How Features Create Knowledge of Kinds

Shohei Hidaka (shhidaka@indiana.edu)

Linda B. Smith (smith4@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University;
1101 East Tenth Street, Bloomington, IN 47405-7007, USA

Abstract

Given a single instance of a novel category, two- and three-year-old children systematically generalize its name to other novel things based on *appropriate* feature dimensions. We explain this in terms of a prediction of the probabilistic density (category likelihood) in feature space from a single novel instance. In principle, observing more instances from a particular probabilistic density, one can estimate the probabilistic density more accurately. In this sense, children's success in generalization from a single instance seems to go beyond the theoretical limit. We provide a theoretical account for the phenomenon. In our theory, these kind of kind specific generalizations, a fast mapping from a single instance to a whole category is due to the structure of the system of learned categories and a sort of optimization of the category organization.

Keywords: Fast mapping; Probability density estimation; Smooth feature space

Different categories are structured in different ways. For example, colors are relevant to categorizing foods but not to categorizing trucks. Further, some categories are decidedly incoherent and not formed by people. For example, people do not form categories that include fish and elephants but not lions (see also Murphy & Medin, 1985). A key question for a theory of categories, then, is how different features are selected for different kinds of categories, and how some categories but not others are selected. An understanding of very young children's novel word generalizations may provide an answer. Two and 3-year-old children generalize novel names for novel things in the "right" way given just a single instance of the category: generalizing names for novel artifacts by shape, for novel animates by multiple features, and for substances by material. This fast mapping of a name for a single thing to a whole category surely facilitates early word learning.

Kind specific generalizations

The phenomenon of interest derives from a widely used experimental task of novel noun generalization (NNG) (see Carey & Bartlett, 1978). In these tasks, children are shown a single novel thing and are then asked to generalize that name to other things. One experimental variable in these studies is the properties of the objects themselves, for example, whether they have features typical of animates (e.g., eyes, legs, hands), features typical of artifacts (e.g, solid with angular parts, straight edges), or features typical of substances (e.g., nonsolid, rounded, flat forms, with irregular shapes). In general, a large literature indicates that when 2- and 3-year-old children are given a novel never-seen-before thing, told its name ("This is a dax"), and asked what other things have that name, the children systematically extend the name to new instances by different features for different kinds (Imai & Gentner, 1997, etc.) Specifically, they extend the names for things with features indicative of animates by multiple similarities, for solid things with features typical of artifacts by shape, and for nonsolid substances by material. For these different kinds of things, young children have clearly solved the feature selection problem and seem to know that different kinds of features matter for different kinds of things. They know what kinds of categories need to be formed.

Category likelihood and feature selection

Children's use of different features to form different kinds of categories in these tasks appears to directly reflect the category likelihoods of those features for known categories. Samuelson and Smith (see also Colunga & Smith, 2005) examined the category structure of the first 312 nouns typically known by children learning English (and in other studies the first 300 nouns learned by children learning Japanese). They measured category structure by asking adults to judge the characteristic within-category similarities of typical instances of individual noun categories on four dimensions, shape, color, texture, and material. They found that individual artifact categories (e.g., chairs, forks, spoons, cups) were judged to have instances that were highly similar in shape but variable in other properties, that animal categories were judged to have instances that were similar in all properties, and that substance categories were judged to have instances that were similar in material (and color). Thus, the importance of features to different kinds of categories for children may reflect the expected distributions of those features for nearby categories, a point we expand on below.

Distribution of category likelihoods

Several recent studies further indicate that children's different patterns of category generalization for different kinds of features may be geometrically organized in some larger feature space (Imai & Gentner, 1997; Colunga & Smith, 2005). For example, Colunga & Smith (2005) showed that children's generalizations of novel names by shape versus material shifted gradually as the presented novel instances varied incrementally from shapes typical of artifacts (complex, lots of angles) to shapes typical of substances (simple rounded shapes). Similarly, Colunga and Smith (under review, see also Yoshida & Smith, 2003) showed that children's generalization by shape versus material shifted gradually as (identically shaped) instances were incrementally varied from solid

(brick like), to perturbable (play dough like), to nonsolid (applesauce like).

