

Category and feature identification

Charles Kemp

Department of Psychology
Carnegie Mellon University

Kai-min K. Chang

Language Technologies Institute
Carnegie Mellon University

Luigi Lombardi

Dipartimento di Scienze della Cognizione e della Formazione
Università di Trento

Abstract

This paper considers a family of inductive problems where reasoners must identify familiar categories or features on the basis of limited information. Problems of this kind are encountered, for example, when word learners acquire novel labels for pre-existing concepts. We develop a probabilistic model of identification and evaluate it in three experiments. Our first two experiments explore problems where a single category or feature must be identified, and our third experiment explores cases where participants must combine several pieces of information in order to simultaneously identify a category and a feature. Humans readily solve all of these problems, and we show that our model accounts for human inferences better than several alternative approaches.

Keywords: inductive reasoning; identification; Bayesian models; information integration; semantic cognition; naming to description

Suppose that you are watching a German nature program and that you pick up enough of the narrative to learn that a *Schmetterling* is colorful, has wings, and has antennae. Can you guess what a *Schmetterling* might be? Similarly, suppose that you learn that zebras and tigers are both *gestreifet*. Can you guess what *gestreifet* might mean? We will refer to both of these problems as *identification* problems. In the first case, you

This work was supported in part by the National Science Foundation under award number CDI-0835797. We thank Faye Han and Bobby Han for running our experiments and coding the data, Simon De Deyne for providing us with the listing matrix L , and Michael Lee, Daniel Navarro, Amy Perfors and Tim Rogers for valuable suggestions. We also thank Gert Storms and his entire group for creating and sharing the Leuven natural concept database.

need to identify a category—namely, *butterfly*. In the second case, you need to identify a feature—namely, *striped*. Problems like these draw on semantic knowledge about animals and their features, and this paper will consider how this knowledge can be used to address identification problems.

As our opening examples suggest, category identification and feature identification are problems regularly faced by second-language learners. In many cases these learners will already have concepts like *butterfly* and *striped*, and their task is to map novel labels onto these concepts. Identification, however, may play an equally critical role in first-language acquisition. Before learning her first few words, a child may already have formed a category that includes creatures like the furry pet kept by her parents, and learning the word “cat” may be a matter of attaching a new label to this pre-existing category (Fodor, 1975; Mervis, 1987; Chomsky, 1991). Bloom (2000) summarizes this proposal by suggesting that “much of what goes on in word learning is establishing a correspondence between the symbols of a natural language and concepts that exist prior to, and independently of, the acquisition of that language” (p 242).

This paper develops a probabilistic framework that can address a broad family of identification problems. Like all inductive problems, identification problems can only be solved if a learner relies on background knowledge, and our approach offers a formal characterization of the knowledge that guides category and feature identification. We propose that this knowledge is stored in a semantic repository that includes information about the relationship between categories and features (for instance, butterflies have wings) along with information about the frequency with which different categories and features are encountered (a random speaker is more likely to refer to dogs or cats than to chameleons or llamas). We make these ideas concrete by describing a repository built from the Leuven natural concept database (De Deyne et al., 2008).

Prior knowledge plays a critical role in inductive reasoning, but this knowledge must be combined with evidence in order to solve inductive problems. Often multiple pieces of evidence are available, and a reasoner must integrate all of this information. Several accounts of information integration can be found in the psychological literature (N. H. Anderson, 1981), and different approaches combine multiple pieces of evidence by adding (Lombardi & Sartori, 2007), multiplying (Medin & Schaffer, 1978; Oden & Massaro, 1978) or taking the maximum (Osherson, Smith, Wilkie, Lopez, & Shafir, 1990) of a set of numerical scores. We will argue that probabilistic inference provides a principled account of information integration that avoids arbitrary choices of functions like sums and products.

The inductive problems we consider and the modeling approach we pursue both build on previous contributions to the psychological literature. The problem of identification is related to the work of Lombardi and Sartori (2007, see also Sartori and Lombardi, 2004) who developed a computational account of category identification that is known as the additive relevance model. These authors report that their model performs better than a simple Bayesian alternative, but their analysis was based on sparse feature matrices that may not adequately capture what people actually know about categories and their features. Our results suggest that a Bayesian account of category identification performs better than the additive relevance approach when both are supplied with a semantic repository that better captures the knowledge that people bring to the problem.

Several psychologists have developed probabilistic models of inductive reason-

Problem	Form	Example	Example Response
Category identification	Cs have $\{f_1, \dots, f_n\}$.	Cs have stripes.	$C = zebra$
Feature identification	$\{c_1, \dots, c_m\}$ have F.	Rabbits have F.	$F = long\ ears$
Joint identification	$\{c_1, \dots, c_m\}$ have F.	Rabbits have F.	$F = fur$
	Cs have F. Cs have $\{f_2, \dots, f_n\}$.	Cs have F. Cs have stripes.	$C = tiger$

Table 1: Three identification problems. Each problem asks a reasoner to identify a category C, a feature F, or a category and a feature.

ing (Shepard, 1987; J. R. Anderson, 1990; Heit, 1998) and our approach continues within this general tradition. Of the many probabilistic models that have been developed, our approach is related most closely to models of categorization (J. R. Anderson, 1991) and generalization (Kemp & Tenenbaum, 2009) that attempt to explain how inferences about novel objects and properties are guided by semantic knowledge. The identification problems we consider are somewhat different, but our approach is consistent with the idea that probabilistic inference is a domain-general principle that helps to explain how humans solve many inductive problems. Although we focus throughout on identification, we return to the relationship between identification and other inductive problems in the General Discussion.

A probabilistic account of category and feature identification

This paper will focus on the three identification problems in Table 1. Each problem consists of a list of statements about animal categories and their features, and each list includes a hidden category C, a hidden feature F or a hidden category and a hidden feature. In each case the task of the reasoner is to identify the hidden items. Although the problems in Table 1 are simple enough to be experimentally tractable, they are inspired in part by the real-world inductive challenge faced by first- and second-language learners. In real world identification problems, the hidden category or feature will typically be introduced as an unfamiliar component of a linguistic utterance (e.g. *Punda milia* have stripes), and the task of the learner is to identify the meaning of this novel word or phrase.

This paper will develop a unified probabilistic model that addresses all three of the problems in Table 1. For each of these problems, our model specifies a probability distribution over the values of the hidden items given the items that have been observed. We propose that humans choose categories and features that have high probability according to these distributions.

To formally specify these distributions we take a generative approach. More precisely, we specify a probabilistic procedure for generating identification problems like the examples in Table 1. Suppose that we start with a semantic repository that captures knowledge about animal categories and their features. We will specify a procedure that samples a list of statements from this repository, including, for example, the statement that “zebras have stripes.” We now assume that some of the categories and features in the sampled statements are hidden—for example, “zebras have stripes” might become “Cs have stripes.” Given this procedure for generating identification problems, we can now use Bayesian inference to work

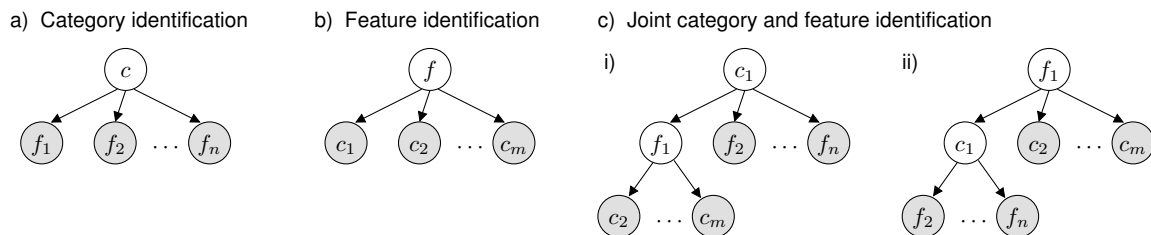


Figure 1. Generative approaches to the three identification problems in Table 1. Shaded nodes indicate variables that are observed, and unshaded nodes indicate variables with values that must be inferred. (a) Category identification. The observed features are assumed to be independently generated from the distribution $p(f|c)$. (b) Feature identification. The observed animal categories are generated from the distribution $p(c|f)$. (c) Joint category and feature identification. This paper evaluates the approach in (ii), which assumes that the hidden feature is generated before the hidden category.

backwards and identify the hidden items in any given problem.

