Constructing a hypothesis space from the Web for large-scale Bayesian word learning

Joshua T. Abbott (joshua.abbott@berkeley.edu)
Joseph L. Austerweil (joseph.austerweil@gmail.com)
Thomas L. Griffiths (tom_griffiths@berkeley.edu)
Department of Psychology, University of California, Berkeley, CA 94720 USA

Abstract

The Bayesian generalization framework has been successful in explaining how people generalize a property from a few observed stimuli to novel stimuli, across several different domains. To create a successful Bayesian generalization model, modelers typically specify a hypothesis space and prior probability distribution for each specific domain. However, this raises two problems: the models do not scale beyond the (typically small-scale) domain that they were designed for, and the explanatory power of the models is reduced by their reliance on a hand-coded hypothesis space and prior. To solve these two problems, we propose a method for deriving hypothesis spaces and priors from large online databases. We evaluate our method by constructing a hypothesis space and prior for a Bayesian word learning model from WordNet, a large online database that encodes the semantic relationships between words as a network. After validating our approach by replicating a previous word learning study, we apply the same model to a new experiment featuring three additional taxonomic domains (clothing, containers, and seats). In both experiments, we found that the same automatically constructed hypothesis space explains the complex pattern of generalization behavior, producing accurate predictions across a total of six different domains.

Keywords: generalization; concept learning; word learning; Bayesian modeling; online databases

Introduction

Many problems solved by the mind conform to the same abstract computational formulation: How should a property be generalized to novel stimuli from a set of stimuli observed to have the property? As there are many ways to extend the property that are consistent with some observed evidence, these are problems of *induction*, where the evidence constrains, but does not determine, the solution to a problem. The Bayesian generalization framework (Shepard, 1987; Tenenbaum & Griffiths, 2001) has been remarkably successful at explaining human generalization behavior in a wide range of domains. However, its success is largely dependent on the choice of a hypothesis space and a prior probability distribution on hypotheses, which are usually hand constructed by the researcher for each specific problem. This is unsatisfying practically, because the models do not scale beyond the originally modeled problem, and theoretically, as it is unclear whether their success is due to the cleverness of the modeler and not because of a deep mathematical property of the computational problem that people solve.

One possible solution is to use existing sources of information about the organization of a domain as the basis for specifying a hypothesis space and prior. This helps address both the practical and the theoretical concerns raised by the Bayesian generalization model. In this paper, we use this approach to show how a hypothesis space and prior can be constructed automatically from a large online database, making it possible to apply the Bayesian generalization framework to a wide range of naturalistic stimuli. We focus on one specific generalization problem, word learning, where people learn new words from observing a few objects that can be labeled with that word. Given that the number of possible extensions of a word is essentially infinite, learning the objects referred to by a word is a very difficult inductive problem (Quine, 1975). Xu and Tenenbaum (2007) showed how the Bayesian generalization framework could be used to explain how people learn new words. However, to construct the hypothesis space of their Bayesian model, Xu and Tenenbaum (2007) elicited approximately 400 similarity judgments from their participants. Clearly this is not practical to extend into every domain where people learn words. Thus, word learning is an appropriate setting for exploring novel methods of constructing hypothesis spaces and prior distributions.

We propose a method for automatically constructing the hypothesis space and prior distribution of a Bayesian word learning model using freely available online resources. In particular, we use WordNet (Fellbaum, 2010; Miller, 1995) as an initial source for automatically creating the hypothesis space, and ImageNet (Deng et al., 2009) as a source of naturalistic images that can be used as stimuli to test the resulting model in behavioral experiments. WordNet is a popular lexical database of English comprised of over 100,000 relational sets of synonyms. ImageNet is a large ontology of images conforming to the hierarchical structure of WordNet, with the aim of providing over 500 high-quality images per noun in WordNet. These resources allow us to construct hypothesis spaces and prior distributions for word learning without eliciting a single judgment from participants and test the resulting model on a much larger scale than was previously possible. We demonstrate that the Bayesian model formulated from WordNet captures participant judgments in two behavioral experiments, addressing the practical and theoretical issues with Bayesian models discussed earlier.

