# Word Learning as Category Formation

**Spencer Caplan (spcaplan@sas.upenn.edu)**
Department of Linguistics, University of Pennsylvania

**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 & Medin, 1981; Xu & 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. The Naïve Generalization Model (NGM) presented in this paper offers an explanation of word learning phenomena grounded in category formation (Smith & Medin, 1981) The NGM captures a range of relevant experimental findings (Xu & Tenenbaum, 2007; Spencer, Perone, Smith, & Samuelson, 2011), including those which are in conflict with a Bayesian inference theory (Xu & Tenenbaum, 2007).

**Keywords:** Language Acquisition; Word Learning; Cognitive Modeling; Computational Linguistics

## Word Learning and Generalization

A crucial facet of language acquisition is the development of the lexicon. Language learners need to infer the set of vocabulary items belonging to their particular language based on the patterns of speech produced around them. While much previous work has focused on referential ambiguity resolution (Yu, 2008), the issue of generalization is less well understood. Consider a simple environment for learning the word 'dog': A child hears an adult speaker refer to their pet as /dɔg/. While from the prospective of referential ambiguity the situation is clear, the space of possible meanings for the phonetic label /dɔg/ is still quite large. The word may be the particular pets name, or it could mean pets generally. It might pick out the set of (all and only) dogs.

Experimental work on the acquisition of word meanings has shown that language learners approach the problem with strong biases with respect to referents and concepts that severely limit this potential search space (Markman, 1990; Landau, Smith, & Jones, 1988; Snedeker, Gleitman, et al., 2004; Gillette, Gleitman, Gleitman, & Lederer, 1999). Yet, despite the help of limited search from such biases, word learning is still able to function even in an impoverished "base condition" in which the only direct source of information is the set of referents. A helpful conceptualization of this is that words are invitations to form categories (Waxman & Markow, 1995). It is striking that infants interpret a word as selecting members of some kind, rather than simply naming an individual referent.

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/she 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* that the word selects the set of all dogs.

A popular previous model of word learning functions via Bayesian inference (Xu & Tenenbaum, 2007). While some support for this paradigm is offered by the 'suspicious coincidence effect' (SCE)—that an increase in sample-size corresponds to more narrow word meanings—conflicting experimental findings (Spencer et al., 2011) contradict predictions made by Xu and Tenenbaum (2007). Notably the SCE disappears depending on the temporal presentation style in which stimuli are given to participants (Spencer et al., 2011). While this gap in SCE under such conditions was first highlighted by Spencer et al. (2011), they did not provide a computational model to account for the finding. Nonetheless this difference in performance between sequential and parallel presentation is in fact consistent across a range of related studies (Gelman & Markman, 1986; Lawson, 2017). This variance across presentation style should thus be viewed as an important cognitive effect worthy of explanation rather than simply a nuisance data point to capture.

The Naïve Generalization Model (NGM) presented in this paper offers an explanation of word learning phenomena grounded in category formation (Smith & Medin, 1981). The model explains the mechanism by which hearing novel words invites a learner to create a new category from component 'features'. Once a representation for a novel word has been created, the learner is able to evaluate subsequent labeled objects with respect to this hypothesized meaning. Evaluation of meanings is 'naïve' in the sense that it does not optimize for any particular global value, with both creation and evaluations of word meanings functioning locally. The NGM is consistent with, and offers an explanation of, a range of previously conflicting experimental findings in word learning and generalization (Xu & Tenenbaum, 2007; Spencer et al., 2011; Lawson, 2017).

Below we review the Bayesian account of word learning and discuss relevant experimental findings which cannot straight-forwardly be accounted for on such a theory. Next, we describe the internal mechanisms of the NGM. This includes the representation and computation of features for word learning. We detail the performance of the NGM on two evaluation schemes, one qualitative and quantitative, with respect to modeling experimental data.

## Previous models and experimental findings

Existing models of category generalization in word learning have been built on hypothesis comparison and global optimization (Xu and Tenenbaum (2007) and subsequent work). A large set of hypotheses compete based on the relative probability that each hypothesis would be generate the attested

input data. The task is then re-framed as choosing how words map onto those concepts by ruling out impossible or less probable hypotheses until a consistent hypothesis is reached. The now seminal implementation of this is a Bayesian inference model from Xu and Tenenbaum (2007). Given some set of attested referents, a Bayesian learner evaluates all hypotheses (h) for candidate word meanings according to Bayes rule, by computing their posterior probabilities (the likelihood of each hypothesis given the input data $p(h|referents)$), proportional to the product of prior probabilities $p(h)$ and likelihoods $p(referents|h)$.