The semantic repository plays a critical role in this approach and must specify two kinds of distributions. First, it must specify a prior distribution $p(c)$ over categories and a prior distribution $p(f)$ over features. These distributions can capture factors like the familiarity of a category and the frequency with which a feature is thought about. For example, in most contexts a familiar category like *dog* should receive higher prior probability than a category like *chameleon*. The semantic repository must also provide two additional distributions: $p(f|c)$ which specifies the probability that a person will choose f when asked to list a feature of category c , and $p(c|f)$ which specifies the probability that a person will chose c when asked to list an animal category that has feature f . For example, $p(\textit{barks}|\textit{dog})$ should be greater than $p(\textit{breathes air}|\textit{dog})$, since *barks* is the more characteristic feature of dogs, and $p(\textit{breathes air}|\textit{dog})$ should be greater than $p(\textit{has wings}|\textit{dog})$, since dogs breathe air but do not have wings. Similarly, $p(\textit{robin}|\textit{has wings})$ should be greater than $p(\textit{penguin}|\textit{has wings})$, which in turn should be greater than $p(\textit{dog}|\textit{has wings})$.

The four distributions $p(c)$, $p(f)$, $p(c|f)$ and $p(f|c)$ can be used to generate many kinds of identification problems. Here we focus on three problems that we refer to as category identification, feature identification, and joint category and feature identification.

Category identification

The first problem in Table 1 requires a reasoner to identify an animal category given one or more features of the category. For example, the reasoner might be informed that “Cs have stripes and hooves” and asked to identify Category C. As shown in Figure 1a, we assume that problems of this kind are generated by sampling a category c (here $c = \textit{zebra}$) from the prior distribution $p(c)$, then sampling n features from the distribution $p(f|c)$. As a final step, the value of c is hidden and the reasoner is asked to identify this category.

We model this inference using the posterior distribution $p(c|f_1, \dots, f_n)$, or the distribution over categories given the features that have been observed. This distribution can be

written as:

$$p(c|f_1, \dots, f_n) \propto p(f_1, \dots, f_n|c)p(c) \quad (1)$$

$$= \prod_{j=1}^n p(f_j|c)p(c) \quad (2)$$

where the right hand side is expressed using distributions specified by the semantic repository ($p(f|c)$ and $p(c)$). Equation 2 combines two criteria: the hidden category should have high prior probability ($p(c)$ should be high), and should also be consistent with the observed features ($p(f_j|c)$ should be high for each observed feature j). Note that Equation 2 follows from Equation 1 only if the features f_1 through f_n are conditionally independent given the hidden category c . We make this independence assumption for simplicity, but for some applications it may need to be relaxed.

Equation 2 is strongly reminiscent of a categorization model that is known as the “naive Bayes” approach and is discussed by psychologists (J. R. Anderson, 1991) and machine learning researchers (Mitchell, 1997). This model, however, is typically used for object categorization rather than identification—in the standard setting, an object is observed with certain features (f_1 through f_n), and the inference problem is to assign this object to a category. Although there is a close relationship between the problems of object categorization and category identification, we explain in the General Discussion why it is important to distinguish the two.

Feature identification

The second problem in Table 1 is similar to category identification but now the reasoner is required to identify a feature given one or more categories that have the feature. For example, the reasoner might be informed that “zebras and tigers have Feature F” and asked to identify Feature F. We assume that questions of this kind are generated by sampling a feature f from the prior distribution $p(f)$ then sampling m categories from the distribution $p(c|f)$ (see Figure 1b). Feature f is then hidden, but can be inferred by computing the posterior distribution $p(f|c_1, \dots, c_m)$, or the distribution over features given the categories that have been observed:

$$p(f|c_1, \dots, c_m) \propto p(c_1, \dots, c_m|f)p(f) \quad (3)$$

$$= \prod_{i=1}^m p(c_i|f)p(f) \quad (4)$$

where again we have made an assumption of conditional independence. Intuitively, this distribution will favor features that have high prior probability and that are consistent with the categories observed.

Joint category and feature identification

Many different identification problems can be created by sampling features and categories from the semantic repository then hiding some of these items. Some of the most interesting cases involve problems where multiple items must be identified and these items

are related both to each other and to other items that have been observed. The joint identification problem in Table 1 is one such case where a reasoner must simultaneously identify a category and a feature. Suppose, for example, that rabbits have feature F, that Cs have feature F, and that Cs have stripes. A reasoner may combine all of this information and guess that feature F is *fur* and that category C is *tiger*.

Problems of this kind can be generated according to the method in Figure 1c.ii. We first sample a feature f_1 from the prior distribution $p(f)$ (here $f_1 = fur$). Next we sample m categories from the distribution $p(c|f_1)$: here $m = 2$, and the two categories are *tiger* and *rabbit*. We then sample $n - 1$ features from the distribution $p(f|c_1)$: here $n = 2$, and the additional feature sampled is *stripes*. We now create an identification problem by concealing the identities of the first feature (*fur*) and the first category (*tiger*). These items, however, can be inferred using a distribution over the hidden category and feature given everything else that has been observed:

$$p(c_1, f_1 | f_2 \dots f_n, c_2 \dots c_m) \propto p(f_2, \dots, f_n | c_1) p(c_2, \dots, c_m | f_1) p(c_1, f_1) \quad (5)$$

$$= \prod_{j=2}^n p(f_j | c_1) \prod_{i=2}^m p(c_i | f_1) p(c_1 | f_1) p(f_1) \quad (6)$$

Note that a pair (c_1, f_1) will only receive high posterior probability according to Equation 6 if c_1 is consistent with all of the observed features (f_2 through f_n), f_1 is consistent with all of the observed categories (c_2 through c_m), and c_1 is consistent with f_1 . As in Equations 2 and 4 we have assumed that features f_2 through f_n are conditionally independent given c_1 , and that categories c_2 through c_m are conditionally independent given f_1 .

Figure 1c.i shows an alternative method for generating joint identification problems in which category c_1 is sampled before feature f_1 . These two approaches will be equivalent if the distributions specified by the semantic repository satisfy

$$p(c|f)p(f) = p(f|c)p(c) \quad (7)$$

but this condition need not hold in general, and will not hold for the distributions used in this paper. Although the two approaches in Figure 1c may lead to different inferences about the hidden category and feature, here we evaluate the second approach. All of the joint identification problems in our experiment follow the pattern in Table 1 and mention the hidden feature before the hidden category. The approach in Figure 1c.ii seems appropriate for questions of this form, since it assumes that the hidden feature is generated before the hidden category.

The semantic repository

Equations 2, 4 and 6 specify a formal approach to identification that relies on four distributions: $p(c)$, $p(f)$, $p(c|f)$, and $p(f|c)$. These distributions capture the background knowledge that guides identification and are assumed to be specified by a semantic repository that includes knowledge about categories and their features. For all analyses in this paper we use a semantic repository based on the Leuven natural concept database (De Deyne et al., 2008). Note, however, that this semantic repository represents just one way to formalize the distributions required by our probabilistic framework. Future studies can

		0.027	0.015	0.010	0.009	0.003	$p(f)$
		is small	is an insect	lives in water	is green	bites	...
0.006	alligator	0	0	7	12	4	
0.008	anchovy	16	0	5	0	1	
0.012	ant	12	12	0	0	8	
$p(c)$	⋮						

Figure 2. Our model relies on distributions $p(c)$, $p(f)$, $p(f|c)$ and $p(c|f)$ that are provided by a semantic repository. The latter two distributions are defined by normalizing the rows or columns of a semantic matrix like the example shown here.

explore whether alternative repositories allow our framework to account better for human inferences.