We illustrate this idea in Figure 1 which represents category generalization as a likelihood estimation problem in a set of feature dimensions. Figure 1a shows a category likelihood (i.e., relative probability density of category membership is shown on the z axis) and its contour plot projected in a 2-dimensional feature space. Individual contours of categories are represented as ellipses in a 2-dimensional feature space (Figure 1b). The contours of the category likelihoods–that is the distribution of features across the two dimensions of shape and texture/material–varies systematically as a function of location in that space. This idealized representation illustrates the structure that appears to characterize the nouns that are learned early by young children and also to characterize children's generalizations of a newly learned noun to new instances. That is, instances of categories with highly constructed and angular shapes vary little in shape but vary greatly in texture and material whereas categories of animal-like shapes vary little in shape but are also constrained in their variation in texture/material. Finally, the unconstructed simple shapes of substances are correlated with category distributions of relatively variable shapes but limited texture/materials.

The key insight is this: Similar categories, those categories close in the feature space, have similar patterns of category likelihoods for different features. Put another way, the categories in the same region of conceptual space have similar shapes (their generalization patterns) and there is a gradient of category shapes (category likelihoods) across the space as a whole. We will call a space of categories with these properties *smooth*: near (or categories with similar instances) have similar generalization patterns and far (or categories with dissimilar instances) have dissimilar generalization patterns.

Hidaka, Saiki and Smith (2006) analyzed the relation between the central tendencies and generalization patterns of 48 early-learned noun categories in a 16-dimensional feature space (See also the later simulation section). At issue was whether near categories would have similar variance patterns and far categories would have dissimilar ones. Thus, *smoothness* was defined as a correlation between similarities of paired categories in central tendencies and those in variance patterns. They found a positive correlation ($R = 0.537$, Figure 2) which indicates a smooth space of categories.

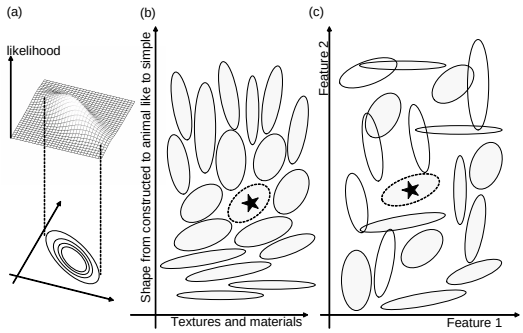

Figure 1: Schematic category organizations having the same likelihoods contours: (a) likelihood pattern is represented as ellipsis (b) "smooth" and psychologically likely organization and (c) randomly distributed organization.

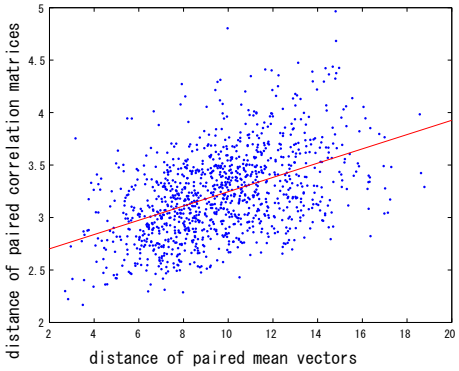

Figure 2: Scatter plot of distances in mean vectors and those in covariance matrices of pairs of categories from adult judgments.

Fast mapping with *smooth* categories

Smoothness may not only be a descriptive property of early learned categories but might also explain children's kind-specific generalizations from a single instance. Smooth categories provide an advantage because the category to be learned is more predictable in a smooth space. To see the relationship between predictability and category organization, let us assume that one knows some categories (shown as solid ellipses in Figure 1b) and observes the first instance (a black star) of a novel category (a broken ellipsis). In the case of Figure 1b, the learner might easily predict the unknown generalization pattern shown by the broken ellipsis. Because nearby categories have similar patterns of likelihoods, the system (through the available space or competition among categories) can predict the likelihood of the unknown category, a likelihood that would also be similar to other known and nearby categories in the feature space. If categories did not have this property of smoothness, if they were distributed like that in Figure 1c, where each category has a variance pattern unrelated to those of nearby categories, the learner has no basis on which to predict the generalization pattern. In sum, the relationship between consistency in category distribution and predictability implies that smooth categories may underlie young children's ability to generalize names for novel things in the right way given just one instance.

