Word Learning as Category Formation[*]

Spencer Caplan
University of Pennsylvania
3401 Walnut Street
Philadelphia, PA 19104
spcaplan@sas.upenn.edu

Abstract

A fundamental question in word learning is how, given only evidence about what objects a word has previously referred to, children are able to generalize the total class (Smith, 1979; Xu and Tenenbaum, 2007). E.g. how a child ends up knowing that *'poodle'* only picks out a specific subset of dogs rather than the whole class and vice versa. Here we present a computational model of word learning which accounts for a wide range of previously conflicting experimental findings.

1 Generalization in Word Learning

Words are invitations to form categories (Waxman and Markow, 1995). It is striking that infants interpret a word as selecting members of some kind, rather than simply naming an individual referent. Put clearly in (Waxman, 2003): *"Novel words invite infants to assemble together objects into categories that would otherwise (without linguistic context) be perceived as disparate and distinct."*. This is not to say that categorization functions solely through the learning of words, but rather that the process receives a notable boost from linguistic support.

If hearing a novel word like *'fep'* prompts the learner to create a category, we would like to know what knowledge ends up encoded by that process and how. Once a child has seen that 'poodle' can refer to whatever instances of poodles they were exposed to, how does he know that 'poodle' can refer to all (and only) items in the real class of poodles? This is in contrast to both *failing to generalize* sufficiently, e.g. erroneously positing that the word only refers to their pet, as well as *overgeneralizing* to the set of all dogs.

The Naïve Generalization Model (NGM) presented in this paper offers an explanation of word learning phenomena grounded in category formation (Smith and Medin, 1981) and learning theory (Gallistel, 1990). The NGM captures relevant experimental findings (Xu and Tenenbaum, 2007; Spencer et al., 2011) which cannot straight-forwardly be accounted for on a Bayesian inference theory (Xu and Tenenbaum, 2007).

2 Experimental Findings in Generalization

The most popular experimental setup for investigating the mechanisms behind learners' behavior in word generalization tasks stems from (Xu and Tenenbaum, 2007). In an ostensive labeling task, participants' are asked to extend the category presented in a training set by selecting matching items from a miniature 'test-world'.

This setup consists of photographs of objects distributed across different broad categories or genres (animals, vegetables, and vehicles) to be used as stimuli. For any particular item, we operationally define a 'basic-level' term (Markman, 1990) as the label which would most likely be given to it in isolation (e.g. a dog) . In relation to the basic-level term, that same item might also be referred to using a 'subordinate-category label' such as 'poodle' or a 'superordinate-category label' such as animal.

On each trial, participants are presented with one or several *training* objects below the test grid along

[*] I would like to thank Charles Yang for consistent guidance and advice throughout this project.

Proceedings of the Society for Computation in Linguistics (SCiL) 2018, pages 174-177.
Salt Lake City, Utah, January 4-7, 2018

with an accompanying nonce word-label. For instance, a participant may be shown a picture of a dalmatian with the label 'fep' and asked to pick out all the other 'feps' from the simultaneously displayed test grid. While popularized by (Xu and Tenenbaum, 2007), this paradigm has been replicated and extended numerous times.

The broad experimental results are as follows: When only a single object is presented with a label, then subjects most commonly generalize to the basic-level category (e.g. selected all *dogs* rather than only *dalmations* given that the single training item was a dalmatian) (Xu and Tenenbaum, 2007; Spencer et al., 2011). When multiple training examples are presented simultaneously, then generalization is made narrower (e.g. selecting only dalmatians). This 'suspicious coincidence effect' has been presented in favor of global evaluation in word learning. Yet, this faces empirical challenges from conditions under which the 'effect' is not obtained. In particular, when those same training items are given a single label but displayed to participants in sequence rather than all at once then this effect disappears (Spencer et al., 2011). i.e. all dogs are chosen rather than only dalmatians. Models of Bayesian inference do not predict this effect of presentation style which is well-captured under the NGM.

3 Bayesian Inference Models

Existing models of category generalization in word learning have been built on hypothesis comparison and indirect negative evidence ((Xu and Tenenbaum, 2007) and subsequent work). Multiple innate hypotheses compete based on the relative probability that each hypothesis would be generated by the attested input data.

The *'suspicious coincidence effect'* (SCE) is that, when presented with a single exemplar, learners tend to assume a 'basic-level' of generalization (e.g. DOG) whereas when multiple like exemplars are presented simultaneously then only a narrow generalization is most common (such as POODLE).

This, however, is inadequate to capture the experimental facts from (Spencer et al., 2011). When participants are presented the same stimuli in sequence (including a short temporal gap between items) rather than simultaneously the SCE disappears. The Bayesian model presented in (Xu and Tenenbaum, 2007) does not predict this effect of presentation style.

In the next section, we introduce the Naïve Generalization Model (NGM, which implements a system of word learning as category formation. Learners extract properties of objects and store a mental record of them. Grounded in literature on category formation (Smith and Medin, 1981), inductive reasoning (Lawson, 2017) and learning theory (Gallistel, 1990), these mental representations serve as the basis of word meanings and generalization. The NGM captures a range of experimental findings with respect to word learning, including the effect of presentation style that is unpredicted by models of hypothesis evaluation or Bayesian inference.

