels of the hierarchy, and a unique "instance" feature, as shown in Table 1. (The "instance" feature is added to simulate the variations in the different objects of the same subordinate category in the X&T experiments.)

1	**U:**	{ *fep* }
	S:	{ INSTANCE$_1$, DALMATIAN, DOG, ANIMAL }
2	**U:**	{ *fep* }
	S:	{ INSTANCE$_2$, TABBY, CAT, ANIMAL }
3	**U:**	{ *fep* }
	S:	{ INSTANCE$_3$, POLAR BEAR, BEAR, ANIMAL }

Table 1: An example of a sequence of utterance-scene pair trials in the 3-super. condition.

Testing the model. After training on a novel word, in order to assess its level of generalization within the category hierarchy, we compare the model's learned meaning of the word to test objects that constitute various types of matches to the training conditions: *i.e.*, subordinate matches, basic-level matches, and superordinate matches. Table 2 gives an example of each type of match:

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

Table 2: An example of each level match from the test objects, given the training condition in Table 1.

To assess whether the model generalizes the learned meaning of a word w to the various types of test matches, we first consider the probability of a test object Y at time t given the learned meaning of w:

$$P_t(Y|w) = \prod_{y_i \in Y} P_t(y_i|w) \qquad (8)$$

where y_i are the features in Y, and $P_t(y_i|w)$ is calculated using Eqn. (5) for features y_i observed with w during training, and using Eqn. (6) for y_i not observed with w. (Recall that Eqn. (5) reduces to Eqn. (6) when a feature has not been seen with the word.) From $P_t(Y|w)$, we subtract the predictive probability of the test object before the model has observed any data, $P_0(Y|w)$, which gives us its increase in preference attributable to the word learning trials.[4]

Calculating $P_t(Y|w) - P_0(Y|w)$ is informative about one test object, but we need to measure generalization of the learned word to all the objects of a certain type of match – *i.e.*, subordinate, basic-level, or superordinate. We formulate the probability of generalization to a type of test match as the relative average increase in preference for test items of that type of match, using the Shepard-Luce choice rule (Shepard, 1958; Luce, 1959):

$$P_{gen}(m|w) = \frac{\text{avg}_{Y \in m}\,[P_t(Y|w)\text{-}P_0(Y|w)]}{\sum\limits_{m'} \text{avg}_{Y' \in m'}\,[P_t(Y'|w)\text{-}P_0(Y'|w)]}$$

where m is the set of test objects at a certain level of match, and m' ranges over subordinate matches, basic-level matches, and superordinate matches.

Using $P_{gen}(m|w)$ to communicate our models results has the advantage of using the learned word meanings in a very direct way to assess the preference for the various types of test matches in the X&T experiments. However, the disadvantage is that this measure is not directly comparable to the reported figures from the human data, which are the percentage of test objects selected of a particular type of match. Hence, in presenting our results below, we focus on the general patterns of preferences indicated by the different measures.

Parameters. To model children, whom we assume to have no bias towards generalization to specific category levels, we equate all parameters $k_{\mathcal{G}}$ and $\gamma_{\mathcal{G}}$ across all feature groups, reflecting that all category levels are treated equivalently. Here we use values of $k_{\mathcal{G}} = 100$ and initial values of $\gamma_{\mathcal{G}} = 0.5$ for all $\mathcal{G}$ as the "child" parameter settings.[5] In contrast, we assume that adults, through word learning experience, have accumulated biases that reflect observed differences in feature groups. More specifically, we assume that the probability of observing a new feature for a group $\mathcal{G}$ depends on the degree of specificity of that group: That is, over time, it is less likely to observe a completely new kind of animal, e.g., than a new breed of dog. We simulate these biases by us-

[4] $P_0(Y|w) = \prod_{\mathcal{G}} \frac{1}{k_{\mathcal{G}}}$ is the prior probability of any object instance, given parameters drawn from the Dirichlet prior, because Eqn. (6) yields the value $\frac{1}{k_{\mathcal{G}}}$ when all assoc$_t$ scores are 0 – *i.e.*, no features from $\mathcal{G}$ have been observed with the word.

