Semantic Packing As a Core Mechanism of Category Coherence, Fast Mapping and Basic Level Categories

Shohei Hidaka (hidaka@cog.ist.i.kyoto-u.ac.jp)
Graduate School of Informatics, Kyoto University and JSPS Research Fellow

Jun Saiki (saiki@cv.jinkan.kyoto-u.ac.jp)
Graduate School of Human and Environmental Studies, Kyoto University

Linda B. Smith (smith4@indiana.edu)
Department of Psychological and Brain Sciences, Indiana University

Abstract

In the present study, category coherence, a question why intuitive groupings for natural categories exist, is considered by applying a computational theory that models discrimination and generalization. The computational process is what we called "semantic packing". In the model, category learners' two conflicting constraints on discrimination and generalization are optimized simultaneously as if it packed knowledge into memory. The model also revealed a computational structure of efficient categories, called basic categories, by mathematical proof and exemplification of the relationship between the past proposed theory and our theory. Furthermore, the empirical evidence of the theory from human semantic rating was shown and used for testing predictability of novel natural categories. The results suggest that semantic packing could reproduce the configuration of natural categories from only their generalization without any knowledge as fast mapping in which children can generalize a novel instance without trial and error. In summary, the semantic packing could be a core mechanism of the three essential categorization processes, category coherence, fast mapping, and basic level category.

Introduction

Why Does "Category Coherence" Emerge?

Murphy and Medin (1985) define "category coherence" as the intuitive and useful groupings that characterize natural categories, and claim that this coherence is one of most important aspects of semantic cognition. For example, color is more important for discrimination when the item is a pea rather than when it is a ball. How do we learn this? And how do we use the knowledge that we have learned? How, when we see an object-a potential pea or ball-do we know to attend to color or not?

This problem of feature selection has played a key role in theoretical discussions of the mechanisms that underlie category learning. In past studies, some measures of a learner's category representation were proposed to explain the basic level category advantage. The basic category is known as the most efficient level for various cognitive processes, such as picture naming, category or feature listing, and speech frequency (Rosch, Mervis, Gray, Johnson & Boyes-Braem, 1976). This concerns the relative frequency of features or categories, thus, these measures depend on what features or categories are selected (Murphy and Lassaline, 1997). For these and other reasons, Murphy and Medin (1985) claimed that categorization based only similarity and correlations is not enough to solve the coherence problem. Insead, they suggested that categorization is based on folk theories and causal induction, or what some call "theory theory". However, their major criticism of similarity-based accounts, circularity, is also a problem for theory theories. To what theory to apply to any thing – a theory about food and the relevance of color or a theory about artifacts and the nonerelevance of color – one has to already know what kind of thing is at hand.

The Outline

The main goal of this study is to propose a computational theory that connects the emergence of category coherence, basic level categories, and fast mapping (see also the next paragraph). To this end, we propose two metrics. The first is what we call *smoothness*. This concerns the relationship between features and generalization of categories found in developmental studies and is a measure how cohesive categories are. Categorization with the smooth feature space makes a novel instance generalized precisely, which is considered as fast mapping. The second metric concerns discrimination, which yields category coherence as the result. There is an essential trade-off between generalization and discrimination. Both cannot be maximized simultaneously. The joint optimization is obtained by what we call "semantic packing". This process is analogous to a task in which various sized and shaped containers (categories) are efficiently packed into a larger but finite container (memory and the attentional and retrieval processes that apply to memory). The packing process is related to processes such as cue validity (Rosch et al. 1976) and category utility (Corter & Gluck, 1992), which are measures for basic level categories. We first present the theory metaphorically and then the formal mathematical specification later.