4 Naïve Generalization Model

Counter to the Bayesian inference account, we argue that word learning is a dynamical process in which hypothesized representations are generated and only locally revised (as needed) based on input data. On this account, not all plausible hypotheses are simultaneously available. Meanings are built incrementally; any evaluation metric functions only over what is generated from input by the learner. As this does not necessarily maximize global probability of the output vocabulary, we term this model the Naïve Generalization Model (NGM).

The difference in performance between parallel and sequential experimental presentations of the same stimuli is driven by creating a category from multiple examples rather than comparing an extant category to new data. Starting from the assumption that encountering a word is an invitation to form a category, the first instance of a word is qualitatively different from other instances. Upon initial occurrence there are no prior hypothesized meanings to compare against, and so a representation must be created. At all future instances, however, the learner must ask decide whether a new token is consistent with the current mental representation or not. If the prior hypothesized representation is inconsistent with current input, then an alternative hypothesis may be created.

4.1 Features

Our implementation follows the classic literature on categories (Smith, 1979; Smith and Medin, 1981) by representing concepts as salient *features*. What we call 'features' are simply properties that hold for some item. While any two properties may be equally true of an object, some properties are more salient than others to an observer. To simulate the degree to which a property is noticed by a learner, we model two normal distributions over salience differing only in mean; one for features consistent with a basic-level bias and one for all other features.

When a learner encounters a new word, the model samples from the appropriate salience distribution for each feature present. The result is a mental representation as a gradient vector of features. The learner iterates over the items displayed (if more than one present) and each feature present in the real world will be stored in mental representation at a proportion relative to that feature's salience.

A representation R is computed for a label w based on an example set of training items T by sampling all features $\forall f$ with salience $S(f)$. This is adapted from classic approaches to category membership calculation (Smith & Medin 1981).

$$R_w = \sum_{t \in T} \forall f \in t_p, S(f) \qquad (1)$$

t_p is the set of features (or *properties*) of the item t. $S(f)$ is the *salience function* for a feature f which returns a value samples from the normal distribution with mean μ determined by the hierarchical level of f.

Multiple (simultaneous) exposures for a label causes entrenchment (Lawson, 2017). We sum the values of each present feature (until reaching a ceiling condition). This is in line with previous featural implementations of categories, e.g. (Kruschke, 2008).

When items are presented in sequence, only salient features are encoded in representation. This 'sparse' representation corresponds to a broad category generalization. Simultaneous presentation, on the other hand, entrenches shared features between present items (Gentner and Namy, 1999; Rescorla, 1980). This entrenchment of otherwise non-salient features leads the representations to correspond to more specific, narrow categories.

4.2 Computing Distances

A standard distance calculation is made between any new objects and existent mental representations (Smith and Medin, 1981). The comparison of that value to a fixed parameter threshold determines category membership. Distance is then calculated between a test item and a mental representation for a label by summing the difference of any feature present in the mental representation that is missing in the test object under consideration. However, there is no cost incurred for features which are present in a test item but are missing in the mental representation of a class. For example, every object in the world is going to be perceived as having some color value, but that color plays no role in these items' membership in various natural classes.