The repository considered here includes 113 animal categories and 757 features. The full Leuven database includes 129 categories and familiarity ratings for the entire set. We dropped the 16 least familiar categories, and normalized the familiarity ratings for the 113 remaining categories to create the prior distribution $p(c)$ required by our model. Of the 113 categories in the semantic repository, *cat* and *dog* are the two with highest prior probability, and *llama* and *python* are the two with lowest prior probability.

The Leuven data include two components that specify relationships between categories and features. The first is a listing matrix L collected in an experiment where participants were asked to list features for each category. Entry $L(i, j)$ indicates the number of participants who listed feature j for category i .¹ The listing data specify a set of features, and in a follow-up experiment four participants provided binary acceptability judgments about all category-feature pairs. We organized these responses into a truth matrix T where $T(i, j) = 1$ if any of the four participants indicated that category i had feature j . For example, suppose that no participant spontaneously generated the feature *breathes air* when asked to list features of hamsters, but that *breathes air* was generated for another category such as dolphins. In the follow-up experiment, participants may have indicated that hamsters breathe air when asked directly about this category-feature pair. If so, then matrices L and T will show that $L(\textit{hamster}, \textit{breathes air}) = 0$ but that $T(\textit{hamster}, \textit{breathes air}) = 1$.

We combined the listing matrix L and the truth matrix T to define the conditional

¹Although the Leuven database includes both animal and artifact categories, note that the listing matrix L is currently available only for the animal categories. Our work therefore focuses exclusively on the animal categories.

distributions $p(c|f)$ and $p(f|c)$. Both distributions are defined in terms of a semantic matrix $S = L + T + 0.001$ (see Figure 2). Combining L and T ensures that any true category-feature pair will have a semantic strength of at least 1, and that pairs which are frequently generated in the listing task will have high strengths. Adding a small constant (0.001) ensures that there is a non-zero strength for any category-feature pair, which will be useful when working with noisy experimental data. The distributions $p(c|f)$ and $p(f|c)$ can now be defined by normalizing the rows and columns of semantic matrix S . If we normalize the rows we produce a matrix where row i specifies the distribution $p(f|c_i)$. If we normalize the columns we produce a matrix where column j specifies the distribution $p(c|f_j)$. In other words,

$$\begin{aligned} p(f_j|c_i) &= \frac{S(i, j)}{\sum_{j'} S(i, j')} \\ p(c_i|f_j) &= \frac{S(i, j)}{\sum_{i'} S(i', j)} \end{aligned} \tag{8}$$

The distributions just defined can be given the following probabilistic interpretation. Suppose that the frequencies in row i of matrix L are generated from a multinomial distribution θ_i and that the prior distribution on θ_i is a Dirichlet distribution with a parameter vector equal to 0.001 plus the i th row of matrix T . In other words, matrix T captures our initial expectations about which features may be associated with category i , and the observations in L can be used to update these expectations. Given these assumptions, normalizing the i th row of matrix S produces a distribution $p(f|c_i)$ that is equivalent to the maximum *a posteriori* estimate of the distribution θ_i that generated the listing data for category i . Our specification of $p(c|f_j)$ can be given a similar interpretation.

The remaining distributions in the semantic repository specify prior distributions $p(c)$ over categories and $p(f)$ over features. As described already, our prior $p(c)$ is based on the familiarity ratings included in the Leuven database. We define the prior $p(f)$ using information in the listing matrix L :

$$p(f_j) = \frac{\sum_i L(i, j)}{\sum_{i, j} L(i, j)} \tag{9}$$

Equation 9 indicates that the probability of any feature is proportional to the number of times it was generated in the feature-listing experiment. The resulting prior captures the familiarity of a feature, or the readiness with which it springs to mind. Of the 757 features in the semantic repository, *is small*, *is a bird* and *is an animal* are the three with highest prior probability, and *looks like a snail*, *is difficult to remove from one's body*, and *was used in medical science in earlier times* are three with very low prior probability.

The strategy used for collecting the Leuven data creates an important asymmetry between categories and features. Matrix L was constructed by asking participants to list features of a given categories, not by asking participants to list categories which had a given feature. As a result, our method for estimating $p(f|c)$ from the Leuven data is more principled than our method for estimating $p(c|f)$. Both methods could be placed on an equal footing by supplementing the Leuven data with results from a ‘‘category-listing’’ experiment in which participants are provided with a feature and asked to list which categories have this feature. Here, however, we use the available data—matrices L and T —to create a rough and ready approximation of $p(c|f)$.

A second issue with the Leuven data is that there were slight variations in the number of participants who completed the feature listing task for each category. At least 20 participants generated features for each category, but the exact number for each category cannot be accurately reconstructed. As a result, the distributions $p(c|f)$ and $p(f)$ in Equations 9 and 8 may be somewhat distorted. Note, however, that the total number of features generated for each category is approximately the same, indicating that the variability in the number of participants per category could not have been too large. We therefore believe that this variability introduces some noise into the distributions $p(c|f)$ and $p(f)$, but does not fundamentally compromise the quality of our semantic repository.

The distributions specified by our repository do not satisfy the constraint in Equation 7. As a result, our repository is not consistent with a principle of “order invariance” which states that the distribution on category-feature pairs (c, f) is the same regardless of whether the category or the feature is generated first. Given the conditional distributions in Equation 8, the principle of order invariance is uniquely satisfied by choosing priors

$$\begin{aligned} p(c_i) &= \frac{\sum_j S(i, j)}{\sum_{i,j} S(i, j)} \\ p(f_j) &= \frac{\sum_i S(i, j)}{\sum_{i,j} S(i, j)} \end{aligned} \tag{10}$$

These priors, however, are not consistent with our intuitions about the relative familiarity of different categories and features. For example, the three categories with highest prior probability according to Equation 10 are *bee*, *owl*, and *lion*, and neither *cat* nor *dog* appears among the top 80 categories. This result may be explained in part by the fact that different numbers of participants listed features for different categories. Since the priors in Equation 10 do not make intuitive sense, we chose to abandon the principle of order invariance. Future work can assess this principle in more detail, and can examine whether people produce different distributions over category-feature pairs depending on the order in which they generate members of these pairs.

A relevance account of identification

We know of no alternative framework that will handle all three problems in Table 1, but Lombardi and Sartori (2007) developed a *semantic relevance* model of category identification and the same approach can be applied to feature identification. Consider first the problem of category identification. As before, we assume that a set of features $\{f_1, \dots, f_n\}$ is observed, and that the task is to identify a category that has these features. Roughly speaking, a feature is relevant to a category if it is strongly associated with that category and associated with few other categories. Formally, we transform a semantic matrix S into a relevance matrix R with entries

$$R(i, j) = S(i, j) \log \left(\frac{m}{m_i} \right),$$

where m is the total number of categories and m_i is the number of categories with non-zero associations to feature i . We set $S = L + T$ where L and T are the listing and truth matrices derived from the Leuven data, and smooth the resulting R matrix by adding 0.001 to each entry.

Matrix R specifies the relevance of each feature to each category, but we need some way to combine these relevance scores in cases where multiple features of the hidden category are provided. Lombardi and Sartori (2007) rely on a measure of setwise relevance that captures the extent to which a set of features is relevant to a category. Given such a measure, they propose that people solve category identification problems by choosing the category that is maximally relevant to the provided features under this measure. At least two setwise measures can be considered: the *additive* relevance model uses $R_{\text{set}}(\{f_1, \dots, f_n\}, c_i) = \sum_{j=1}^n R(i, j)$ and the *multiplicative* relevance model uses $R_{\text{set}}(\{f_1, \dots, f_n\}, c_i) = \prod_{j=1}^n R(i, j)$. Lombardi and Sartori (2007) recommend the additive approach, and suggest that an additive model outperforms a multiplicative model on the problem that they consider. Here, however, we evaluate both approaches.