This family of models is *global* in two senses. First, calculations of hypothesis-fit to the data are taken over all input received. The learner would need to track some record of every attested exemplar in order to compute probabilities over them. The second global notion is that all alternative hypotheses are also calculated for goodness-of-fit to the input data. This allows for global comparison not only between total input and some temporary hypothesis but between all hypotheses themselves.

This makes an intuitive prediction dubbed the 'suspicious coincidence effect', that if a learner is exposed to some new word 'fep' (adapted from (Xu & Tenenbaum, 2007, p.249)): "It would be quite surprising to observe only Dalmatians called feps if in fact the word referred to all dogs and if the first four examples were a random sample of feps in the world. This intuition can be captured by a Bayesian inference mechanism that scores alternative hypotheses about a words meaning according to how well they predict the observed data, as well as how they fit with the learners prior expectations about natural meanings."

## Experimental Findings

Experimental support for this prediction is offered from Xu and Tenenbaum (2007). The task is that participants are interacting with an 'alien' puppet, ostensibly a monolingual speaker of 'alien puppet talk'. On each trial, participants are presented with one or several *training* objects below the test grid along with an accompanying monosyllabic 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' for the puppet from the simultaneously displayed test grid. The general findings in this paradigm are consistent across both child and adult participants.

The test grid consists of photographs of real objects distributed across three 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; Mervis, 1984; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) 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 more narrow 'subordinate-category' label such as 'poodle' or a broader 'superordinate-category' label such as 'animal'. Within each genre in the test grid, objects exist within these three hierarchical label levels. The set of 'test' objects is consistent across trials with only their position on the grid randomized.

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. selecting all *dogs* rather than only *dalmations* given that the single training item was a dalmatian) (Xu & Tenenbaum, 2007; Spencer et al., 2011). This is consistent with the robust effects of a 'basic-level' bias (Markman, 1990). When multiple training examples are presented simultaneously, then generalization is made narrower (e.g. selecting only dalmatians). This 'suspicious coincidence effect', that category narrowness is linked to the size of the training sample, has been presented in favor of the Bayesian model of word learning. Yet, Bayesian inference is not the only family of models which make such a prediction. What's more, global evaluation models face empirical challenges from conditions under which the 'effect' is not obtained. When the same training items are given a single label but displayed to participants in sequence rather than all at once, the SCE disappears (Spencer et al., 2011) (see Table 1 for a summary). i.e. all dogs are chosen rather than only dalmatians. The lack of a 'suspicious coincidence effect' under serial presentation runs counter to the predictions of Bayesian inference. We note that the temporal gap introduced between referents under sequential vs. simultaneous presentation is only a single second between item displays.

Category generalization is simply one of a wide range of cognitive tasks which exhibit a difference in outcome based on presentation style of exemplars. For instance, inductive category learning (Carvalho & Goldstone, 2015), visual pattern differentiation (Lappin & Bell, 1972), relational reasoning (Son, Smith, & Goldstone, 2011), property projection (Lawson, 2017), etc. all show important differences under sequential vs. simultaneous presentation of stimuli. Taken together, the effects of presentation style across this wide range of domains and studies should be understood as an important phenomenon whose root causes make up a core aspect of categorization models.

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 classic literature on category formation (Smith & Medin, 1981), these mental representations serve as the basis of word meanings and generalization. We describe the range of experimental findings captured by this model, including the effects of presentation style which are not accounted for under a model of Bayesian inference.

## Naïve Generalization Model

Word learning is to construct mental representations of words. While the Bayesian inference account of this process posits a global probability optimization over a large set of latent hypotheses, we instead argue that word learning is a dy-

| Trial Type | Level of Generalization | Example Meaning |
|---|---|---|
| Single Exemplar | Broad | "Dog" |
| Multiple Simultaneous Objects | Narrow | "Dalmatian" |
| Multiple Objects in Sequence | Broad | "Dog" |

Table 1: Basic generalization patterns from (Xu & Tenenbaum, 2007; Spencer et al., 2011). Both the size of the training set as well as the temporal manner of presentation have notable effects on the meanings posited by participants.

namical process. 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 Naive Generalization Model (NGM). The term 'naive' here is intended to highlight the lack of an explicit optimization function. Rather, empirical pattern in word learning arise from largely mechanistic means. The NGM is able to capture a range of previous unaccounted for empirical findings in meaning generalization with respect to word learning. This includes the basic-level bias, the 'suspicious coincidence effect' that multiple simultaneous exposures to labeled training instances narrows hypothesized meanings, as well as the effects of presentation style which seemingly *block* the 'suspicious coincidence effect'. On the NGM, word learning is fundamentally a *local* mechanism by which mental representations of words are *constructed* rather than strictly evaluated.