Working Definition: "Semantic Smoothness" in Natural Categories

Many developmental studies using novel word generalization tasks have shown that children systematically attend to different properties when generalizing different types of entities, a process known as "fast mapping". For example, children generalized solid artifacts and non-solid substances based on the similarity of shape and material, respectively (Soja, Carey & Spelke, 1991). Therefore, children seem to solve the circularity problem,

knowing to attend to the right properties even though they do not yet know the category. Importantly, neither younger children nor late talkers show the same generalization pattern (Jones, 2003). This finding suggests that these differential weighting patterns are learned. An adult rating study of the similarities that characterize the first 300 nouns learned by children showed that their attentional biases in noun extension tasks reflected the regularities in the corpus of early noun categories. Specifically, there is a high correlation between category generalization (i.e., shape- or material-based category organization) and property (i.e., solidity) (Samuelson & Smith, 1999). In other words, these rating data indicated that "property" (i.e. solid or non-solid) of natural categories predicted "generalization" (i.e. property weighting: shape- or material- based generalization), and vice versa. The semantic space would be in our terms "smooth" if the correlation between property and generalization was universal in any semantic domain, as robust as the correlation between solidity and shape-based generalization in early acquired noun categories, that is, the property difference between any two categories would be correlated to a difference in how those categories are generalized (see also Equation 12 for the mathematical definition). Furthermore, the smooth semantic space would form clusters that have a correlated property-generalization relationship. In other words, because of the property-generalization correlation, similarly distributed categories would be grouped near each other (i.e., domain specific property weighting: Figure 1 (b)) [1]. Thus "smoothness" of the semantic space may be considered as a quantitative measure of category coherence. Here, in an empirical study, we investigate the smoothness of the semantic space of early-acquired nouns. The results indicate that natural categories have "smoothness", that some categories (e.g., "cat" and "tiger") share similar properties and generalization patterns, and that other categories (e.g., "cat" and "chair") share dissimilar ones. Before a more detailed presentation of the theory, we consider how one recent theory of category development has dealt with this issue.

Efficiency in Semantic Cognition

Past proposed measures of the basic category advantage concentrate on the discriminability of categories. More specifically, cue and category validity are respectively maximized in subordinate and superordinate categories (Murphy & Lassaline, 1997). Furthermore one variety of category utility is equivalent to mutual information (i.e. degree of probabilistic independence) between categories and features (Gluck & Corter, 1985). This is maximized with dependent categories and features (e.g., one-to-one mapping between features and categories: an extreme case of subordinate categories). These measures are (positively or negatively) discriminabilities of category. Importantly, the surface differences in these measures (Corter & Gluck, 1992) could just reflect the differences in the selected set of categories and features (Murphy & Lassaline, 1997).

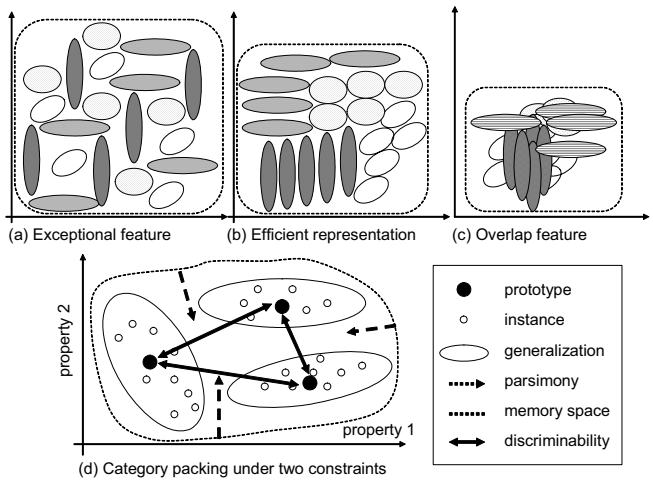

Figure 1: 20 categories (ellipses represent contour of probabilistic distributions) in property space. (a) Category coding with exceptional feature: More property space (dashed line) is needed for categorization. (b) Efficient category coding: Less property space and more discriminability. Categories with similar generalization are localized. (c) Category coding with overlapped feature: Feature is useless for categorization. (d) Category packing is considered the balanced optimal state under two constraints: memory parsimony and discriminability of category.