Category packing

But why should categories be smooth? The consistency and predictability of categories could derive from the dense interaction among adjacent categories. Starting with this idea, we

propose a theoretical account, the *packing* model, so-called because the category configuration is formed by competition among categories for feature space with the result being that categories are organized in feature space like things are organized in a well-packed suitcase. The main ideas of the packing model are (1) probabilistic densities of categories should not "overlap" and (2) there should be no "gaps" in the feature space in which no category is likely but in which some uncategorized instances do occur. In this sense, Figure 1c, which has many gaps (blanks among categories) and overlaps (intersection among categories), is not well packed. On the other hand, Figure 1b, which has fewer gaps and overlaps, is well packed. More formally, (1) as joint probabilities of paired categories indicate overlap probability, the total sum over feature space of joint probabilities of all paired categories should be smaller , and (2) the probability distribution of all categories should be well fitted to given instances' probability distribution. We call computational condition (1) and (2) *discriminability* and *generalizability* respectively. In general, discriminability and generalizability are in a trade-off relationship: more discriminable categories tend to have more gaps but less overlap, and more generalizable categories tend to have more overlap but less gaps. The optimally *packed* category configuration would be the middle of these two extremes. Next we give a formal description of the packing model and show that an optimal solution for the model has smooth categories as a general trend.

Theoretical Formulation of Packing

We define discriminability as the probability of a discrimination error among categories and we define generalizability as the likelihood of instances given categories. Next we define the *packing cost* function as the sum of discriminability and generalizability, and then derive the category distribution that minimizes the packing cost function.

Consider first a simple case that includes only two categories A and B in one feature dimension (Figure 3). The likelihood of Category A (B) has a single central tendency at mean μ_A (μ_B) and varies with variance σ_A (σ_B). An optimal category discrimination is to judge an instance as the most likely category. That is the probability of discrimination error probability over feature space is the minimum probability in Category A and B (i.e., colored area in Figure 3). That is formally described as $\varepsilon_{AB} = \int_\Omega \min\{P(\theta|A), P(\theta|B)\} d\theta$, where $\min x,y$ is the minimum value in x and y, θ and Ω are respectively a particular feature value and feature space. We define discriminablity as the total error probability between category A and B (the colored area in the figure) when category discrimination is optimal. Because our goal is minimizing ε_{AB}, hereafter we use the upper bound F_{AB}, where $\exp(F_{AB}) = \int_\Omega \{P(\theta|A)P(\theta|B)\}^{\frac{1}{2}} d\theta \geq \varepsilon_{AB}$, instead of the error per se. In particular, when $P(\theta|A)$ and $P(\theta|B)$ are normal distributions, the upper bound of error is $F_{ij} = -\frac{1}{4}(\mu_i - \mu_j)^t(\sigma_i+\sigma_j)^{-1}(\mu_i-\mu_j) - \frac{1}{2}\log(\frac{1}{2}|\sigma_i+\sigma_j|) + \frac{1}{2}\log(|\sigma_i||\sigma_j|)$, which is called the Bhattarcheryya bound (Duda, Hart, and Stork, 2000). In the following derivations, we assume each probabilistic density is a normal distribution, and utilize the Bhattarcheryya bound as discriminability of categories. In the more general case of N categories in D dimensional space $\Omega \supset \theta$, we assume a likelihood $P(\theta|c_i)$ of category c_i with feature θ defined as a normal distribution having mean μ_i vector and covariance matrix σ_i:

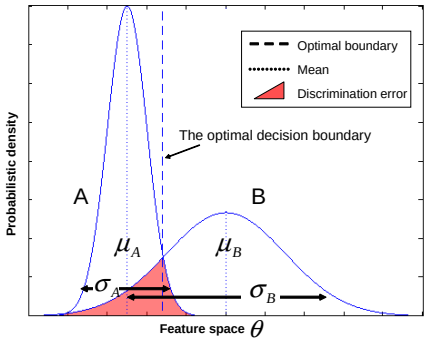

Figure 3: Category likelihoods of two categories

$$P(\theta|c_i) = ((2\pi)^D|\sigma_i|)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\theta-\mu_i)^t\sigma_i^{-1}(\theta-\mu_i)) \quad (1)$$

where superscript t indicates transposition. In the case of N categories, the discriminability is defined as the sum of Bhattercherrya bound of all category pairs: $\bar{F_N} = \log\left[\sum_i \sum_j \exp(F_{ij})\right]$.