The plan of the rest of the paper is as follows. In the next sections we review the Bayesian generalization model and then examine how Xu and Tenenbaum (2007) constructed the hypothesis space for their Bayesian word learning model. We then show how to build a hypothesis space from WordNet that can be used to evaluate word learning models on a large scale. Afterwards, we present two experiments utilizing this hypothesis space: one that replicates a previous study of adult word learning, and one that investigates word learning for a set of complex concepts in novel domains. Finally, we discuss the implications of our work and future directions for research.

The Bayesian Generalization Framework

The Bayesian word learning model is a special case of the Bayesian generalization framework. This framework has been used to model generalization in a number of domains including dimensional concepts (Austerweil & Griffiths, 2010; Shepard, 1987; Tenenbaum, 1999), word learning (Xu & Tenenbaum, 2007), numerical concepts (Tenenbaum, 2000), sequential rules (Austerweil & Griffiths, 2011) and rule-based categorical concepts (Goodman, Tenenbaum, Feldman, & Griffiths, 2008). Typically, problems are formulated in this framework as follows: Assume we observe n positive examples $\mathbf{x} = \{x_1, \ldots, x_n\}$ of concept C and want to compute $P(y \in C|\mathbf{x})$, the probability that some new object y belongs to C given the observations $\mathbf{x}$. We compute this probability by using a hypothesis space $\mathcal{H}$, which is a set of hypothetical concepts, where each hypothesis is defined by the objects that would be members of the concept if the hypothesis were true, $P(\mathbf{x}|h)$.

Defining a Bayesian generalization model amounts to defining a hypothesis space $\mathcal{H}$, a prior probability distribution over hypotheses, $P(h)$, and for each hypothesis, a likelihood function, $P(\mathbf{x}|h)$, indicating the probability of observing a set of objects x given that the hypothesis is true. A typical definition of the likelihood follows from assuming strong sampling, where objects are generated uniformly at random from the true hypothesis (Tenenbaum & Griffiths, 2001)

$$P(\mathbf{x}|h) = \begin{cases} 1/|h|^n & \text{if } \mathbf{x} \subset h \\ 0 & \text{otherwise} \end{cases}. \tag{1}$$

This likelihood function instantiates the *size principle* for scoring hypotheses: hypotheses containing a smaller number of objects assign greater likelihood than hypotheses with more objects to the same set of objects (Tenenbaum, 1999; Tenenbaum & Griffiths, 2001). The prior distribution over hypotheses, $P(h)$ depends on the domain and in previous literature has ranged from a simple uniform distribution over the hypothesis space (Shepard, 1987) to a stochastic process over tree structures (Kemp & Tenenbaum, 2009). Given the prior and likelihood, the posterior probability that a hypothesis is true given a set of objects belonging to a novel concept, $P(h|\mathbf{x})$, follows from Bayes' rule: $P(h|\mathbf{x}) \propto P(\mathbf{x}|h)P(h)$. From this, we can compute the probability that a new object y is also a member of the concept C by averaging the predictions of all hypotheses weighted by their posterior probabilities:

$$P(y \in C|\mathbf{x}) = \sum_{h \in \mathcal{H}} P(y \in C|h)P(h|\mathbf{x}), \tag{2}$$

where $P(y \in C|h) = 1$ if the new object y is in hypothesis h, and 0 otherwise.

Word Learning as Bayesian Inference

Xu and Tenenbaum (2007) derived the hypothesis space for their Bayesian word learning model by applying hierarchical clustering (see Duda & Hart, 1973) to the perceived similarity of every pair of objects. The hypothesis space, prior and likelihood are defined by the tree resulting from hierarchical clustering. Using a tree is well justified from a psychological perspective as children assume the possible referents of novel nouns are tree-structured (Markman, 1991). Nodes in the tree represent potential words (hypotheses) which extend to all the leaves they cover, where the leaves of the tree correspond to the domain of possible objects. The height of a node h (minimal distance from the node to a leaf) is a measure of the average pairwise dissimilarity of objects covered by node h and approximates the heterogeneity of the objects that can be called that word. The intuition that more distinctive clusters are more likely to have distinguishing names, was incorporated by defining the prior $P(h)$ to be proportional to the branch length separating node h from its parent:

$$P(h) \propto \text{height}(\text{parent}(h)) - \text{height}(h), \tag{3}$$

where parent(h) returns the parent of node h. To incorporate a *basic-level* bias (Markman, 1991; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976) in which new words tend to refer more often to a word at an intermediate level in a taxonomy, the prior probability of hypotheses at the basic level were 10 times the value given by Equation 3 (see below for examples). As the height of node h also approximates the number of objects in the extension of the possible word h, the likelihood of observing n objects called word h is defined as

$$P(\mathbf{x}|h) \propto \left[\frac{1}{\text{height}(h) + \varepsilon}\right]^n, \tag{4}$$

where ε is a small constant so that the leaf hypotheses (those that refer to only a single object) do not have infinite likelihood (as their height is zero).