The generalization model does not function in isolation. The NGM is embedded within a larger understanding of word learning and is consistent with previous work regarding other stages in learning required for vocabulary acquisition. Notably this includes the mechanisms behind referent mapping posited in (Stevens, Gleitman, Trueswell, & Yang, 2016) The contribution of the NGM is to explain the way in which representations of meaning are created, updated, and maintained.

## Features

Our implementation follows the classic literature on categories (Smith & Medin, 1981) by representing concepts as salient *features*. What we call 'features' are simply properties that hold for some item. While any two properties will be equally true of an object, in the sense that they are formal operators, it should be clear intuitively that some properties are more salient than others. Consider the number 73. It is probably easier to determine that 73 is odd than it is to determine that 73 is a prime; it's not that its prime-ness is less valid than its being odd, rather it is simply a matter of salience (i.e. how noticeable it is to an average person quickly).

To simulate the degree to which a property is noticed by a learner, we model two normal distributions over salience. These 'salience distributions' differ only in mean; one for features with elevated prominence (the driving force behind the basic-level bias) and one for all other features. Of course, the prototypical-ness of items within a class, or the salience of certain features depends on the class and the objects themselves. But this is simply a way of formally implementing the notion that some levels of generalization are privileged compared to others. It is of theoretical interest that the model functions with such an impoverished feature space. For instance, the features in use are 'flat'—without inherent hierarchical relation between them—from the perspective of the learner. Yet the combination of these 'flat' features results in hierarchically nested extensions for word meanings.

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 (Figure 1). Values are allowed to be any decimal between zero and one. The upper-bound of one is important because, conceptually, this corresponds to the feature being as present mentally as it is in the physical world. 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).

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

$t_p$ 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 $\mu$ determined by the hierarchical level of $f$.

While features for an object in the world are formal operators, the mental stored values for a given feature are gradient. 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): 'the simplest way [to learn associative strengths] is adding a constant increment to the weight whenever both its source and target node are simultaneously activated.'

### Computing distances

The NGM makes a standard distance calculation between any new objects and extant mental representations. 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 (Smith & Medin, 1981).

There is a distance penalty for any feature present in the mental representation that is missing in the test object under consideration. This value is in accordance with the represented featural salience. However, there is no cost incurred for features which are present in a test item which 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 any of various natural classes being learned here.

Mutual exclusivity is a powerful and well-established constraint in word learning (Markman, 1990). We formally implement a feature-level adaption of this in the NGM by allowing properties in conflict to block addition to a single mental representation.

### Learning by presentation style

When trained on a single exemplar, the experimental finding is that learners' generalization to basic-level items occurs a substantial proportion of the time. This is driven by the privileged status of certain features for generalization over others. When training objects are *initially* presented simultaneously, whether that is a single exemplar or many, then a hypothesis category needs to be formed in a single shot. Thus when they are co-present, the function which extracts features from a scene is able to essentially compare exemplars to exemplars. When features are activated multiple times, they undergo entrenchment —creating stronger links in mental representation (Smith & Medin, 1981). Properties which, when encountered in isolation, would not have a significant effect on stored meaning can, through this entrenchment, lead to more narrow-generalization. The NGM's mechanistic account of featural entrenchments thus makes the same predictions are Bayesian inference with respect to the 'suspicious coincidence effect' under parallel presentation.

When the same stimuli are presented in sequence rather than in parallel, learners' generalization occurs primarily at the basic-level rather than the subordinate level. Even though training objects are shown to learners multiple times, the learner only constructs an initial hypothesis only once. After the first exemplar has disappeared from view, the learner needs to construct some mental representation for the presented word. Once a mental representation exists, there is no onus to change it significantly so long as subsequent objects picked out by the word are congruent with what's stored. This process is analogous to *localist* models of referent learning such as *Pursuit* (Stevens et al., 2016). Learners select a single hypothesis and either stick with it if evidence is consistent, or move to a new hypothesis when faced with inconsistent evidence. When subsequent training instances appear, the original exemplar(s) have disappeared from view with only the generated category remaining. This means that learners are essentially comparing new exemplars to a category representation rather than directly comparing exemplars with each other. Since all of these trials concern levels of generalization, no new training item will disprove an over-generalized hypothesis. Therefore, learners will simply continue along with whatever initial hypothesis was created. Repeat exposures increase a learner's *confidence* in the hypothesized meaning rather than triggering any change in the word's internal contents. This continues until some 'convergence point' is reached and a semantic representation is more or less fixed. Such a convergence point is a required component of any model of word learning. The cause of the 'basic-level bias' on sequential presentation trials is the same as in the single-exemplar trails: certain types of features lead to privileged levels of generalization.