[5] To determine the parameters for the "child" learner, we examined a number of settings with equal parameter values for all the feature groups, and observed similar results in these settings. (We did not perform an exhaustive search over the parameter space.)

ing various values for the parameter $\gamma_{\mathcal{G}}$, which determines the prior probability of a word being observed with new (previously unseen) features in $\mathcal{G}$ (cf. Eqn. (6)). We assume that the expected number of features ($k_{\mathcal{G}}$) is the same across groups. We perform a non-exhaustive search on the parameter space of $\gamma_{\mathcal{G}}$ to select a set of values that yield the patterns of X&T's adult experiments. The "adult" parameter values are given in Table 3:[6]

$\gamma_{\text{inst}} = 1.2$	$\gamma_{\text{subord}} = 1.0$	$\gamma_{\text{basic}} = 0.5$	$\gamma_{\text{super}} = 0.2$
$k_{\text{inst}} = 100$	$k_{\text{subord}} = 100$	$k_{\text{basic}} = 100$	$k_{\text{super}} = 100$

Table 3: "Adult" parameter settings.

6 Experimental Results

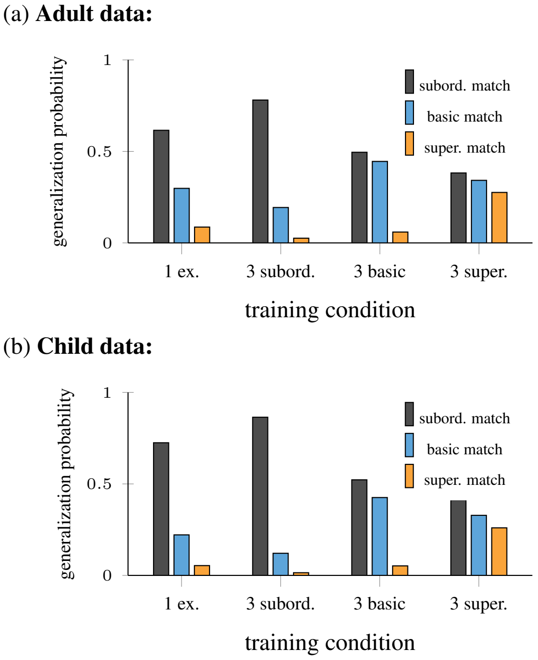
We present results of the model using both child settings (Figure 3b) and adult settings (Figure 3a). Recall that these values do not correspond to the percentages reported in the human data; to evaluate the patterns of generalization, we look at the relative preference for the various types of test match. Note also that since the generalization probabilities sum to 1.0 within each of the 4 training conditions, we can only compare the *pattern* of generalization across conditions (and not the actual value of the probabilities).

We discuss each of the child and adult sets of results in detail below.

6.1 The Child Learner

Recall that in the simulations of a child, we use equal values across all feature groups for the $k_{\mathcal{G}}$ and initial $\gamma_{\mathcal{G}}$ parameter settings, to reflect that the learner has no bias towards generalization to specific category levels.

Looking at the results in Figure 3b, we can see that the child learner generally replicates the patterns of results observed in X&T's experiment on children (cf. Figure 2b). Given multiple training items (the 3-subord., 3-basic, and 3-super. conditions), the model, like children, generalizes to the lowest level category in the hierarchy that is consistent with the training items, roughly equally preferring items from that category or lower, with slight preference for the lower categories. In contrast, after seeing a single training example (the 1-ex. condition), the model shows some tendency to generalize to the basic-level, demonstrating a small but notable basic-level bias — e.g., the tendency to consider the word as referring to any dogs (but less so to other animals) after seeing just a single example of a particular kind of dog. As in children, the difference in the model between the preference for subordinate vs. basic-level matches is much smaller when trained on 1 instance as opposed to 3 subordinates. (In Figure 3b, compare the difference between the 1st bar [subord. match] and 2nd bar [basic match] of the 1-ex. training condition to that of the 3-subord. training condition.)