We argue that learning efficiency under the trade off between discrimination and generalization is the core mechanism underlying category coherence. We define "semantic memory" for the purposes of this paper as a set of categories, which is formulated as a probabilistic distribution in psychological feature space (Figure 1). A reasonable question to ask is what features should be represented for any category (e.g. A feature "with four limbs" for both "dog" and "cat"). Categories in an overlapping feature space need less memory. But overlapping features also mean less discriminability of the categories (Figure 1 (c): E_1 gets smaller, but G gets bigger in Equation 7). One solution to this is to store only exceptional features or conjunctive features such that the conjunctions are unique for each category(e.g., "four-limbed, eyed, a pet, meat-eating, gather in crowds, ..." as single feature only for "dog"). The categories in this exceptional feature space must be sparse and will therefore need more memory (i.e., larger dashed-line enclosure). This is, by definition, all categories must be defined by unique conjunctive feature sets. (Figure 1 (a): G gets smaller, but E_1 gets bigger in Equation 7). However, this means that what one knows about one category can not help in learning or making decisions

[1] Consistent with the mathematical term "smooth", "smoothness" here refers to the probabilistic degree of local linearity of manifold (category-feature space) where generalization σ_i is curvature around μ_i .

about another. This is precisely where the developmental evidence is most compelling: The categories children already know help them learn new similarly structured categories. This insight indicates that discriminability and generalization from knowledge about one category to another trade off. An efficient semantic memory may try to optimize both, and that may work at some midlevel between these two extremes (Figure 1 (b)).

What then would emerge in such a case? We describe the optimal state as "category packing", where the system packs categories of a particular shape in feature space close together, thus taking up less feature space overall (Figure 1 (d): both G, E_1 and L get smaller in Equation 7). Assume that one creates optimally organized categories by moving the prototypes or, alternatively, the distributions (i.e. $\frac{\partial L}{\partial \mu_i} = 0$ or $\frac{\partial L}{\partial \mu_i} = 0$: note that this is not "category learning"). This process is analogous as packing things into smaller space (categories or things avoid probabilistic or solid "collision", respectively). The most efficient packing of different sizes and shapes of things (or categories, we propose) consists of packing similarly shaped things together (i.e. emergence of semantic smoothness: Figure 1 (b) and Equation 12). Next we briefly introduce the detailed formulation of our theory in a simple case in which categories are defined by prototypical representations.

Theoretical Formulation of Packing

We prove the equivalence between semantic packing and smoothness, under the simplification that each category is represented by its prototype and generalization pattern. Note that this simplification does not assume any predefined specific "feature" or "category" in the packing process and also that the prototypical representation is not a necessary assumption but an application. Instead of specifying categories and features, we investigated what category organization emerges as the result of the efficient categorization. Before this proof, we also prove the approximate equivalence among our discriminability measure, cue validity, and category utility.

Assume that there are n categories $c_1, c_2, ..., c_i, ..., c_n$ in feature space $\theta \subset \Omega$. Assume the conditional probability of feature θ given category c_i as $P(\theta|c_i)$. F defined as the equation below indicates measure of discriminability among categories, which is upper bound of minimum error ratio under optimal decision making.

$$F_n = \int_\Omega \prod_i P(\theta|c_i)^{\frac{1}{n}} d\theta \quad (1)$$

Let the joint probability of category c_i in feature θ be $P(c_i, \theta)$. If one evaluated that it is category c_i when $\theta \subset \Omega_i$, then the correct ratio is $\sum_i \int_{\Omega_i} P(c_i, \theta) d\theta$. In particular $n = 2$, the minimum error ratio of the optimal decision (i.e. Bayes decision) is $\hat{\epsilon} = \int_\Omega \min(P(c_1, \theta), P(c_2, \theta)) d\theta$, because, to maximize the correct ratio, one must choose the category with largest probability given θ. To estimate analytically the exact minimum or maximum distribution is difficult. Accordingly we used instead the upper bound of the minimum, called the "Bhattacheryya bound" (Duda, Hart & Stork, 2000). The inequality $\min(a, b) \leq a^\beta b^{1-\beta}$ is true when $a, b \geq 0$ and $0 \leq \beta \leq 1$. Therefore, Using Bayse' theorem $P(c_i, \theta) = P(\theta|c_i)P(c_i)$, where $P(\theta|c_i)$ is conditional probability θ given c_i.