## Results

### Qualitative Evaluation

When evaluating the output of a computational cognitive model with respect to human experimental performance it is important to keep in mind the status of qualitative effect presence. The evidence that results from experiments such as Xu and Tenenbaum (2007); Spencer et al. (2011) is informative largely on the basis of indicating which experimental conditions drive a significant difference in participant performance. It is the presence of the performance gap rather than the exact percentage of test items that some sample of participants selected which we should primarily be concerned with. The gap in basic-level items selected when training objects were presented in parallel vs. sequentially happens to be approximately 40% (Spencer et al., 2011). The interpretation of the experiment, however, would be the same whether the size of that gap turned out to be 35% or 65% instead.

It is important for the validity of a parameter-dependent cognitive model that there exist a set of input parameters which results in approximating true human performance on a task. However, another crucial question is to determine the degree to which qualitative effects of model performance are driven by factors internal to the model itself or dependent on specific parameter inputs.

To investigate the parameter independent performance of the NGM, we measured the proportion of parameter configurations which result in qualitatively the same trends as empirical output from Spencer et al. (2011). This is measured in two parts. First that the 'suspicious coincidence effect' is present under parallel presentation trials. This is defined as the proportion of basic-level test items selected being at least 15% lower in the parallel presentation trial compared with the single exemplar trial. Secondly that the sequential presentation demonstrates the same basic trend as baseline generalization. This means that the proportion of test items selected in the sequential condition be within 15% of the single ex-

| | 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 (0.0) | 100 (0.0) | 93.5 (14.6) | 86.7 (19.3) |
| Actual Sequential | 99.12 (3.82) | 88.33 (16.31) | 94.17 (22.47) | 90.83 (26.19) |
| Model Sequential | 100 (0.0) | 100 (0.0) | 100 (0.0) | 91.5 (15.4) |
| **Basic-level objects** | | | | |
| Actual Parallel | 48.24 (40.40) | 10.53 (24.97) | 92.10 (20.31) | 85.09 (27.72) |
| Model Parallel | 49.4 (29.2) | 18 (7.1) | 91.6 (15.3) | 84.8 (20.4) |
| Actual Sequential | 48.24 (40.40) | 53.33 (36.11) | 90.00 (13.68) | 86.67 (26.27) |
| Model Sequential | 49.4 (29.2) | 50.2 (29.7) | 94.4 (4.4) | 91.5 (15.8) |
| **Superordinate-level objects** | | | | |
| Actual Parallel | 7.02 (16.01) | 0.88 (2.62) | 15.35 (11.20) | 81.31 (23.54) |
| Model Parallel | 7 (14.1) | 0 (0.0) | 1 (3.0) | 86.9 (18.1) |
| Actual Sequential | 7.02 (16.01) | 2.5 (4.75) | 13.33 (21.01) | 75.41 (30.04) |
| Model Sequential | 7 (14.1) | 6.4 (13.9) | 22.1 (25.2) | 93.3 (12.7) |

Figure 1: Table showing results comparison between experiments run in Spencer et al. (2011) and output of the NGM. Standard deviations are given in parentheses.

emplar trials. 15% was chosen has a representative sample standard deviation based on the results reported in Spencer et al. (2011).

With multiple parameters in the NGM (means for the salience distributions as well as standard deviation, category distance cutoff) a large number of parameter configurations are possible for the model to be seeded with. A grid search with step-size of 0.1 resulted in 432 tested configurations each run with 1000 simulated 'participants'. The output trends of the NGM were qualitatively consistent with human performance on all trials. The mean size of the 'suspicious coincidence effect' under parallel presentation was $\mu = 58.18\%$ with standard deviation $\sigma = 17.29\%$. Under sequential presentation the mean gap in generalization from baseline was $\mu = 0.8\%$ with standard deviation $\sigma = 0.7\%$. The qualitative trends required to be captured by the model are, on the whole, independent of individual parameter setting.

## Quantitative Evaluation

Parameter tuning and quantitative 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. There are seven different trials types (single exemplar trial, three trials with objects presented in parallel, and three trials with objects presented simultaneously) which we would like to model the experimental findings for. To ensure fair evaluation (and avoid over-fitting), we train the model on only two of the cases originally described in 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. Parameter tuning was performed by running a five-way stepwise (*step size* = 0.1) grid search (two salience distributions means, salience standard deviation, distance threshold, mutual exclusivity threshold).