(a) **Adult data:**

(b) **Child data:**

Figure 3: Our model data for (a) adults and (b) children. Each bar is the probability of a type of test match: *i.e.*, subord(inate), basic(-level), or super(ordinate).

Interestingly, our child learner exhibits the observed basic-level bias in the absence of any difference in the model in how it treats different category levels. The observed pattern arises from a type/token frequency interaction of the kind often noted to influence generalization of linguistic categories (*e.g.*, Bybee, 1985; Croft and Cruse, 2004): here, the interaction between the token frequency of word–feature pairs in the input and the type frequency of different features within a group of dependent features. For example, having seen 3 types of animals ("3 super." condition), the model can readily accommodate that *fep* refers to another kind of animal, in contrast to the "3 basic" condition, where it has seen the same number of tokens but only a single feature type from the feature

[6] For a certain range of such parameter settings – i.e., with gradually decreasing $\gamma_{\mathcal{G}}$, which determines the prior probability of a word being observed with new (previously unseen) features in $\mathcal{G}$ (cf. Eqn. (6)). for feature groups at successively higher levels in the hierarchy — the model produces similar results to the presented adult learner.

group at that level (3 dogs). We can also clearly see the inverse impact of token frequencies on generalization: the more examples of a single subordinate type are seen, the less the model accepts that *fep* refers to a different kind of subordinate (the "3-subord." vs. "1-ex." conditions). That is, with only 1 token of DALMATIAN, the model can generalize to other types of dogs more readily than when it has seen 3 tokens of DALMATIAN.

In general, interactions between the type and token frequencies of the different feature groups interact to yield the observed patterns in the model. These results indicate that properties of the input data coupled with the model's handling of feature groups can account for children's word generalization behaviour, without the need for an explicit basic-level bias.

6.2 The Adult Learner

Adult participants in X&T exhibited a stronger tendency than children to generalize to the basic-level category, especially after seeing a single exemplar. We explore whether the model can simulate an adult learner as well. As discussed in Section 5, by varying the parameters $\gamma_{\mathcal{G}}$, we can incorporate biases towards different category levels that we assume an adult has learned. More specifically, we set $\gamma_{\mathcal{G}}$ to successively larger values for more specific feature groups $\mathcal{G}$, to ensure successively greater generalization in lower levels of the hierarchy (see Table 3). As shown in Figure 3a, our model (using such settings of the parameters) replicates the patterns of X&T's adult experiments (cf. Figure 2a), including a stronger basic-level bias than that shown by children. That is, in the 1-ex. and 3-subord. conditions, the difference between the 1st bar [subord. match] and 2nd bar [basic match] is smaller for the adult settings of the model (Figure 3a) than for the child settings (Figure 3b), mimicking the stronger basic-level bias found in adults.

6.3 Variations in Basic-level Generalization

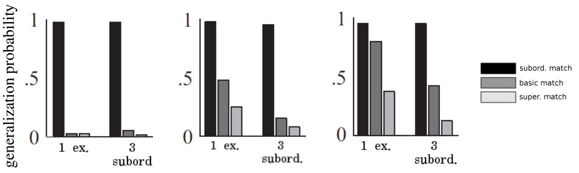
Research shows that people's degree of basic-level generalization depends on the overall category of the objects. Specifically, Abbott et al. (2012) perform the same set of experiments as X&T on adults, exploring three additional superordinate categories (clothing, containers, and seats). Their results are shown in Figure 4; for space reasons, we focus here on the training conditions with 1-example or 3-subordinates, which are the locus of the basic-level effect. The results show that people exhibit no basic-level generalization for containers, moderate generalization for clothing, and strong generalization for seats (compare Figures 4a, 4b and 4c).

Interestingly, the computational experiments of Abbott et al. (2012) also reveal that the Bayesian model of X&T mimics varying levels of basic-level generalization in the 1-example cases, but does not capture the differences that people exhibit across the categories in the 3-subordinate condition (compare "3 subord." in Figures 4 and 5): unlike people, here the X&T model does not exhibit basic-level generalization for any of the categories.