$$\hat{\epsilon} \leq \sqrt{P(c_1)^\beta P(c_2)^{1-\beta}} \int \sqrt{P(\theta|c_1)^\beta P(\theta|c_2)^{1-\beta}} d\theta \quad (2)$$

The right side in this inequality is called the "Charnoff bound", the upper bound of the error. β giving the minimum a Charnoff bound is around $\beta = \frac{1}{2}$, therefore, Chernoff bound with $\beta = \frac{1}{2}$, called Bhattacheryya bound ψ_b can be used as the second best bound. Obviously, $\psi_b = \sqrt{P(c_1)P(c_2)}F_2$. When $n > 2$, for short tailed probabilistic distribution such as normal distribution, $F \cong k_F F_2 = B$ can be approximated with constant k_F in the local of the particular nearest pair of normal distributions. Therefore, suppose $\psi_{ij} = \sqrt{P(c_i)P(c_j)} \int P(\theta|c_i)^{\frac{1}{2}} P(\theta|c_j)^{\frac{1}{2}} d\theta$.

$$\prod_i P(c_i)^{\frac{1}{n}} F_n \cong k_F \sum_i \sum_{j \neq i} B_{ij} \quad (3)$$

The maximum correct ratio of cue validity ($P(c_i|\theta)$: Rosch et al., 1976) model considering frequency of feature $P(\theta)$ (Reed, 1972) is defined as $\hat{Q} = \int \max_i P(c_i|\theta)P(\theta)d\theta = \int \max_i P(\theta, c_i)d\theta$. Therefore,

$$F_n \cong \prod_i P(c_i)^{-\frac{1}{n}} (1 - Q) \quad (4)$$

In other words, minimum F_n indicates the maximum cue validity Q. The category utility of category i is defined as $U(c_i) = P(c_i) \int (P(\theta|c_i)^2 - P(\theta)^2) d\theta$ (Corter & Gluck, 1992). Total category utility $U = \sum_i U(c_i)$ is a similar measure with F as follows (i.e. the order is $O(U) = O(-F^2)$). Applying $P(\theta) = \sum_i P(\theta, c_i)$ and $k_u = P(c_i)^{-1} - 1$.

$$U = \sum_i k_u \int P(\theta, c_i)^2 d\theta - 2 \int \sum_i \sum_{j \neq i} (\frac{dB_{ij}}{d\theta})^2 d\theta \quad (5)$$

Next, we prove the equivalence between semantic packing and smoothness when the probability of feature θ given ith category $P(\theta|c_i)$ is defined as, a prototypical representation, a d-dimensional normal distribution. Then $P(\theta|c_i) = ((2\pi)^d |\sigma_i|)^{-\frac{1}{2}} \exp(-\frac{1}{2}(\theta - \mu_i)^t \sigma_i^{-1} (\theta - \mu_i))$, where a mean vector (i.e., prototype) and covariance matrix (i.e., generalization or feature weighting) are μ_i and σ_i. The superscript t refers to transposition. The discriminability measure (Equation 1) can be rewritten as follows. Assume that $A = \sum_i^n \sigma_i^{-1}$, $B = \sum_i^n \sigma_i^{-1} \mu_i$, $C = \sum_i^n \mu_i^t \sigma_i^{-1} \mu_i$, and $G = \log(F)$

$$G = \frac{1}{2n}(B^t A^{-1} B - C - n \log |A| - \sum_i^n \log |\sigma_i|) \quad (6)$$