Next we define generalizability that is, the likelihood of instances given a category. The logarithm of the likelihood of category c_i (Equation 1) without constant terms can be simply written as $G_i = \log(\sigma_i) + \sum_k^{K_i}(x_{ik}-\mu_i)^t\sigma_i^{-1}(x_{ik}-\mu_i)$ where x_{ik} $(k = 1,2,\ldots,K_i)$ are observed instances of novel category c_i. The packing cost L_n consists of combination among discriminability F_n and fitting to given instances G_i. This is mathematically formulated as the minimization of the Lagrange equation L_n with a constant λ as follows: $L_N = 4\bar{F_N} + \sum_i \lambda(G_i - \log(C))$, where C is a constant.

Optimally-packed categories are smooth Next we show a structural property of optimally-packed categories by solving the differential of the packing cost with respective to statistical parameters of categories. The differential of $\bar{F_N}$ with respective to a parameter X is $\frac{\partial \bar{F_N}}{\partial X} = \mathrm{E}_{\bar{F}}\left[\frac{\partial F_{ij}}{\partial X}\right] = \bar{F_N}^{-1}\left\{\sum_{i,j} Q_{ij}\frac{\partial F_{ij}}{\partial X}\right\}$ where $\mathrm{E}_{\bar{F}}\left[\frac{\partial F_{ij}}{\partial X}\right]$ is the expectation for $Q_{ij} = P(c_i)P(c_j)\exp(F_{ij})$ $(i,j = 1,2,\ldots,n)$ as a probabilistic density. Since the differential of L_N with respect to σ_i $(i = 1,2,\ldots,n)$ is zero, we obtain the following equation: $\frac{\partial L_N}{\partial \sigma_i^{-1}} = \mathrm{E}_{\bar{F}}[(\mu_i - \bar{\mu}_{ij})(\mu_i-\bar{\mu}_{ij})^t + \bar{\sigma}_{ij} - \sigma_i] + \lambda\{S_i - \sigma_i\} = 0$ where $\bar{\sigma}_{ij} = 2\sigma_i(\sigma_i+\sigma_j)^{-1}\sigma_j$, $\bar{\mu}_{ij} = \frac{1}{2}\bar{\sigma}_{ij}(\sigma_i^{-1}\mu_i + \sigma_j^{-1}\mu_j)$ and $S_i = \sum_k^{K_i}(x_k-\mu_i)(x_k-\mu_i)^t$ is the scatter matrix. By solving the above equation $(i = 1,2,\ldots,N)$, we obtain the follow-

ing relationship [1].

$$\sigma_i = \frac{\sum_j Q_{ij}\left\{\hat{S}_{ij} + \bar{\sigma}_{ij}^{-1}\right\} + \lambda S_i}{\sum_j Q_{ij} + \lambda K_i} \quad (2)$$

where $\hat{S}_{ij} = (\mu_i - \bar{\mu}_{ij})(\mu_i - \bar{\mu}_{ij})^t$. Equation (2) indicates covariance σ_i consists of the weighted average of three components (i.e., Q_{ij} and λ as its probabilistic density), the scatter matrix of categories $\hat{S}_{ij}$, the harmonic mean of covariance matrices $\bar{\sigma}_{ij}^{-1}$ and the scatter matrix of observed instances S_i. Note that Q_{ij} exponentially decays in proportion to the distance between category c_i and c_j. Thus, the scatter matrix of categories $\hat{S}_{ij}$ reflects the distribution of nearby categories c_j around category c_i. Conceptually, it means that nearby categories constrain the "niche" in the feature space for category c_i to spread out (The broken ellipsis in Figure 1a or b). The harmonic means of covariance matrices $\bar{\sigma}_{ij}$ indicate that σ_i would be similar to those of other "closer" categories. Therefore, Equation (2) implies that the general structure of optimally-packed categories is *smooth*, i.e., closer categories in feature space have similar covariance patterns.

Estimation of a novel category from the first instance Here we derive the covariance estimation for "novel word generalization" that one only knows the first instance. In this case, we approximately obtain $S_i \approx 0$ by assuming the first instance is close to the true mean (K_i=1 and $x_k = \mu_i$). In addition, we can obtain λ by solving the constraint equation $\frac{\partial L_N}{\partial \lambda} = G_i - \log(C) = 0$. Then the optimal covariance of the novel category is given as follows.

$$\sigma_i = C \left| \sum_j^N Q_{ij}\left\{\hat{S}_{ij} + \bar{\sigma}_{ij}^{-1}\right\} \right|^{-1} \sum_j^N Q_{ij}\left\{\hat{S}_{ij} + \bar{\sigma}_{ij}^{-1}\right\} \quad (3)$$

The EM algorithm (Dempster et al, 1977) is available for iterative minimizing the packing cost with Equation (2) or (3).