A relevance model of feature identification can be defined similarly. The main difference is that entry $R(i, j)$ in the relevance matrix R should now specify the relevance of category i to feature j rather than the relevance of feature j to category i . As a result, the strength matrix S is normalized differently to create the relevance matrix:

$$R(i, j) = S(i, j) \log \left(\frac{n}{n_j} \right),$$

where n is the total number of features and n_j is the total number of features with non-zero associations to category i . As before, we smooth the relevance matrix by adding 0.001 to each entry, and consider two strategies—additive and multiplicative—for combining relevance values.

Experiments 1 and 2: Category and feature identification

We developed three experiments to explore the identification problems in Table 1. Our first two experiments explore the two simplest identification problems: category identification and feature identification. Category identification has been previously studied (Lombardi & Sartori, 2007) but feature identification has received little attention, and our first and most basic goal is to explore whether people can solve both problems. A second goal is to evaluate our Bayesian approach and to compare it with the additive relevance approach. Lombardi and Sartori (2007) found that the additive relevance model outperforms a Bayesian account of category identification, but this result may depend in part on the semantic repository that they used, and the resources included in the Leuven database should allow a more accurate test of competing models than has previously been possible.

Experiment 1: Category identification

Our first experiment required participants to identify a hidden category given up to three features of this category.

Participants. 20 adults participated for course credit.

Stimuli and Procedure. Participants were asked to fill out a written questionnaire. The instructions informed them that each question would list “between 1 and 3 features that describe a certain animal” and that they should make “three guesses about what that animal might be.” A demonstration question was included where the categories were fruits

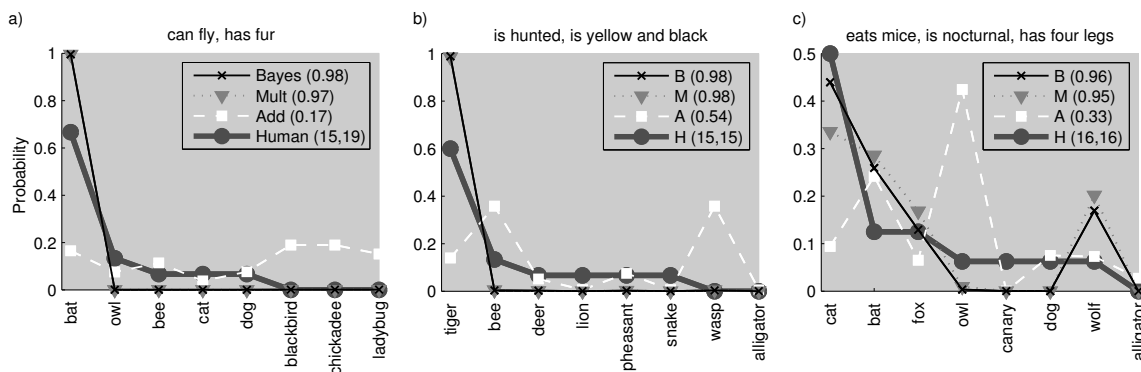


Figure 3. Responses to three of the category identification problems in Experiment 1. The curves for the Bayesian and multiplicative models are near identical in all three plots. The first three numbers in the legend show the amount of probability mass assigned by each model to the eight items in each plot. The remaining numbers show the absolute number of human responses for each question, and the number of responses that chose one of the eight categories shown.

rather than animals, and the features provided were *yellow* and *grows on trees*. Several possible guesses were listed, including *lemon*, *banana*, *mango*, and *grapefruit*.

The experiment included 50 questions: 20 listed three features, 20 listed two features, and 10 listed a single feature. A complete list of questions is shown in Appendix A. Note that some of the translations included in the Leuven database were adjusted in order to make the feature labels as idiomatic as possible. For example, *has a fur* (in dutch: *vacht*) was replaced by *has fur*. Participants generated a ranked list of three categories in response to each question, and were asked to leave some of these slots blank if they could not think of three categories for a given question. The order of the questions was randomized across participants.

Results.

The data were coded by manually identifying the category in semantic matrix S that best matched each response. 29% of the responses were left uncoded because they did not correspond to any category in the S matrix, and all subsequent analyses will consider only the responses that were coded. Most of the coded responses were cases where a participant provided the exact label for one of the categories in the repository. A small number of cases, however, were less straightforward. The semantic repository includes some subordinate categories (e.g. *viper* and *cobra*), but subordinate categories that do not appear were coded using an appropriate basic-level category (e.g. *taipan* was coded as *snake*, and *great white shark* was coded as *shark*). Any modifiers were dropped—for example, *baby kangaroo* was coded as *kangaroo*, and *furless cat* as *cat*. A small number of pairs were not included under any of the previous criteria but were treated as equivalent—for example, *puppy* was coded as *dog*, and *pony* as *horse*.

Responses to three questions are summarized in Figure 3. In each plot, the human curve shows an empirical probability distribution where the probability of a category is proportional to the number of times it was named in the experiment. Given, for example, that Cs eat mice, are nocturnal and have four legs, the most common response is that $C = \text{cat}$ (Figure 3c). Note that there are several good responses to many questions in

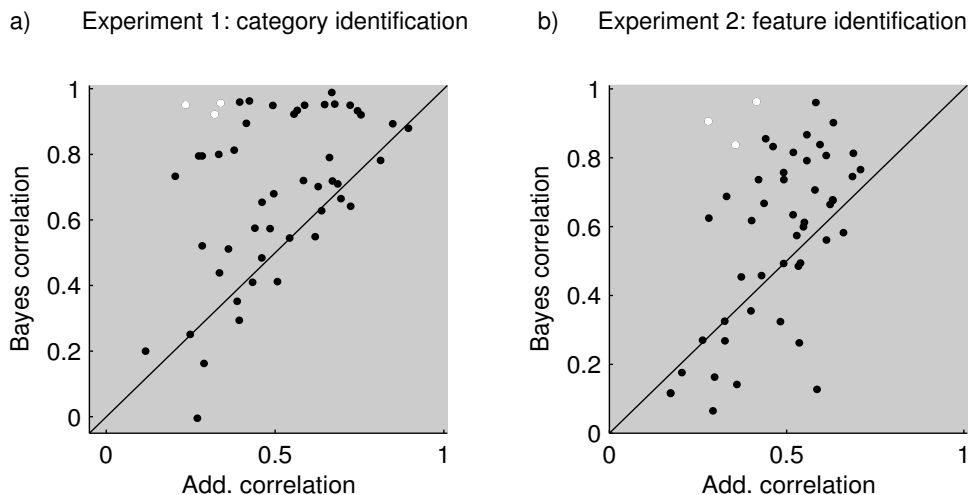


Figure 4. Model correlations for the individual questions in Experiments 1 and 2. For each question we constructed an empirical distribution where the probability of an item (a category or feature) is proportional to the number of times it was named in the experiment. The x and y axes show correlations between these empirical distributions and the predictions of the additive relevance and Bayesian models. In each plot, the trio of white points shows the three questions that led to the greatest difference in performance between the two models. Responses to these trios are plotted in Figures 3 and 6.

the experiment. In Figure 3c, for example, *fox* is a sensible guess given the information provided.

The remaining curves in Figure 3 show the predictions of three models: the Bayesian model, the additive relevance model, and the multiplicative relevance model. Each of these models orders the 113 categories in the S matrix in response to each question. The Bayesian model generates a probability distribution $P(c|\{f_1, \dots, f_n\})$ over these categories, and the semantic relevance models (additive and multiplicative) generate relevance scores $R_{\text{set}}(\{f_1, \dots, f_n\}, c)$ for each category. For our purposes, we convert the relevance scores to a pseudo-probability distribution by normalizing so that each set of scores sums to one.