Optimization for only the constraint $\frac{\partial G}{\partial \mu_i}$ or $\frac{\partial G}{\partial \sigma_i}$ (i.e., discriminability in Figure 1) gives $(\mu_i - \mu_j)^t (\mu_i - \mu_j) \to \infty$ or

$|\sigma_i| \rightarrow 0$, indicating an immense amount of feature space or an instance as a category (i.e., no generalization), respectively. Therefore, constraints to normal distributions $E_1 = \sum_i^n ||\mu_i||^2 = \sum_i^n \mu_i^t \mu_i$ and $E_2 = \log|A^{-1}|$ are necessary. For the cognitive process, the constraints E_1 and E_2 refer to maintenance of constant memory space (i.e. parsimony in Figure 1) and generalization ranges, respectively. The Lagrange multiplier method is used for optimization of the constraints. The Lagrange equation with multiplier λ is $L = G + \lambda_1 E_1 + \lambda_2 E_2$, which indicates semantic packing (L) optimizes both discriminability (G) and generalization (E_1 and E_2).

$$\frac{\partial L}{\partial \mu_i} = \frac{\partial}{\partial \mu_i}(G + \lambda_1 E_1) = -\sigma_i^{-1}(\mu_i - \bar{\mu}) + \lambda_1 \mu_i = 0 \quad (7)$$

where $\bar{\mu} = A^{-1}B = (\sum_i^n \sigma_i^{-1})^{-1} \sum_i^n \sigma_i^{-1} \mu_i$.

$$\mu_i = -(\lambda_1 \sigma_i - I)^{-1} \bar{\mu} \quad (8)$$

where I is the identity matrix. Therefore the relationship between a pair of categories when L is optimized as a function of μ is

$$\Delta\mu_{ij} = \lambda_1(\lambda_1\sigma_i - I)^{-1} \Delta\sigma_{ij} (\lambda_1 \sigma_j - I)^{-1} \bar{\mu} \quad (9)$$

where $\Delta\mu_{ij} = \mu_i - \mu_j$ and $\Delta\sigma_{ij} = \sigma_i - \sigma_j$. Next, in addition to μ_i, L is optimized as a function of σ_i. As $\frac{\partial L}{\partial \sigma_i} = \frac{\partial}{\partial \sigma_i}(2nG + \lambda_2 E_2)$, thus applying $\frac{\partial 2nG}{\partial \sigma_i} = \sigma_i^{-1}(\bar{\mu} - \mu_i)(\bar{\mu} - \mu_i)^t \sigma_i^{-1} + n\sigma_i^{-1} A^{-1} \sigma_i^{-1} - \sigma_i^{-1}$

$$\sigma_i \frac{\partial L}{\partial \sigma_i} \sigma_i = (\bar{\mu} - \mu_i)(\bar{\mu} - \mu_i)^t + (n + \lambda_2) A^{-1} - \sigma_i \quad (10)$$

As $\sigma_i \frac{\partial L}{\partial \sigma_i} \sigma_i - \sigma_j \frac{\partial L}{\partial \sigma_j} \sigma_j = 0$

$$\Delta\sigma_{ij} = \sum_{k=i,j} (-1)^{\delta_{ki}} (\hat{\mu} - \mu_k)(\hat{\mu} - \mu_k)^t \quad (11)$$

where $\delta_{ii} = 1$ when $i = j$, otherwise $\delta_{ij} = 0$. Notice that σ_i is constant in Equation 9, and Equation 11 is $\Delta\sigma_{ij} \cong O(\Delta\mu_{ij})$. Consequently, the approximate monotonic relationship between $\Delta\mu_{ij}$ and $\Delta\sigma_{ij}$ with a given constant α (i.e. "smoothness") emerges, when $\frac{\partial L}{\partial \mu_i} = 0$ or $\frac{\partial L}{\partial \sigma_i} = 0$ (i.e. "packing").

$$||\mu_i - \mu_j|| \approx \alpha ||\sigma_i - \sigma_j|| \quad (12)$$

In other words, semantic smoothness, which is the correlation between feature and generalization (Equation 12), is approximately equivalent to semantic packing. A learning system with smooth categories that optimize the packing principle, and vice versa.