Simulation of novel word generalization

In this simulation, we demonstrate that the model can predict the likelihoods (feature distribution patterns) of natural categories which were unknown to the model. This is similar to the problem of how children can predict the right generalization pattern from a single instance. We formulated fast mapping in this sense as a prediction of an unknown probabilistic density (category likelihood) in feature dimensions from a single novel instance. Specifically, in this simulation, the model "knows" 47 natural noun categories and observe the first instance of the 48th novel category. Then the task of the model is to predict the probabilistic density pattern of this new category. The model's prediction of the probabilistic density was calculated by an optimal solution for the packing cost with respect to the configuration of surrounding known categories (See also the theoretical formulation). We used 48 natural categories sampled from the Mac Arthur-Bates Communicative Development Inventory (MCDI; Fenson et al., 1994), which is a vocabulary list of words normatively known by 50% of 30-month-olds, and 16-dimensional features provided by adults in a judgment task. The model's prediction of the probability density of novel categories is compared to actual statistics of natural categories. Because the structure of *known* categories has been shown to play a major role in children's kind specific generalizations (e.g., Colunga & Smith, 2005), we manipulate the number of samples from each category the model knows, using norms of the acquisition age of the 48 nouns from 16 to 30-month-olds (from the MCDI). The goodness of fast mapping is discussed in light of this simulated word development.

Procedure

In each trial of the simulation, one category is assigned as unknown, and the other 47 categories are known. Each category is assumed as a normal distribution, and the model predicts the unknown covariance matrix based on the given means and covariances of known categories. This process is simulated for each noun category as unknown and for each acquisition rate of nouns corresponding from hypothetical sixteen to thirty month of age. In sum, novel word generalization was simulated 50 times for 15 ages by 48 categories.

Prediction We used Equation (3) to calculate a covariance matrix of a novel category σ_i from an instance sampled from the category ($\mu_i = x_{i1}$) and other known categories (μ_j and σ_j, $j \neq i$). The calculated σ_i minimizes the packing cost in terms of adjacent other categories. The coefficient λ in each simulation is calculated by assuming the scaling constant as the determinant of the covariance matrix of the unknown category ($C = |(N-1)^{-1} S_i|$).

Word development To mimic the likely growth in knowledge about known categories over 16 to 30 month old, we assumed that the number of instances observed by the model increases in proportion to these acquisition rates. The logic behind this manupulation is this: Statistical properties in the adult judgments on feature of categories are products of development, and that children learn a subset of instances adults know. Specifically, we generated 2000 random instances for each category which has the same mean and covariance matrix as that of adult judgments of these categories on the 16 dimensions (i.e., adults' knowledge as a parent population). Then we assumed 50 out of 2000 instances were "known" (samples) at the age the norms indicated the word was in 100% of the children's productive vocabulary. In this learning scheme, the accuracy of estimation of statistics for hypothetical known categories increases in proportion to the acquisition rates in the MCDI list. Figure 4 shows the mean acquisition rates in the MCDI (solid line) and smoothness index (broken line) over 48 nouns. The smoothness index in the hypothetical known categories gradually increases from 0.1 to 0.4 along increment of known instances (from approximately 10 to 45 instances per category).

[1] The precise solution is given as a quadratic eigenvalue problem.