Using this framework, Xu and Tenenbaum (2007) accurately predicted how people extend words to new objects depending on the diversity and number of objects labeled with that word. In a set of experiments on both adults and children, they showed participants one or more positive examples of a novel word while manipulating the taxonomic relationship of the objects the word referred to. For example, participants might observe one Dalmatian, three Dalmatians (exemplars at the subordinate-level), a Dalmatian, terrier, and mutt (exemplars at the basic-level), or a Dalmatian, pig, and toucan (exemplars at the superordinate-level) being labeled with a novel word (e.g. "fep"). After observing a word refer to one or three example objects at the subordinate, basic, or superordinate-level, they were asked whether the word referred to novel subordinate, basic, superordinate, and out-of-domain objects.

When participants were given one example of an object that refers to a word (e.g. one Dalmatian), they tended to select the subordinate-level matches (e.g. the two other Dalmatians) and the basic-level matches (e.g. the two non-Dalmatian dogs). However, when they were shown three subordinate-level examples of a concept (e.g. three Dalmatians), the participants tended to choose only the subordinate-level matches (e.g. they only believed the word referred to the two other Dalmatians). The Bayesian word learning model captured this phenomenon because the prior favors words at the basic-level, but the likelihood favors words at the subordinate-level, and the likelihood's weight increases exponentially in the number of objects.

Unfortunately, the manner in which the hypothesis space was constructed (through hierarchical clustering on pairs of similarity judgments) poses a serious constraint to assessing the model's validity. To construct the hypothesis space in the three domains tested by Xu and Tenenbaum (2007), where there are 15 images per concept, each participant had to provide roughly 400 similarity judgments. To test how well this framework extends to new concepts and domains using their method for constructing the hypothesis space, an impractically large quantity of human judgments would need to be elicited. In the following section, we introduce an alternative method of constructing a hypothesis space for the Bayesian word learning model, which allows for testing the framework without eliciting any judgments from participants.

Large-Scale Word Learning

Using an online word ontology, we can automatically construct the hypothesis space of a Bayesian word learning model. WordNet is a large lexical database of English represented as a network of words linked by directed edges denoting semantic relatedness (Fellbaum, 2010; Miller, 1995). Its structure was manually designed to group lexical concepts in an "is-a" hierarchy based on the many-to-one mapping of synonyms. For example, a Poodle "is-a" type of dog, thus WordNet has a directed edge from the node for *dog* to the node for *Poodle*. As WordNet is hierarchically structured like the hypothesis space used by Xu and Tenenbaum (2007), it is an ideal candidate for constructing our hypothesis space.

Using a hypothesis space derived from WordNet, we can better test the predictions of different generalization theories for word learning by examining their predictions for a large range of concepts. In the rest of this section, we present the method used to construct a hypothesis space from WordNet and outline the implementations of three generalization models using this hypothesis space for large-scale word learning.

Constructing a Hypothesis Space

In the context of the Bayesian generalization framework, the hypotheses correspond to subsets of the universe of objects that are psychologically plausible candidates as extensions of concepts (Tenenbaum & Griffiths, 2001). Using WordNet as the basis of our hypothesis space, the set of objects is the set of leaf nodes from the noun-space of the directed graph and the hypotheses correspond to both the inner nodes of the directed graph and the leaf nodes, which distinguish between objects at the subordinate-level. To construct a hypothesis space from WordNet, we first extracted a tree from the 82,115 noun nodes of WordNet.[1] The nodes are hypotheses, which represent possible words, and form the hypothesis space for the model. From this graph we create a hypothesis space that is a binary matrix, $\mathcal{H}$, whose rows are the objects (64,958 leaf nodes from the graph) and columns are the hypotheses (82,115 nodes, 17,157 of which are inner nodes and 64,958 are leaf nodes). Each entry (i, j) of the matrix $\mathcal{H}$ denotes whether or not hypothesis node j is an ancestor of leaf node i in the WordNet graph (with a 1 indicating it is). The leaf nodes are included as hypotheses so that the model distinguishes between subordinate objects.