An analytic solution to $\frac{\partial L}{\partial \mu_i} = 0$ is demonstrated as follows. Assume that $E_1' = \sum_i^n \nu_i^t \nu_i$ where $\nu_i = \mu_i - A^{-1}B$ to be the constraint instead of E_1, and note that the replacement does not lose generality. Solving the Lagrange equation $L = \frac{G}{2} + \frac{\lambda}{2} E_1'$, we get $\frac{\partial L}{\partial \mu_i} = -\sigma_i^{-1} \nu_i + \lambda \sum_j^n (\delta_{ij} - \sigma_i^{-1} A^{-1}) \nu_j$ where $\delta_{ii} = 1$ when $i = j$, or $\delta_{ij} = 0$. Let $\nu = (\nu_1, \nu_2, ..., \nu_n)^t$ be the d-by-n-dimensional vector having ν_i as its ith elements. In addition, let $\mathbf{\Sigma}$ be the super matrix having σ_i as its ith diagonal elements, and $\mathbf{A}^{-1}$ be a super matrix having n^2 A^{-1} as its all elements. Then,

$$\nu - \lambda(\mathbf{\Sigma} - \mathbf{A}^{-1})\nu = 0 \quad (13)$$

Thus, Equation (7) ($i = 1, ..., n$) can be solved by ν as an eigenvector of $(\mathbf{\Sigma} - \mathbf{A}^{-1})$ in Equation (13).

Method

Survey Procedure

The first step in the simulation study was to collect data on the similarities of 48 nouns that are among the earliest learned by children (Fenson et al, 1993). To determine the relevant similarities across a broad range of properties, 104 Japanese undergraduates rated each noun category using 16 pairs of adjectives (Hidaka & Saiki, 2004). The goal here is to place the categories in a relatively (16 dimensions) large feature space. These adjective pairs are the potential features. Subjects used a 5-point scale to indicate how well the pair of adjectives described the items in the category (e.g., *large* = 5, *small* = 1). The 16 pairs of adjectives were selected by a pilot survey using 41 pairs collected from prior studies. We created questionnaires of 5 different orderings to cancel out the order effect. Participants completed the survey in about an hour.

Stimuli

- **Adjective pairs (linguistic scales)**
  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

Analysis and Simulation

Correction of survey data The rating value was corrected by a logistic function to make the correlation between mean and variance zero. The original rating showed a small positive correlation between the deviation from the median and the variance, because an extreme rating (i.e., a rating near one or five) has a smaller variance than a rating near the median. More specifically, the parameters of the logistic function $f(x) = (1 + \exp((x-b)c^{-1}))^{-1}$ are estimated to have zero correlation between $|x - b|$ and a standard deviation of rating x, and estimated parameters are $b = 3$ and $c = 1.2$. The corrected mean and variance is used for analysis and simulation.

Index of semantic smoothness Semantic smoothness, as predicted by Equation 12, was specifically calculated by norms of the mean vector and covariance matrix in the model. The mean vector and covariance (or correlation) matrix represent the mean and covariance across the 16-adjective ratings for all subjects. The correlation and contribution of the norms of the mean vector and the covariance were used as an index of smoothness. The contribution of the major axis is calculated by the principal component analysis, because the norms of both the mean and the covariance have variances. In other words, the coefficient of determination in the regression analysis underestimates this contribution because it supposes that only the dependent variable has error.