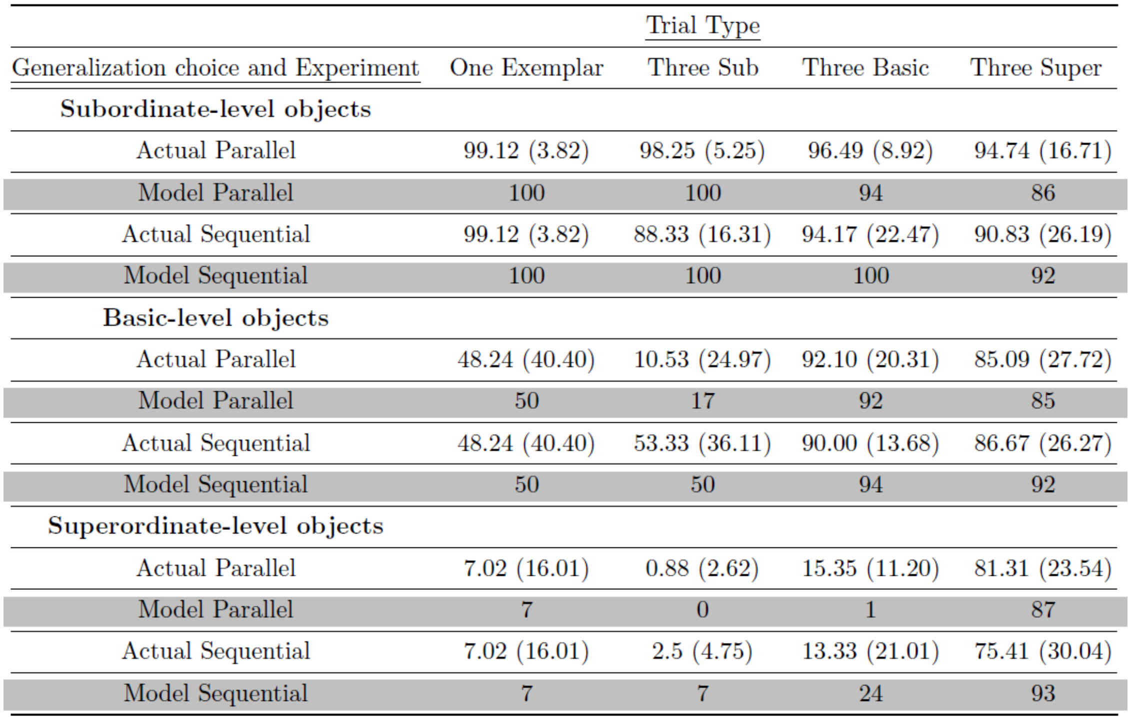
5 Results

Tuning and testing of the computational model was performed by feeding in the same input data from published experiments and scoring the resultant output like the empirical findings. Parameter tuning was performed by running the model over the two basic experimental conditions from (Xu and Tenenbaum, 2007) – training over a single exemplar and training over three basic-level matches in parallel. Testing was then performed on all experimental conditions from (Spencer et al., 2011) varying the hierarchical organization and presentation style of the input.

This model captures the broad range of experimental findings in category generalization as shown in (Table 1). The mean divergence per trial between the experimental data and the output of the model is 5.67%. 96% of trial configurations were within a single standard deviation of the empirical finding.

6 Discussion

The Naïve Generalization Model presented here is able to account for a range of experimental findings in word learning. This includes the 'suspicious coincidence' effect from (Xu and Tenenbaum, 2007). Importantly though, the NGM captures effects of presentation style (Spencer et al., 2011) which are not predicted on a model of Bayesian Inference. The



	Trial Type			
Generalization choice and Experiment	One Exemplar	Three Sub	Three Basic	Three Super
Subordinate-level objects				
Actual Parallel	99.12 (3.82)	98.25 (5.25)	96.49 (8.92)	94.74 (16.71)
Model Parallel	100	100	94	86
Actual Sequential	99.12 (3.82)	88.33 (16.31)	94.17 (22.47)	90.83 (26.19)
Model Sequential	100	100	100	92
Basic-level objects				
Actual Parallel	48.24 (40.40)	10.53 (24.97)	92.10 (20.31)	85.09 (27.72)
Model Parallel	50	17	92	85
Actual Sequential	48.24 (40.40)	53.33 (36.11)	90.00 (13.68)	86.67 (26.27)
Model Sequential	50	50	94	92
Superordinate-level objects				
Actual Parallel	7.02 (16.01)	0.88 (2.62)	15.35 (11.20)	81.31 (23.54)
Model Parallel	7	0	1	87
Actual Sequential	7.02 (16.01)	2.5 (4.75)	13.33 (21.01)	75.41 (30.04)
Model Sequential	7	7	24	93

Figure 1: Table showing results comparison between experiments run in (Spencer et al. 2011) and output of the present model.

local computations of the NGM are consistent with other types of word learning models such as Pursuit (Stevens et al., 2016) with respect to referent-mapping. Word learning is a dynamical process in which hypothesized representations are generated and only locally revised (as needed) based on subsequent input.

References

Charles R Gallistel. 1990. *The organization of learning*. MIT press Cambridge, MA.

Dedre Gentner and Laura L Namy. 1999. Comparison in the development of categories. *Cognitive development*, 14(4):487–513.

John K Kruschke. 2008. Models of categorization. *The Cambridge handbook of computational psychology*, pages 267–301.

Chris A Lawson. 2017. The influence of task dynamics on inductive generalizations: How sequential and simultaneous presentation of evidence impact the strength and scope of property projections. *Journal of Cognition and Development*, 36(1).

Ellen M Markman. 1990. Constraints children place on word meanings. *Cognitive Science*, 14(1):57–77.

Robert A Rescorla. 1980. Simultaneous and successive associations in sensory preconditioning. *Journal of Experimental Psychology: Animal Behavior Processes*, 6(3):207.

Edward E Smith and Douglas L Medin. 1981. *Categories and concepts*. Harvard University Press Cambridge, MA.

Linda B Smith. 1979. Perceptual development and category generalization. *Child Development*, pages 705–715.

John P Spencer, Sammy Perone, Linda B Smith, and Larissa K Samuelson. 2011. Learning words in space and time probing the mechanisms behind the suspicious-coincidence effect. *Psychological science*, 22(8):1049–1057.

Jon Scott Stevens, Lila R Gleitman, John C Trueswell, and Charles Yang. 2016. The pursuit of word meanings. *Cognitive Science*.

Sandra R Waxman and Dana B Markow. 1995. Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, 29(3):257–302.

Sandra R Waxman. 2003. Links between object categorization and naming. *Early category and concept development: Making sense of the blooming, buzzing confusion*, pages 213–241.

Fei Xu and Joshua B Tenenbaum. 2007. Word learning as bayesian inference. *Psychological review*, 114(2):245.