The eight categories along the x-axis of each plot were chosen by taking the top two choices according to each model and according to our participants. If some of these choices overlapped leaving fewer than eight distinct choices in total, we filled out the set by including the next most common human responses. The additive relevance model often assigns scores to the different categories that differ very little in absolute magnitude. To ensure that the relative preferences of each model were visible, we normalized each curve in Figure 3 so that the scores sum to 1 across the eight items in each plot. The legend indicates the total probability mass assigned to the eight categories by each model. Figure 3a shows, for instance, that the eight categories plotted account for 0.98 of the probability mass according to the Bayesian model, indicating that very little probability is assigned to the 105 categories not shown. The additive model, however, assigns only 0.17 of its probability mass to the eight categories shown, indicating that a substantial amount of probability mass is reserved for the remaining categories in the data set. The legend for each plot also shows the absolute number of human responses that referred to one of the eight categories shown.

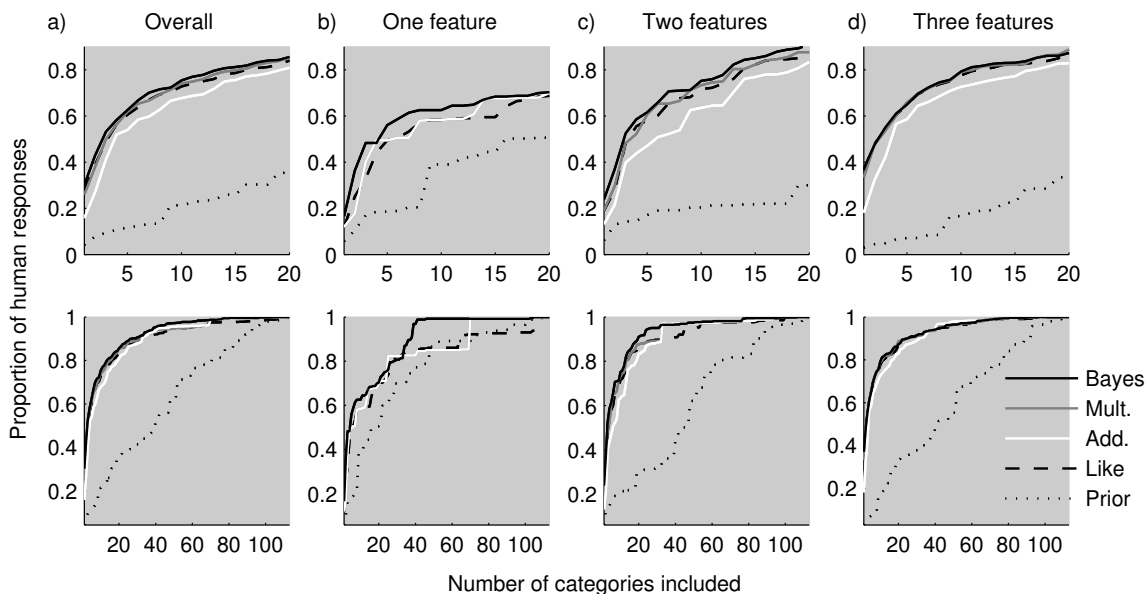


Figure 5. Rank analysis for Experiment 1. Each plot shows the proportion of human responses included when we take the best n responses according to a model. The first column shows results across all questions, and the next three columns show results for questions including one, two and three features respectively. The top row shows results for values of n between 1 and 20, and the bottom row shows the complete curve for each model. *Bayes* is our Bayesian model and *Mult* and *Add* are the multiplicative and additive relevance models. Note that the two relevance models generate identical results for the one feature questions.

In Figure 3a, for example, 19 human responses matched a category in our data set, and 15 of these responses matched one of the eight categories shown.

Figure 3 suggests that the Bayesian and multiplicative relevance models perform similarly, but that the additive relevance model often prefers categories that are linked with only some of the features provided. In Figure 3a, for example, the top choice according to the additive model is *blackbird*, which can fly but does not have fur. The Bayesian and multiplicative models successfully predict that *bat* is the best response to this question.

The questions in Figure 3 are the three that best discriminate between the Bayesian model and the additive relevance model. For each of the 50 questions in the experiment, we computed correlations between model predictions and the empirical distributions generated by participants. Correlations for the Bayesian and additive relevance models are shown in Figure 4a, and the three white points represent the three questions shown in Figure 3. Figure 3 shows that the correlations achieved by the Bayesian model vary across a wide range, but that most of the points lie above the diagonal, indicating that the Bayesian model tends to perform better than the additive relevance model. Average correlations confirm this conclusion, and indicate that the Bayesian (0.70) and multiplicative (0.68) models both perform substantially better than the additive model (0.50).

We ran a second analysis to assess the predictions of the three models. For any given question, the top few categories according to a successful model should account for the majority of the responses provided by participants. The “top- n inclusion curves” in Figure 5 are based on the proportion of human responses that fall within the top n categories chosen

by each model. The first column shows the combined results across all 50 questions in the experiment, and the remaining columns show results for the 10 one-feature questions, the 20 two-feature questions, and the 20 three-feature questions.

The top row of Figure 5 shows results for the critical regime where n is small—intuitively, we hope that most human responses to a given question will belong to the top few categories according to a model. The Bayesian model outperforms both relevance models on the one-feature questions, and the Bayesian and multiplicative models both outperform the additive model when two or more features must be considered. Figure 5a shows, for example, that the single best response according to the Bayesian model accounts for 30% of human responses overall, but that the best response according to the additive model accounts for only 16% of human responses.

The remaining two models shown in Figure 5 are similar to the Bayesian model, but include only the likelihood term $p(f_1, \dots, f_n|c)$ or the prior $p(c)$. Both models perform worse than the Bayesian model—the likelihood model by a small margin, and the prior model by a large margin. This result suggests that both components of the Bayesian model are needed to account for human inferences.

Experiment 2: Feature identification

We developed a second experiment to explore the problem of feature identification. Experiment 2 was very similar to Experiment 1 except that participants were given categories rather than features, and were asked to list features rather than categories.

Participants. 20 adults participated for pay or course credit.

Stimuli and Procedure. Participants were informed that each question would list “between 1 and 3 animals that have a certain feature,” and that they should make three guesses about what that feature might be. A demonstration question was included where the categories were fruits—*banana* and *pineapple*. Nine possible guesses were listed, including features like *is yellow*, *is a food*, *grows in tropical countries*, and *has a thick skin*.

The experiment included 50 questions: 32 listed three categories, 12 listed two categories, and 6 listed a single category. All questions are shown in Appendix A. Participants generated a ranked list of three features in response to each question, and were asked to skip any question including a category that they had never encountered before.

Results.

The data were coded by manually identifying the feature in matrix S that best matched each response. 18% of the responses were left uncoded because they did not correspond to any feature in the S matrix. For several reasons the coding task was more challenging than the corresponding task for Experiment 1. There are often several ways to describe a given feature. Some descriptions seem very close (e.g. *jumps* and *moves in a jumping pattern*) but other cases are less clear (e.g. *lives in water* and *found in water*.) A second challenge is that the semantic repository contains several groups of features that are similar in meaning (e.g. *is edible* and *is eaten for meat*), and we chose just one of these features for any given response. A third challenge is that the features were provided by Dutch speakers, and some of the feature labels (e.g. *has a sharp view*) are rather different from the labels that American undergraduates might provide (e.g. *has good eyesight*). Many coding

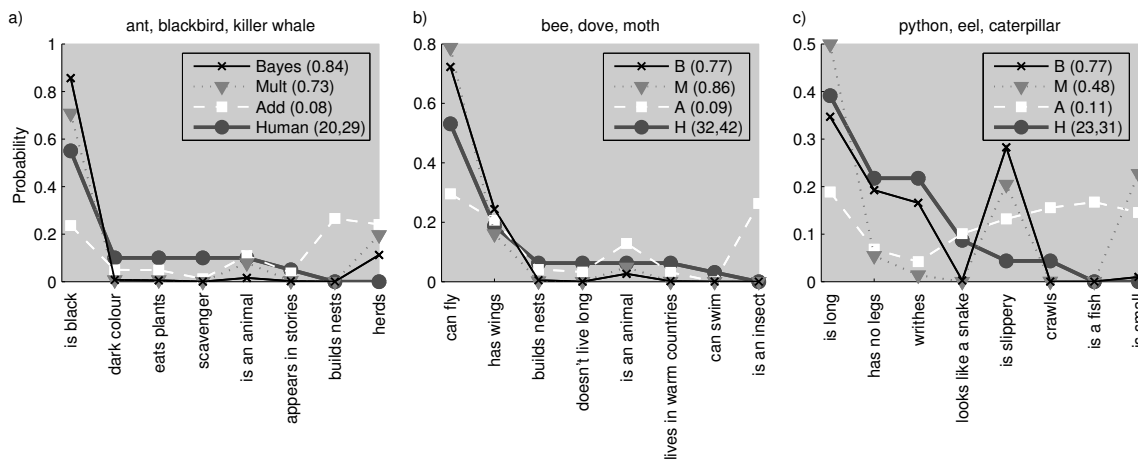


Figure 6. Responses to three of the feature identification problems in Experiment 2.