Generalization Models

With a hypothesis space derived from WordNet, we now have the ability to test the Bayesian model of word learning on a much larger scale. In addition, we can use the hypothesis matrix as a feature space for testing alternative models. We compare the Bayesian model against two similarity models: a prototype model and an exemplar model. Given a set of examples $\mathbf{x} = \{x_1, \ldots, x_n\}$ representing some concept C (where the elements of $\mathbf{x}$ correspond to rows in the hypothesis matrix $\mathcal{H}$), we can compute a score for each row $y \in \mathcal{H}$ denoting the probability that y is also a member of C. We present the different ways to compute this score below.

Bayesian model. This is the Bayesian generalization framework that we discussed earlier. We used strong sampling for the likelihood, $P(\mathbf{x}|h)$, which is computed via Equation 1, where the size of h is the number of nodes that can be reached by a directed path from h. This simply corresponds to the sum of the elements in the column corresponding to h.

The prior $P(h)$ was defined to be Erlang distributed in the size of the hypothesis (a standard prior over sizes in Bayesian models; Shepard, 1987; Tenenbaum, 2000)

$$P(h) \propto (|h|/\sigma^2)\exp\{-|h|/\sigma\}, \qquad (5)$$

where the σ parameter was set to 200 by hand fitting the model predictions to all human responses (the same value was used in both experiments). This value favors medium sized hypotheses, which is roughly equivalent to a basic-level bias. The probability that word C extends to object y after observing a set of objects called C is

$$\text{Bscore}(y) = P(y \in C|\mathbf{x}) = \sum_{h \in \mathcal{H}} P(y \in C|h)P(h|\mathbf{x}), \qquad (6)$$

where $P(y \in C) = 1$ if $y \in h$ and 0 otherwise, and $P(h|\mathbf{x})$, is the posterior distribution over hypotheses.

Prototype model. In this model, we define the prototype of a set of objects, x_{proto}, to have those features owned by a majority of the objects in the set. The generalization measure for

[1] Technically WordNet is a directed acyclic graph because some nodes have multiple parents (the method still works in these cases).

an object y is

$$\text{Pscore}(y) = \exp\{-\lambda_p \, \text{dist}(y, x_{\text{proto}})\}, \tag{7}$$

where $\text{dist}(\cdot,\cdot)$ is the Hamming distance between the two vectors and λ_p is a free parameter (for all of the results presented here, $\lambda_p = 0.15$, optimized by hand using half-interval search). Pscore was then normalized over all objects y in the hypothesis space (all leaf nodes).

Exemplar model. We define the exemplar model using a similar scoring metric as the prototype model, except rather than computing the distance of object y to a single prototype vector, we compute a distance for each item x_j in the set of observations $\mathbf{x}$. The exemplar generalization measure is thus computed as

$$\text{Escore}(y) = \sum_{x_j \in \mathbf{x}} \exp\{-\lambda_e \, \text{dist}(y, x_j)\}, \tag{8}$$

where $\text{dist}(\cdot,\cdot)$ is the Hamming distance between two vectors and λ_e is a free parameter (for all of the results presented here, $\lambda_e = 0.20$, optimized by hand using half-interval search). Escore was then normalized over all objects y in the hypothesis space (all leaf nodes).

Behavioral Experiments

To evaluate the performance of our models using the WordNet-based hypothesis space, we conducted two experiments using the paradigm of Xu and Tenenbaum (2007). The first experiment replicates Xu and Tenenbaum (2007) on their three object taxonomies (animals, vehicles, and vegetables), which validates our approach for constructing a hypothesis space from WordNet and using images from ImageNet as stimuli. The second experiment extends the paradigm into three previously unexplored domains (clothing, containers, and seats), which have hierarchical structure, but it is not as clear how well they conform to a natural basic-level taxonomy (Rosch et al., 1976).