For each trial, there are three different generalization levels (sub, basic, super) each with a different proportion. To compute the distance from a parameter setting for the model and the empirical data we sum the absolute value of the difference for the proportion for each level. Each trial configuration was run with 1000 simulated 'participants' in the model.

This model captures a 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.

Overall, the output of the NGM is strikingly consistent with human performance on generalization tasks in word learning. Several general patterns are captured here; the strong basic-level bias in generalization from a single, labeled training instance, the 'suspicious coincidence effect' that generalization is more narrow when multiple labeled training items are presented simultaneously, as well as the fact that this ef-

fect is sensitive to temporal manner of presentation. While for practical reasons the NGM was evaluated on a set of seven particular experimental conditions, the underlying trends in generalization are robust under numerous related conditions (Gentner & Namy, 1999; Lawson, 2017; Spencer et al., 2011) Capturing and explaining these trends in a single model is an important contribution.

## General Discussion

Previous 'hypothesis evaluation' models of word learning such as (Xu & Tenenbaum, 2007) attempt to solve the problem of generalization by globally computing the posterior probability of each potential meaning compared to an accumulated set of attested exemplars. While some experimental evidence seems to support this type of globally optimized computation (Xu & Tenenbaum, 2007), other experimental findings (Spencer et al., 2011; Lawson, 2017) are in conflict. The manner in which a fixed set of stimuli is presented to learners (whether simultaneously or in quick succession for instance) induces a large difference in inferred word meanings. Models which attempt to maximize the output probability over hypothesized lexicons cannot account for this effect in a straight-forward manner. To date, no model of word learning has been able to fully capture the range of learner behavior on these tasks.

The Naïve Generalization Model (NGM) presented in this paper offers an explanation of word learning phenomena grounded in category formation (Smith & Medin, 1981). We argue that word learning is fundamentally to *construct* mental representations of words rather than strictly evaluate them. This is a *mechanistic* yet *dynamical* process in which hypothesized representations are generated and only locally revised (*as needed*) based on input data. This does not necessarily maximize global probability of the output vocabulary, but rather the evaluation metric for meanings functions only over what is generated from input by the learner. The NGM explains the mechanism behind meaning generation and generalization for word learning and category formation in a manner that is consistent with and complementary to localist models of referent mapping. Taken together, a more complete picture of word learning begins to emerge.

The NGM correctly predicts the sensitivity of learners to presentation style. These effects of presentation style are robust across related domains, so the explanation offered by the NGM is a real contribution and not simply a method of making rational-level models fit a set of data.

## Acknowledgments

I would like to thank to Charles Yang for helpful advice and feedback throughout this work. Also thank you to John Trueswell and the Penn Language Development and Language Processing Lab for important discussion.

## References

Carvalho, P. F., & Goldstone, R. L. (2015). What you learn is more than what you see: what can sequencing effects tell us about inductive category learning? *Frontiers in psychology*, *6*.

Gelman, S. A., & Markman, E. M. (1986). Categories and induction in young children. *Cognition*, *23*(3), 183–209.

Gentner, D., & Namy, L. L. (1999). Comparison in the development of categories. *Cognitive development*, *14*(4), 487–513.

Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135–176.

Kruschke, J. K. (2008). Models of categorization. *The Cambridge handbook of computational psychology*, 267–301.

Landau, B., Smith, L. B., & Jones, S. S. (1988). The importance of shape in early lexical learning. *Cognitive development*, *3*(3), 299–321.

Lappin, J. S., & Bell, H. H. (1972). Perceptual differentiation of sequential visual patterns. *Attention, Perception, & Psychophysics*, *12*(2), 129–134.

Lawson, C. A. (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*(just-accepted).

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

Mervis, C. B. (1984). Early lexical development: The contributions of mother and child. *Origins of cognitive skills*, 339–370.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive psychology*, *8*(3), 382–439.

Smith, E. E., & Medin, D. L. (1981). *Categories and concepts*. Harvard University Press Cambridge, MA.

Snedeker, J., Gleitman, L., et al. (2004). Why it is hard to label our concepts. *Weaving a lexicon*, *257294*.

Son, J. Y., Smith, L. B., & Goldstone, R. L. (2011). Connecting instances to promote childrens relational reasoning. *Journal of experimental child psychology*, *108*(2), 260–277.

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.

Stevens, J. S., Gleitman, L. R., Trueswell, J. C., & Yang, C. (2016). The pursuit of word meanings. *Cognitive Science*.

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

Xu, F., & Tenenbaum, J. B. (2007). Word learning as bayesian inference. *Psychological review*, *114*(2), 245.

Yu, C. (2008). A statistical associative account of vocabulary growth in early word learning. *Language learning and Development*, *4*(1), 32–62.