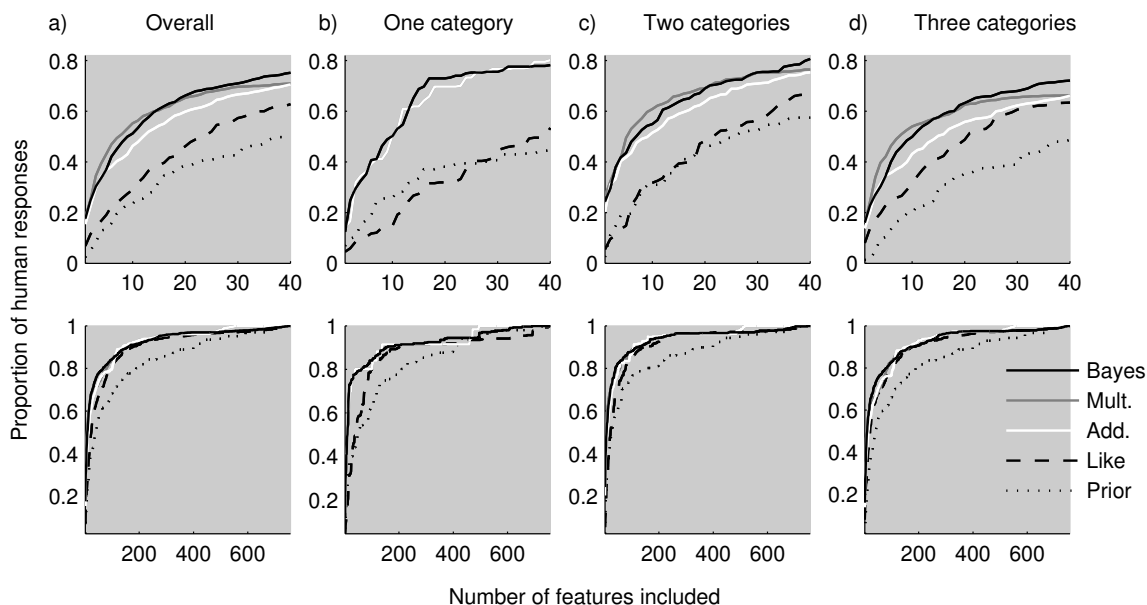


Figure 7. Rank analysis for Experiment 2. Each plot shows the proportion of human responses included when we take the best n responses according to a model. The first column shows results across all questions, and the next three columns show results for questions including one, two and three categories respectively. *Bayes* is our Bayesian model and *Mult* and *Add* are the multiplicative and additive relevance models. The remaining two models are similar to *Bayes*, but include only the likelihood term (*Like*) or only the prior term (*Prior*).

decisions were straightforward, but for those that remained we made our best attempt to capture the meanings intended by our participants. The coded data files are available online so that future researchers can examine and perhaps adjust some of our decisions.²

Responses to three questions are summarized in Figure 6. As for Experiment 1, there are several good responses to most questions in the experiment. Given, for example, that

²See www.charleskemp.com

pythons, eels, and caterpillars share a certain feature, participants may say that the feature is *long*, *has no legs* or *writhes*. The questions in Figure 6 are again the three that best discriminate between the Bayesian model and the additive relevance model. The Bayesian and the multiplicative relevance models make similar predictions, but the additive model often prefers features that are linked with only some of the categories provided. For example, given the categories *frog*, *lizard*, and *spider*, the top choice according to the additive model is *makes a web*, but the Bayesian and multiplicative models both successfully predict that people will prefer *eats insects*.

Figure 4b shows the correlations achieved by the Bayesian and the additive models across the 50 questions in the experiment. The majority of the points fall above the diagonal, indicating that the Bayesian model tends to predict human judgments better than the additive model. Average correlations confirm this conclusion, and indicate that the Bayesian (0.58) and multiplicative (0.63) models both perform substantially better than the additive model (0.47).

Figure 7 provides further evidence that the Bayesian model outperforms the additive model when all questions in the experiment are taken into account. There is no real difference between the two models for questions where a single category is provided, but the Bayesian model outperforms the additive model in cases where participants must combine information about two or more categories. Figure 7 also shows that the Bayesian model outperforms the two related models which include only the likelihood term $p(c_1, \dots, c_m|f)$ or the prior $p(f)$. This time the margin in both cases is substantial, and we can conclude that both components of the Bayesian model make an important contribution.

Discussion

Our first two experiments suggest that people are able to solve both category and feature identification problems. The Bayesian and multiplicative models achieve levels of performance that seem roughly similar, but the additive model is substantially worse at accounting for human inferences. This result may seem surprising, since Lombardi and Sartori (2007) reach the opposite conclusion and show that the additive approach provides a better account of their data than the multiplicative approach. There are several reasons, however, why these studies may have arrived at different conclusions.

One important difference between these studies is that our semantic matrix S is substantially less sparse than the matrix used by Lombardi and Sartori. Around 23% of the entries in our semantic matrix are nonzero, but the corresponding figure for the Lombardi and Sartori matrix is around 6%.³ The reason for this four-fold difference is that our semantic matrix incorporates a listing matrix L and a truth matrix T , but the Lombardi and Sartori matrix includes only feature-listing data. Several other semantic matrices in the literature are based entirely on feature-listing data (McRae, Cree, Seidenberg, & McNorgan, 2005), and all of these matrices are likely to underestimate the knowledge that people bring to identification problems. In particular, these matrices are likely to have gaps

³The density of the entire Lombardi and Sartori matrix is around 2%, but this matrix includes categories from several domains, including animals, musical instruments, vegetables, and artifacts. If we include only the animals and the features that were listed for at least one animal, the density increases to around 6%. The density of the Leuven feature-listing matrix L used in our experiments is also around 6%.

corresponding to facts that people know to be true but that are unlikely to be generated during a feature listing task (e.g. the fact that hamsters breathe air).

Lombardi and Sartori’s results may suggest in part that the additive approach is more robust than the multiplicative approach when the semantic matrix contains gaps. Given, for example, that Cs are pets, have fur, and breathe air, the additive approach may assign a high score to *hamster* even though the semantic matrix does not acknowledge that hamsters breathe air. For the multiplicative approach, one near-zero component (e.g. the association between *hamster* and *breathes air*) is enough to ensure that the overall score assigned to *hamster* is near zero.

Tolerating gaps in the semantic matrix may be a strength when working with an incomplete matrix but can introduce some fundamental problems. Consider, for example, the feature identification problem in Figure 6b. Given that bees, doves, and moths have a certain feature, the additive approach assigns a high score to *is an insect* since this feature is strongly associated with bees and moths. Even though the feature is not associated with doves, the additive approach is willing to overlook this gap in the semantic matrix. The multiplicative approach, however, assigns a low score to *is an insect*: one near-zero component (doves are not insects) is enough to ensure that the overall score for *is an insect* is near zero. In other words, the multiplicative approach alone is able to filter out features that are not shared by all of the categories provided.