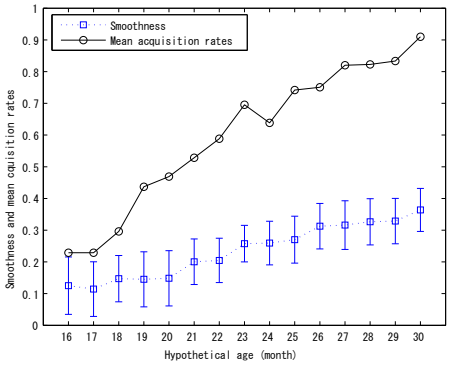

Figure 4: Average word acquisition rates and the smoothness index of 48 categories

Categories and features

We used adult judgment data on early acquired noun categories using adjectives as feature dimensions (Hidaka & Saiki, 2004). The survey includes 48 nouns selected randomly from 312 nouns in the MCDI. They also span a variety of kinds including food, animals, vehicles, tools, furniture, and so forth. In the survey, 104 Japanese undergraduates rated noun's typical features using sixteen five-point-scale pairs of adjectives. For example, subjects rated the noun category "cat" as either "very large", "large", "neither", "small", and "very small". The sixteen adjective pairs having larger variance across noun categories were selected out of the initial 41 pairs, thus these adjectives would characterize the current set of noun categories (see Hidaka & Saiki, 2004 for the detail).

- **Adjective pairs (feature dimensions)**
  dynamic-static, wet-dry, light-heavy, large-small, complex-simple, slow-quick, quiet-noisy, stable-unstable, cool-warm, natural-artificial, round-square, weak-strong, rough hewn-finely crafted, straight-curved, smooth-bumpy, hard-soft.

- **Noun categories**
  butterfly, cat, fish, frog, horse, monkey, tiger, arm, eye, hand, knee, tongue, boots, gloves, jeans, shirt, banana, egg, ice cream, milk, pizza, salt, toast, bed, chair, door, refrigerator, table, rain, snow, stone, tree, water, camera, cup, key, money, paper, scissors, plant, balloon, book, doll, glue, airplane, train, car, bicycle

Results

The logic of our analysis of the simulations is this: We gave the model 47 of the 48 categories (whose generalization gradients were generated by adult judgments) and then asked, given the packing cost, if it could generate from a single instance the generalization of the gradient of the 48th category, thus a simulation of fast mapping. To assess how well the model generated the generalization pattern of the missing category, we examined the correlations, category by category, between the predicted variances, covariances, and joint variance/covariance of the predicted category with its actual generalization pattern from the adult judgments. Considering all the categories (and thus a hypothetical 30-month-old), the correlations were 0.58, 0.56 and 0.88 respectively. To assess whether knowing the whole organization of categories enabled better category learning than merely learning single categories independently of the structure of the whole, we also compared the three measures of generalization for categories defined by 3 randomly selected instances with the measures from the adult judgments. These correlations are low: that for variances, covariances, and joint variance/covariance are 0.23, 0.23 and 0.45 respectively. Moreover, for a series of vocabularies (at monthly intervals from 16 to 30 months by the MCDI), the covariance matrices predicted with an only instance by the packing optimization have significantly higher correlation than those estimated from three instances. These finding illustrate the main idea of the packing account: the location of a to-be-formed category in a geometry of categories –because of the local interactions of nearby categories– constrains the possible shape of the category in the feature space. The packing model, however, depends on these being a densly packed set of known categories. Figure 5 shows the correlation changing over the hypothetical development trajectory. Each point and bar shows mean and standard deviation of 50 simulations with different random values. Since the increase in the accuracy of known category estimations is due to increase in the number of known instances along with the hypothetical age (Figure 4), we analyzed the correlation among these variables. The accuracy in prediction of novel categories (for covariance matrices) increases in proportion to the number of instances (correlation to acquisition rate is 0.911, $p < 0.01$) and smoothness (0.816, $p < 0.01$) of known categories. It suggests that novel word prediction gets more accurate along with the growth of known categories. One reason of this increment may be due to the smoothness of known categories, because smoother categories would have more predictability. In sum, the prediction by packing optimization succeeded in a novel word generalization as accurate as known categories. The accuracy in prediction of novel words increases as accuracy in estimation of known categories increases. These results suggest that the packing optimization is powerful enough for earlier word learning.