Simulation of packing category Three simulations were run. The first simulation involved the semantic packing of randomly generated categories with the goal of visualizing coherent categories. The second simulation attempted to reproduce category organizations based on the adjective ratings for the corpus of early learned nouns and in doing so demonstrates how packing might explain fast mapping. The third simulation involved a Monte Carlo simulation investigating the relationship among measures of discrimination.

In the semantic packing simulation, we optimized the mean and covariance of several categories with randomly generated initial means and covariances (i.e. the gradient method: updating the parameters based on Equation 13 and 10) . The smoothness index was measured after updating was performed 100 times. The updated final state refers to optimization in terms of balanced constraints.In the simulation of the adjective rating data for early-learned nouns, the means of categories were reproduced by a solution of Equation 13 for a given covariance matrix of survey data. This simulation investigates the predictability of a prototype configuration in real data based on the generalization pattern. The results were evaluated based on the correlation between the distances between all pairs of categories in the reproduced and original prototype configuration. The degrees of freedom of the configuration to be estimated is 752 (the number of categories without pivot of rotation by property dimension $(48-1)\times 16$). In the Monte Carlo simulation, 20 one-dimensional normal distributions with uniform-random means (-3 to 3) and variances (0.2 to 2) were generated 500 times. The maximum (Q), paired minimum ($\sum_{i,j\neq i}\min(P(\theta|c_i),P(\theta|c_j))$), overlap ($F$), category utility ($U$), paired Bhattacheryya bound ($\sum_{i,j\neq i}B_{ij}$) of the generated distributions were calculated theoretically (F, U, and B) or numerically (max and min: integral range from -10 to 10 and sample resolution 0.01), and the correlation was analyzed.

Results

Figure 2 shows the relationship between the mean norms and the correlation norm for the adjective-rating The correlation and contribution are .466 and .733, respectively. The correlation and contribution of the smoothness index using covariance, rather than correlation, are

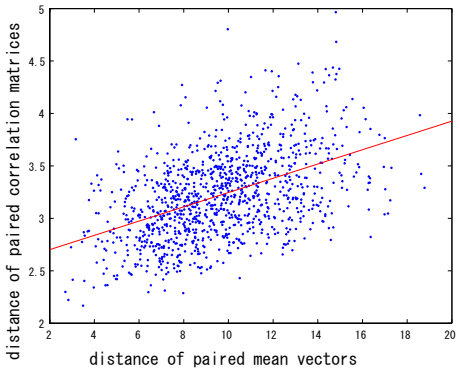

Figure 2: Scatter plot of mean vector norm (x axis: prototype dissimilarity) and correlation matrix norm (y axis: generalization dissimilarity) in survey data

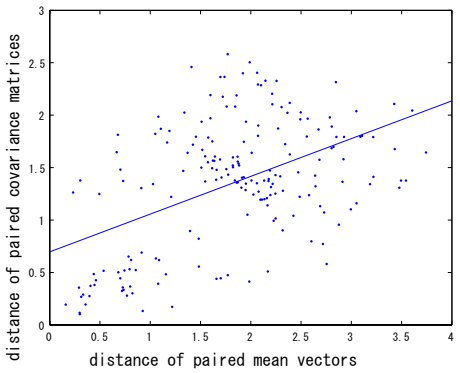

Figure 3: Scatter plot of mean vector norm and correlation matrix norm in simulation

.357 and .688, respectively. These results suggest that the investigated category set has smoothness. The mean norm and covariance norm of paired categories (each category paired with every other category) are shown (Figure 3). The average and standard deviation of the smoothness index for 100 simulations were .490 and .181, respectively. The correlation and contribution between the reproduced mean matrix in the simulation and the mean of survey data were .430 and .715, respectively. These results suggest that semantic packing could reproduce half of the categories *from only their generalization without any knowledge of the category configuration.* This is the kind of result needed to explain feature selection and fast-mapping in children.

The results of Monte Carlo simulation are shown in Table 1. Most of the absolute correlations $|R|$ were greater than .8, which indicated that the five measures were approximately equivalent as proven.