Abbott et al. (2012) note that a domain like containers may not follow a "natural taxonomy" in having a clear basic-level category. This suggestion is compatible with our view that a basic-level bias arises in response to the particular pattern of co-occurrence of features across the category hierarchy. We looked more closely at the training stimuli of their experiment, and observe that the examples of the category "containers" (with the least basic-level generalization) vary greatly, while those of "clothing" and "seats" are less differentiated. Examples from "containers" include a cigar box, trash can, and mailbox, whereas "seats" are restricted to different types of chair (such as a dining chair and an armchair; see Table 1 in Abbott et al. (2012)).

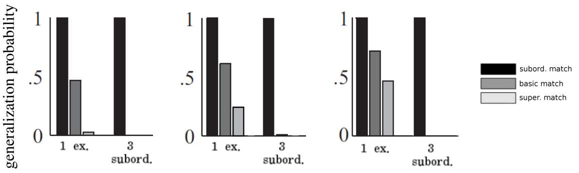
Based on this observation, we hypothesize that people generalize less to a basic-level category when their mental representations for that category's instances have more distinguishing features. Specifically, we assume that people differentiate the given instances of the category "containers" more than those for "clothing" and "seats". We model this difference in the granularity of representations by varying the number of feature groups used in representing an object. Recall that in our earlier experiments, each object was represented as a set of features drawn from 4 different feature groups. We take this representation as the least fine-grained representation and use it for the category "seats". We assume that the objects from the categories "clothing" and "containers" (that exhibit less basic-level generalization) are represented with more feature groups (8 and 12, respectively).

Figure 6 shows the results of running our model



(a) Containers (b) Clothing (c) Seats

Figure 4: Abbott et al. (2012) subject response data.



(a) Containers (b) Clothing (c) Seats

Figure 5: Abbott et al. (2012) model data.

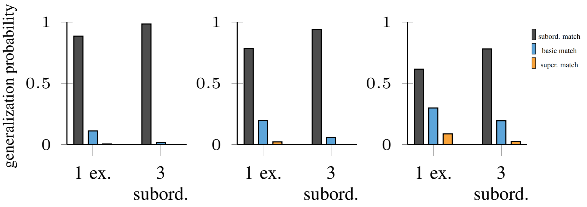
on these three categories using the "adult" parameter settings. As expected, the generalization to the basic-level category is high for the least distinguished category "seats", moderate for the category "clothing", and low for the most distinguished category "containers".[7]

Our results suggest that the observed variation across categories in basic-level generalization could arise from differences in the granularity of representations of categories. This is particularly interesting since the model of X&T, despite encoding an elaborated taxonomy, does not capture the observed behaviour across all training conditions.

7 Conclusions

A key challenge faced by children in vocabulary acquisition is learning which of many possible meanings is appropriate for a word, based largely on ambiguous situational evidence. One aspect of this is what we term the "word generalization" problem, which refers to how children associate a word such as *dog* with a meaning at the appropriate category level in a taxonomy of objects, such as Dalmatians, dogs, or animals.

We present extensions to a cross-situational learner that enable the first computational study of word generalization that is integrated within a word learning model. The model mimics child behavior found by Xu and Tenenbaum (2007): it shows a "basic-level" bias – a preference for word meanings that refer to basic-level objects (like dogs), in contrast to higher-level (animals) or lower-level (Dalmatians) categories – and does so under parameter settings that treat all levels of category the same in the model (*i.e.*, with no built-in basic-level bias). Other (unequal) parameter settings, which could reflect learned knowledge leading to differential treatment of categories, yield behavior that mimics that of adults, who show a stronger basic-level bias. Moreover, similarly to people (Abbott et al., 2012), our model exhibits variations in generalization to the basic level for different types of objects, a behavior that the model of Xu and Tenenbaum (2007) does not fully replicate.



(a) Containers (b) Clothing (c) Seats

Figure 6: Data from our model.