Experiment 1: Validating Our Approach

Participants. Thirty four participants were recruited via Amazon Mechanical Turk and compensated $0.05 for each trial (training set) completed out of twelve possible. Each participant completed as many trials as he or she wished, and twenty unique participants completed each trial. All participant responses were used.

Stimuli and Procedure. Within each taxonomy, the stimuli consisted of the images of objects distributed across the superordinate, basic and subordinate-levels, and subsequently split into training and test sets. The training sets were the labeled objects given to participants of which there were four conditions: a single subordinate-level example (e.g. a Dalmatian); three examples of the same subordinate-level object (e.g. three Dalmatians); the subordinate-level object and two basic-level objects (e.g. a Dalmatian, a Shih Tzu, and a Beagle); and the subordinate object and two superordinate-level objects (e.g. a Dalmatian, a hippopotamus, and a toucan). This corresponds to twelve trials total (four conditions for each of the three object taxonomies).

The test sets were the same regardless of the training set and consisted of eight objects matching the currently tested taxonomy: two subordinate examples (e.g. two other Dalmatians); two basic-level examples (e.g. a Cocker Spaniel and a Corgi); and four superordinate examples (e.g. a cat, a bear, a sea lion, and a horse). There were also sixteen non-matching objects in the test set corresponding to the objects that match the two other taxonomies.

For each trial, participants were instructed that they needed to help a cartoon frog who speaks a different language from us, pick out objects that he wants. The frog shows one or more examples of a novel word (e.g. "dak") and the participant is instructed to select other items that are a "dak" from the objects comprising the test set. A unique novel word was associated with each of the twelve trials.

Results. Figure 1 shows the results of this experiment, along with the predictions of the different generalization models. For each training set condition, the data for each test item has been averaged over participants and domains. The generalization judgments of participants (left-most panel of Figure 1) follows the same qualitative trend as those reported in Xu and Tenenbaum (2007). There is a sharp drop in generalization to basic-level objects when seeing only a single subordinate example compared to the condition when seeing three subordinate examples.

The Bayesian model predictions (second panel from the left) exhibits this same generalization pattern ($r^2 = 0.98$), while the prototype and exemplar models do not ($r^2 = 0.66$ and $r^2 = 0.84$, respectively). This validates our method of automatically creating hypothesis spaces with WordNet.

Experiment 2: Novel Domains

Participants. Thirty six participants were recruited via Amazon Mechanical Turk and compensated $0.05 for each trial completed out of twelve possible. As in Experiment 1, each participant completed as many trials as he or she wished, and twenty unique participants completed each trial. All participant responses were used.

Stimuli and Procedure. Table 1 contains the objects we used for training in the three hierarchical domains (clothing, containers, and seats).[2] As in Experiment 1, the same test objects were used for every training set, and the "non-match" test objects were the objects in the test set which match the two other taxonomies that are not contained in the training set. As before, this corresponds to twelve trials total. The procedure was identical to Experiment 1.

[2] The additional subordinate-level training image and the test images were omitted from Table 1 for brevity.

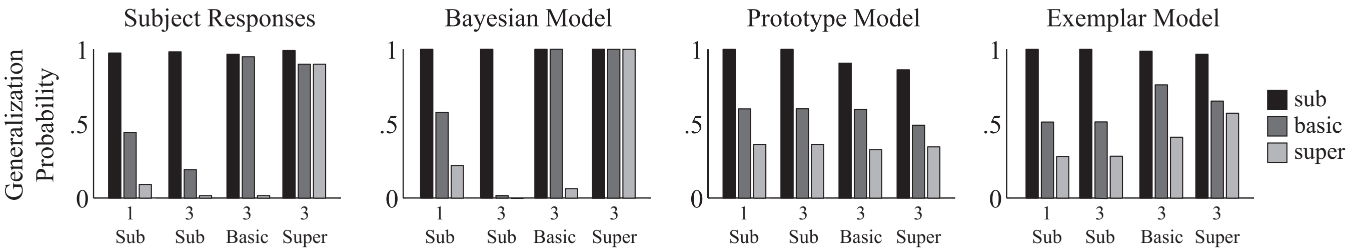


Figure 1: Participant generalization judgments and predictions of the Bayesian, prototype, and exemplar models averaged across the three domains in Experiment 1. The generalizations for non-matching items are omitted for brevity (neither the participants chose nor the Bayesian model predicted non-matching objects, while the prototype and exemplar models predicted non-matches less than 4% of the time for each condition).