Discussion

The results of Monte Carlo simulation exploring some discriminability measures empirically support the the-

$\tau \backslash R$	max	F	U	B	min
max	1	-.867	.855	-.877	-.863
F	-.681	1	-.791	.938	.806
U	.659	-.601	1	-.903	-.949
B	-.686	.785	-.740	1	.938
min	-.658	.620	-.814	.796	1

Table 1: Peason correlation coefficient R (upper triangle) and Kendall rank correlation τ (lower triangle) among measures

oretical relationship, which is the approximate equivalence among F, cue validity and category. In the packing theory, generalization is not just negative discriminability but a limitation on category representation resulting from the whole memory capacity and a lower bound on generalization. With the two conflicting constraints, the computational model predicts that smooth category organization emerges. Consistent with this prediction, the smoothness index of the adjective-rating data for early learned nouns suggests that property-generalization clusters were formed in not only specific domains (e.g. solidity-shape in the survey of Samuelson & Smith, 1999) but also more generally. Indeed from these data, one can predict regions of "fast mapping" involving property-category organization correlations that are unknown to and unexplored by researchers in early category development. The success in a quantification of category coherence using smoothness index provides empirical evidence of the role of semantic packing in human natural categories. Granted the results could be due to the specific properties and noun categories selected. However, the adjectives were selected by their discriminability (i.e. variance of the adjectives) to the category set (Hidaka & Saiki, 2004), and smoothness can be observed in feature space with discriminability. Therefore, the adjective-rating results may be taken as making predictions about the feature space as it relates to early acquired noun categories (i.e. basic categories).

The success in reproducing the organization of early learned nouns suggests that a category system constrained by semantic packing principle could generalize a category to new instances without trial and error. This implies that the system would "know" the generalization pattern of a novel thing in a certain region of feature space. Young children show precisely this kind of knowledge in generalizing names for novel categories. This is typically referred to in the developmental literature as "fast mapping." Notice that the system has no meta-knowledge , as theory-theory claims, about specific domains, but smoothness, a property of the whole system in which categories are learned and represented, does the work of such meta-knowledge.

In summary, the proposed theory suggests that category packing process, balanced discriminability and generalization (i.e. basic categories), leads smooth category-feature organization (i.e. category coherence) consistent to human natural categories, and the smoothness help learners' generalization to novel categories (i.e. fast mapping).

Acknowledgments

This work was supported by grants from Grants-in-Aid for Scientific Research from JMEXT (No. 15650046), JSPS Research Fellowships for Young Scientists, Kyoto University Foundation and the National Institutes of Mental Health, R01 MH60200-06.

References

Corter J. E., & Gluck M. A. (1992). Explaining Basic Categories: Feature Predictability and Information. *Psychological Bulletin*, *111*, 291–303.

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.

Gelman, R., & Williams, E. (1997). Enabling constraints on cognitive development. In D. Kuhn & R. S. Siegler (Ed&.). Cognition, perception and language. Vol. 2. Handbook of child development. (5th ed.) (pp. 575-630). (W. Damon, Ed.) New York: Wiky.

Gluck, M. A. & Corter, J. E. (1985). Information, Uncertainty, and the Utility of Categories, , *Proceedings of the Seventh Annual Conference of the Cognitive Science Society* (pp. 283–287). CA: Lawrence Erlbaum Associates.

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

Jones, S.S. (2003). Late talkers show no shape bias in object naming. *Developmental Science*, *6(5)*, 477–483.

Murphy, G. L., & Lassaline, M. E. (1997). Hierarchical Structure in Concepts and the Basic Level of Categorization. in Lamberts, K. & Shancks, D. (Eds), *Knowledge concepts and categories*, UK: Psychology Press.

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

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*, 382–407.

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

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

Soja, N. N., Carey, S. & Spelke, E. S. (1991). Ontological categories guide young children's inductions of word meanings: object terms and substance terms., *Cognition*, *38*, 179–11.