Discussion

These results show how children's solution to the feature selection problem –to selecting the right features for a given but as yet unknown category– may be a geometrical property of the *system* of known categories in feature space. For this to work, however, two computational conditions seem necessary. First, the feature space itself must be organized in a predictable way. There are hints -though more is needed - that this is so (Hidaka et al., 2006; Colunga & Smith, 2005). If natural categories comprise a smooth space –for whatever reason– it means that a learner of a new category can predict the likelihood pattern of a novel category. This can be done because of a correlation between similarity of cat-

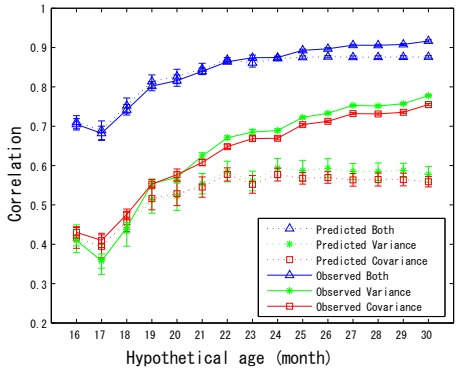

Figure 5: Correlation of the predicted variance patterns to the real data and that of estimation by samples

egories and similarity of category likelihoods. This idea is supported by the simulation results that show a relation between the accuracy of known categories and generalization of a novel category. The second computational requirement is that the learning system needs to utilize not only the observed instances of any given category, but also the consistency of the distribution of those instances in the whole configuration of categories. The core mechanism of the packing model is an optimization of category configuration in terms of discriminability and generalizability. The simulation suggests that the prediction based on this structural consistency was enough powerful. Meanwhile, fast mapping based on the category packing would not work very well when a learner has only a small number of inaccurately estimated categories. These two points raise a number of testable questions about children's novel noun generalizations. We know these become more systematic with the number of known nouns. But the present results suggest that it is not just number of known nouns but how representative the known instances of those nouns are with respect to that kind of category. This in turn suggests that one might be able to speed up noun learning by presenting very young children -not just with many categories in a region of feature space -but with a *few* well packed instances of those categories, a result that has been empirically demonstrated in the case of artifacts (Smith et al, 2002), within a region of a feature space in the earliest developmental stage.

The explanation offered here is consistent with several other approaches to this phenomenon (Kemp et al., 2006; Colunga & Smith, 2005) which also suggest that these kind specific generalizations are based on higher order correlations among feature dimensions. Colunga & Smith (2005) showed that a connectionist model, fed the feature regularities characteristic of early learned nouns, could generalize names for things with different features by different properties. There are a number of similarities and differences across the models that merit investigation. This analysis however, points to new aspects of the system of knowledge that may underlie children's smart feature selection and point in new directions for research - how dense known categories are in a given area of feature space and how well the known instances of those categories portray the generalization pattern for that category.

References

Carey, S. and Bartlett, E. (1978) Acquiring a single new word. Papers reports on child language development. *15*, 17-29.

Colunga, E. and Smith, L. B. (2005). From the lexicon to expectations about kinds: A role for associative learning. *Psychological Review*, *112*, 347-382.

Dempster, A., Laird, N., and & Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, *39* (1, Series B), 1-387.

Duda, R. O., Hart, P. E., & Stork, D. G. (2000) *Pattern Classification* (2nd ed) , New York: John Wiley & Sons.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., & Pethick, S. J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59* (5, Serial No. 242) Chicago: University of Chicago Press.

Hidaka, S. & Saiki, J. (2004). A mechanism of ontological boundary shifting. In *Proceedings of the Twenty Sixth Annual Conference of the Cognitive Science Society*, 565–570.

Hidaka, S., Saiki, J. & Smith, L. B. (2006). Semantic packing as a core mechanism of category coherence, fast mapping and basic level categories. In *Proceedings of the Twenty Eighth Annual Conference of the Cognitive Science Society*, 1500–1505.

Imai, M. & Gentner, D. (1997). A cross-linguistic study of early word meaning: universal ontology and linguistic influence., *Cognition*, *62*, 169-200

Kemp, C., Perfors, A. & Tenenbaum, J. B. (2006) Learning overhypotheses. In *Proceedings of the Twenty Eighth Annual Conference of Cognitive Society*, 417-422.

Murphy, G.L. & Medin, D.L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92,* 289–316.

Samuelson, L. & Smith, L. (1999). Early noun vocabularies: do ontology, category structure and syntax correspond?, *Cognition*, *73*, 1–33.

Smith, L. B., Jones, S. S., Landau, B., Gershkoff-Stowe, L. & Samuelson, L. (2002). Object name learning provides on-the-job training for attention. *Psychological Science*, *13*, 13-19.

Yoshida, H. & Smith, L. B. (2003). Shifting ontological boundaries: how Japanese- and English- speaking children generalize names for animals and artifacts. *Developmental Science*, *6*, 1–34.