Results. Figure 2 presents the averaged results of how participants [3] and the Bayesian model generalized the learned words to the test objects based on the observed training set across the different domains in Experiment 2.[4] Across the three domains, the generalization probabilities of the participants and Bayesian model with the same parameters are extremely similar. This is exemplified in the very good quantitative model fit on the averaged data ($r^2 = 0.95$). Furthermore, the hypothesis space constructed automatically from WordNet explains the idiosyncrasies of participant generalization behavior in each domain ($r^2 = 0.97, 0.88$, and 0.91, for clothing, containers, and seats respectively). For example, the model accurately predicts that participants would generalize most broadly in the seats domain for the single exemplar and three basic-level exemplar training sets. Additionally, the model captures that people generalized the least in the containers domain for the three subordinate-level exemplar training sets. This would not have been possible if the hypothesis space for each domain had the same structure.

Note that there is a larger amount of variance between model predictions and human performance in Experiment 2 than Experiment 1. We believe that this is due to the domains not conforming to a natural taxonomy. For example, it is unclear if box should be the basic-level category for a mail box and a cigar box; however, this is the basic level of these objects provided by WordNet. Regardless, the good quantitative fit of the Bayesian model's predictions provides evidence that using WordNet as a hypothesis space for word learning can capture people's generalizations even for hierarchies without clearly defined basic-level concepts.

Discussion

Although the Bayesian generalization framework has been extremely successful in explaining human generalization behavior, the hypothesis spaces are typically hand-constructed, which is unsatisfying. In this paper, we explored automatically constructing the hypothesis space using an online resource as a potential solution to the methodological challenges posed by this problem. In the first behavioral experiment, we validated that the Bayesian model using this hypothesis space can capture previously found word learning phenomena. In the second behavioral experiment, we showed that the same Bayesian model explains how participants learned words in three novel domains. Using the automatically constructed hypothesis space, the model predicted the subtle changes in participants' word learning behavior across three domains, thus demonstrating the practical and theoretical benefits of our approach.

In the future, we hope to perform a large scale empirical test of the Bayesian word learning model using more heterogeneous training sets (e.g. one subordinate-level and one basic-level object) and more domains with varied conceptual structure. The larger set of empirical results would enable us to perform a more detailed investigation of the prior knowledge over the types of conceptual structures that people use when they learn words (e.g. do people prefer shallow or deep taxonomies?). Additionally, we hope to incorporate how participant behavior is affected by the visual similarity of the images in the training and tests sets (and its interaction with conceptual structure), which at the moment would not be possible to explore with the Bayesian word learning model.

As word learning is a special case of the more general problem of generalization, our approach potentially could be applied to automatically construct hypothesis spaces for generalization problems in other domains. For example, a Bayesian model of commonsense reasoning could be formulated by automatically deriving hypothesis spaces from ConceptNet (Liu & Singh, 2004) or OpenCyc (Matuszek, Cabral, Witbrock, & DeOliveira, 2006). This follows a development in modern machine learning, which has leveraged online resources to make more successful learning algorithms (Medelyan, Legg, Milne, & Witten, 2009; Ponzetto & Strube, 2006). We hope that this draws a closer connection between computer science and cognitive science, which can lead to more psychologically valid, yet still scalable, artificial intelligence systems.

Acknowledgments. This work was supported by the DARPA BOLT contract HR0011-11-2-0009 and grant number IIS-0845410 from the National Science Foundation.

[3] Non-matching objects were only chosen twice (both in the containers domain) and so, they were omitted from Figure 2.

[4] The prototype and exemplar models were omitted from Figure 2 for brevity ($r^2 = 0.80$ and $r^2 = 0.90$ averaged over domains, respectively).

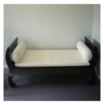
	Clothing		Containers		Seats	
Object level	1	2	1	2	1	2
Subordinate						
Basic						
Superordinate						

Table 1: Training images for Experiment 2.