To explore whether the truth matrix T is the critical difference between our work and the work of Lombardi and Sartori, we set this matrix to 0 and repeated our analysis of Experiment 1. We found that the average correlations achieved by the Bayesian (0.68) and multiplicative (0.68) models were lower than the values previously reported, but that both models still performed better than the additive model (0.50). A second difference between the studies is that we smooth the semantic matrix S by adding a small constant to each entry, but Lombardi and Sartori do not use any smoothing procedure. To explore the role of smoothing we repeated our analysis of Experiment 1 and set the smoothing constant to zero. Again the performance of the Bayesian (0.62) and the multiplicative models (0.59) dropped, but both still performed better than the additive model (0.49). We ran our analysis one more time and set both the truth matrix and the smoothing constant to zero. The Bayesian (0.57) and multiplicative (0.56) models still achieved higher correlations than the additive model (0.51), but a rank analysis similar to Figure 5 now suggested for the first time that the additive model was superior to the others. We can therefore conclude that the truth matrix T and the smoothing procedure are both required in order for the Bayesian and multiplicative models to achieve their full potential.

Although smoothing makes an important contribution to our results, we reanalyzed the data collected by Lombardi and Sartori and found that their conclusions do not significantly change when smoothing is applied. This result suggests that the relative performance of the additive and multiplicative models must depend on at least one additional factor that distinguishes our study from the work of Lombardi and Sartori. There are at least two likely candidates. First, the dependent variable considered by Lombardi and Sartori is naming accuracy, or the probability that a target category will be correctly identified given the limited information available. We propose that there are often many good responses to a given question, and therefore work with the distribution across possible responses (Figures 3, 4, and 6) or the rank order of the possible responses (Figures 5 and 7). A second notable

difference is that Lombardi and Sartori generated their stimuli by sampling from their semantic matrix, but we generated our stimuli somewhat independently of the contents of our matrix. Sampling stimuli from a feature-listing matrix will tend to minimize the impact of gaps in this matrix, since a feature like *breathes air* will never be sampled if the target category is *hamster*.

Although our results raise challenges for the additive relevance model they do not clearly discriminate between the multiplicative model and our Bayesian model. Figure 5 suggests that the Bayesian model outperforms the multiplicative model by a small margin in Experiment 1, but Figure 7 suggests that these models perform comparably in Experiment 2. We propose, however, that the Bayesian approach may be preferred on two grounds. First, the Bayesian approach provides a principled account of information integration that does not rely on arbitrary decisions about whether to combine scores using a product or a sum. There are some attempts to provide principled derivations of relevance-style approaches (Robertson, 2004), but most of these derivations lead to additive rather than multiplicative models. Second, the Bayesian approach relies on general-purpose probabilistic inference, and can therefore be extended to handle a wide range of inductive problems. Our third experiment considers one of these problems, and helps to illustrate the generality of the Bayesian approach.

Experiment 3: Simultaneously identifying categories and features

Identification problems are interesting in part because they often require multiple pieces of evidence to be combined in order to yield a solution. Our first two experiments include some simple examples of this idea. Given, for instance, that polar bears and swans have a certain feature, a reasoner must combine what she knows about polar bears and swans in order to identify the feature.

Our third experiment explores a family of problems where multiple pieces of evidence must be combined in more sophisticated and subtle ways. We consider joint identification problems, or problems where a category C and a feature F must be identified simultaneously (see Table 1). The category and feature are known to be related (C s have feature F) and additional information about the category and feature is also provided: for example, rabbits have feature F and C s have stripes. Experiment 3 explores whether all of these constraints can be combined in order to identify the category and the feature. In the case just described, a good guess might state that feature F is *fur* and that category C is *tiger*.

Participants

30 adults participated for course credit.

Stimuli

Participants completed a written questionnaire with 60 questions. Each question listed three category-feature pairs: for example, the joint identification problem in Table 1

was represented as

$$\begin{aligned}
 & \text{rabbit, feature F} \\
 & \text{animal C, feature F} \\
 & \text{animal C, has stripes}
 \end{aligned} \tag{11}$$

For each question, participants made a single guess about the identity of feature F and animal category C. The order of the questions was randomized across participants, but the three pairs in each question were always listed in the order shown above. A demonstration question was included where the categories were fruits rather than animals:

$$\begin{aligned}
 & \text{lemon, feature F} \\
 & \text{fruit R, feature F} \\
 & \text{fruit R, is sweet}
 \end{aligned} \tag{12}$$

Three possible guesses were listed, including *is a citrus fruit* for the hidden feature and *orange* for the hidden fruit.

The experiment included 60 questions which are shown in Appendix A. The first 52 questions are organized into six *category groups* and seven *feature groups*. Each of these groups includes four questions. The questions in any given category group all mention the same category (e.g. *cow*) but mention four different features. The questions in any feature group mention the same feature (e.g. *eaten as meat*) but mention four different categories. In addition to these category and feature groups, eight extra questions were chosen to round out the set of 60.

Models

Our Bayesian model integrates all of the available information when guessing feature F and category C in Question 11. We can compare this approach to a baseline that does not combine the three statements. The baseline model uses the first statement (rabbits have F) to guess feature F, and uses the final statement (Cs have stripes) to guess category C, but does not attempt to integrate the three statements. More formally, given a joint identification problem as shown in Table 1, the baseline model chooses a category and a feature that maximize the distribution

$$p(c, f | f_2 \dots f_n, c_2 \dots c_m) \propto p(c | f_2, \dots, f_n) p(f | c_2, \dots, c_m) \tag{13}$$

where the distributions on the right hand side are specified by the Bayesian models used for Experiments 1 (Equation 2) and 2 (Equation 4). This baseline combines two probabilistic models that performed successfully in Experiments 1 and 2, but we predict that it will be less successful than our Bayesian model at predicting people’s responses to joint identification problems.

As defined in Equations 6 and 13, the Bayesian and the baseline models both allow items to be repeated. In Question 11, for example, both models could infer that feature F is *stripes* and that category C is *rabbit*. These hypotheses seem inconsistent with the

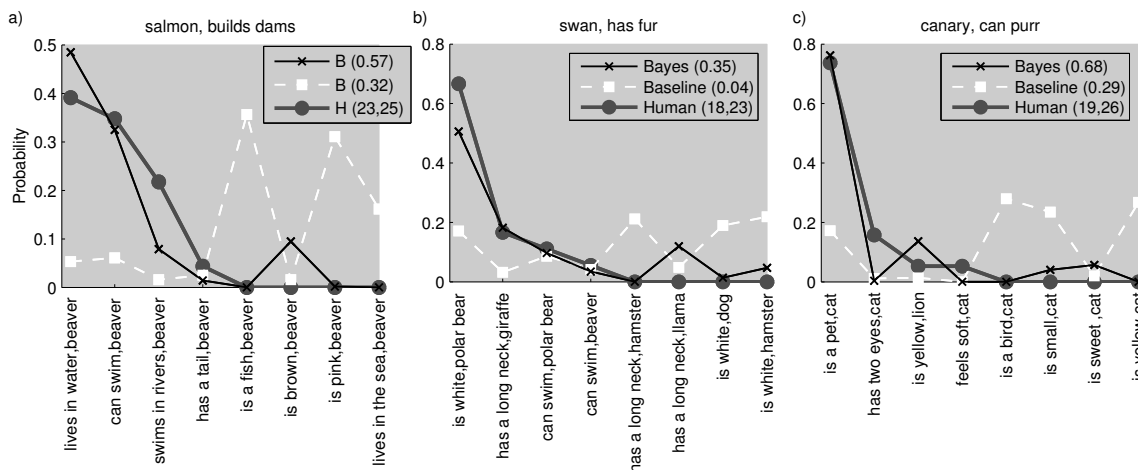


Figure 8. Responses to three of the joint identification problems in Experiment 3.

pragmatic expectations that participants bring to our task and we therefore adjust both models to rule out repeats. The right side of Equation 5 is replaced by

$$p(f_2, \dots, f_n | c, f) p(c_2, \dots, c_m | c, f) p(c, f) = \prod_{j=2}^n p^*(f_j | c, f) \prod_{i=2}^m p^*(c_i | c, f) p(c, f). \quad (14)$$

We previously defined a distribution $p(c_i | f)$, and now convert this into the distribution $p^*(c_i | c, f)$ by setting the probability of a repeat to zero and renormalizing:

$$p^*(c_i | c, f) \propto \begin{cases} 0 & \text{if } c_i = c \\ p(c_i | f) & \text{otherwise} \end{cases} \quad (15)$$

The distribution $p^*(f_j | c, f)$ is defined similarly, and the new distributions are used for both the Bayesian and the baseline models.