Overall, the results of our model arise from the interaction of type and token frequencies of features in the input data, which impact the model's evolving word representations. This mechanism in the model captures the type-token influence often observed to underlie people's generalization of linguistic categories – *i.e.*, their linguistic productivity (*e.g.*, Bybee, 1985; Croft and Cruse, 2004).

One shortcoming of the current model is its built-in ability to "detect" in the input that DOG and CAT features are more specific than ANIMAL features. The next step is to consider how the model might learn these relationships from its evolving knowledge of co-occurring features.

Finally, a similar problem to that of word generalization in humans arises in computational linguistics: how to appropriately generalize a set of concepts to an overarching concept that subsumes the set. For example, this problem underlies one way to determine the selectional preferences of a verb: extract the set of nouns that occur as objects of the verb, map them to the concept nodes in a hierarchy such as WordNet, and then determine the best overarching WordNet category for capturing the salient properties of the object nouns overall (*e.g.*, Li and Abe, 1998; Clark and Weir, 2001). An interesting future direction is to explore how an extension of our work can be applied to such problems in computational linguistics.

[7] Similar results obtain using "child" parameter settings, but (as expected) the basic-level generalization is lower.

8 Acknowledgements

We would like to thank Jackie Chi Kit Cheung and the anonymous reviewers for their valuable feedback. We gratefully acknowledge the support of NSERC of Canada, and of an Ontario Graduate Scholarship to the first author.

References

Joshua T. Abbott, Joseph L. Austerweil, and Thomas L. Griffiths. 2012. Constructing a hypothesis space from the web for large-scale bayesian word learning. *Proceedings of the 34th Annual Conference of the Cognitive Science Society*.

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

Stephen Clark and David Weir. 2001. Class-based probability estimation using a semantic hierarchy. In *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies*, pages 1–8. Association for Computational Linguistics.

William Croft and Alan Cruse. 2004. *Cognitive linguistics*. Cambridge University Press.

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, Noah D. Goodman, and Joshua B. Tenenbaum. 2007. A Bayesian framework for cross-situational word-learning. In *NIPS'07*, volume 20.

Roberta M. Golinkoff, Carolyn B. Mervis, and Kathryn Hirsh-Pasek. 1994. Early object labels: The case for a developmental lexical principles framework. *Journal of child language*, 21(01):125–155.

George Kachergis, Chen Yu, and Richard M. Shiffrin. 2012. An associative model of adaptive inference for learning word–referent mappings. *Psychonomic Bulletin & Review*, pages 1–8.

Hang Li and Naoki Abe. 1998. Word clustering and disambiguation based on co-occurrence data. In *Proceedings of the 17th International Conference on Computational Linguistics - Volume 2*, COLING '98, pages 749–755. Association for Computational Linguistics.

Robert D. Luce. 1959. *Individual choice behaviour*. Wiley, NY.

Ellen M. Markman. 1991. *Categorization and naming in children: Problems of induction*. Mit Press.

David Marr. 1982. Vision: A computational investigation into the human representation and processing of visual information.

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

Steven Pinker. 1989. *Learnability and Cognition: The acquisition of Argument Structure*. Cambridge, Mass.: MIT Press.

Willard Van Orman Quine. 1960. *Word and Object*. MIT Press.

Eleanor Rosch, Carolyn B. Mervis, Wayne D. Gray, David M, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive Psychology*.

Eleanor Rosch, 1973. *On the Internal Structure of Perceptual and Semantic Categories*, pages 111–144. Academic Press, New York, NY.

Roger N. Shepard. 1958. Stimulus and response generalization: Tests of a model relating generalization to distance in psychological space. *Journal of Experimental Psychology*, 55(6):509.

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

Linda B. Smith and Chen Yu. 2008. Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106(3):1558–1568.

Fei Xu and Joshua B. Tenenbaum. 2007. Word learning as Bayesian inference. *Psychological Review*, 114(2):245–272.

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

Daniel Yurovsky, Damian C. Fricker, Chen Yu, and Linda B. Smith. 2014. The role of partial knowledge in statistical word learning. *Psychonomic bulletin & review*, 21(1):1–22.