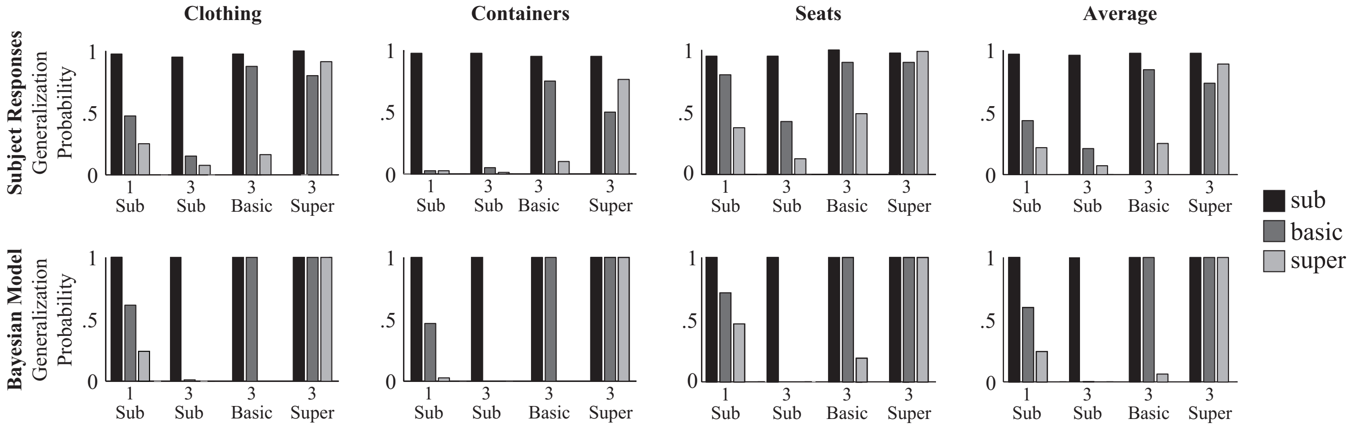


Figure 2: Participant generalization judgments and the predictions of the Bayesian model for Experiment 2. From left to right, the columns present the results for the three taxonomies (clothing, containers, and seats) and average results.

References

Austerweil, J. L., & Griffiths, T. L. (2010). Learning hypothesis spaces and dimensions through concept learning. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 73–78). Austin, TX: Cognitive Science Society.

Austerweil, J. L., & Griffiths, T. L. (2011). Seeking confirmation is rational for deterministic hypotheses. *Cognitive Science*, *35*, 499–526.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255).

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis*. New York: Wiley.

Fellbaum, C. (2010). WordNet. *Theory and Applications of Ontology: Computer Applications*, 231–243.

Goodman, N. D., Tenenbaum, J. B., Feldman, J., & Griffiths, T. L. (2008). A rational analysis of rule-based concept learning. *Cognitive Science*, *32*(1), 108–154.

Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.

Liu, H., & Singh, P. (2004). ConceptNet - A practical commonsense reasoning tool-kit. *BT Technology Journal*, *22*(4), 211-226.

Markman, E. M. (1991). *Categorization and naming in children: Problems of induction*. MIT Press.

Matuszek, C., Cabral, J., Witbrock, M., & DeOliveira, J. (2006). An introduction to the syntax and content of cyc. In C. Baral (Ed.), *Proceedings of the AAAI 2006 Spring Symposium* (p. 44-49). Menlo Park, CA: AAAI Press.

Medelyan, O., Legg, C., Milne, D., & Witten, I. H. (2009). Mining meaning from Wikipedia. *International Journal of Human-Computer Studies*, *67*(9), 1-76.

Miller, G. (1995). WordNet: a lexical database for english. *Communications of the ACM*, *38*(11), 39–41.

Ponzetto, S. P., & Strube, M. (2006). Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of the HLT Conference of the NAACL* (p. 192-199).

Quine, W. V. O. (1975). *Word and object*. MIT Press.

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.

Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317-1323.

Tenenbaum, J. B. (1999). Bayesian modeling of human concept learning. In M. S. Kearns, S. A. Solla, & D. A. Cohn (Eds.), *Advances in Neural Information Processing Systems 11* (Vol. 11, p. 59-65). Cambridge, MA: MIT Press.

Tenenbaum, J. B. (2000). Rules and similarity in concept learning. In S. A. Solla, T. K. Leen, & K.-R. Muller (Eds.), *Advances in Neural Information Processing Systems 12* (Vol. 12, pp. 59–65).

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences*, *24*(4), 629–640.

Xu, F., & Tenenbaum, J. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.