Results

The data were again manually coded. 15% of the categories and 6% of the features were left uncoded because they did not correspond to entries in the S matrix. 19% of the pairs provided by participants included at least one feature or category that does not appear in the S matrix. Responses to three questions are summarized in Figure 8. Given, for example, that salmon have feature F , that animal C has feature F , and that animal C builds dams, the most common response is that feature F is *lives in water* and category C is *beaver*.

To explore whether participants use all of the available information when choosing each component of their response, we examined the pairs chosen for each category and feature group. Figure 9a shows the responses for one category group (four arguments that mention *canary*) and 9b shows responses for one feature group (four arguments that mention *has fur*). The first group of bars in Figure 9a.i shows inferences about feature F

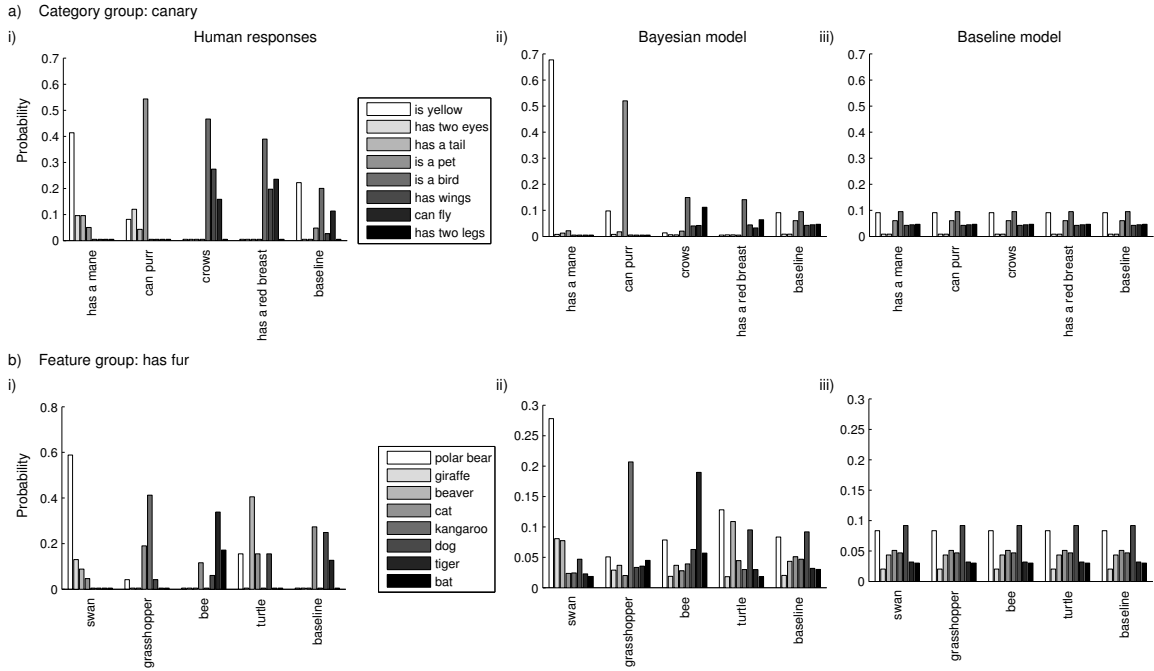


Figure 9. Joint identification responses across a category group and a feature group. (a)(i) All four questions in the category group mention the category canary, and the feature is either *has a mane*, *can purr*, *crows* or *has a red breast*. Participants were required to identify one feature of canary and the first four groups of bars summarize their responses to the four questions. The final group of bars summarizes responses from Experiment 2 when participants were told that canaries have F and asked to identify this feature. (ii) Predictions of the joint identification model. (iii) Predictions of the baseline model. (b) (i) All four questions in the feature group mention the feature *fur*, and the category mentioned is either *swan*, *grasshopper*, *bee* or *turtle*. Participants were required to identify one category with the feature *fur*, and the first four groups of bars summarize their responses. The final group summarizes responses from Experiment 1 when participants were told that Cs have fur and asked to identify category C.

in the question:

$$\begin{aligned}
 & \text{canary, feature F} \\
 & \text{animal C, feature F} \\
 & \text{animal C, has a mane}
 \end{aligned} \tag{16}$$

The most common response is *yellow*, but the next group of bars shows that the top response changes to *is a pet* when *has a mane* is replaced by *can purr* in Question 16. The final group of bars in Figure 9a.i shows responses from Experiment 2 when participants are simply told that “Canaries have feature F” and are asked to guess what the feature might be. Comparing the groups of bars, it is clear that inferences about feature F change depending on the feature mentioned in Question 16, suggesting that participants combine all components of the question when choosing their response.

Figure 9b shows a similar pattern of results, suggesting that inferences about category C are also shaped by all components of the question. The first group of bars in Figure 9b.i

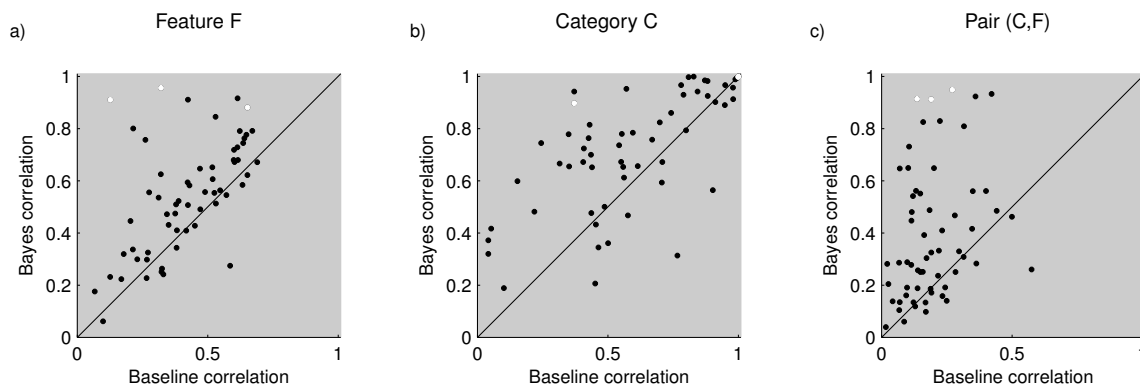


Figure 10. Model correlations for the individual questions in Experiment 3. Each point in each plot represents a single question and shows the correlation between human responses and the predictions of the Bayesian and baseline models (y and x values respectively). The first two plots consider marginal distributions over features and categories, and the final plot considers distributions over the set of all category-feature pairs. The three white points in each plot show the questions which led to the biggest difference between the two models according to the analysis in (c). Responses to these three questions are shown in Figure 8.

shows inferences about category C in the problem

$$\begin{aligned}
 & \text{swan, feature F} \\
 & \text{animal C, feature F} \\
 & \text{animal C, has fur}
 \end{aligned} \tag{17}$$

The most common response is *polar bear*, but the top response changes to *kangaroo* when *swan* is replaced by *grasshopper* in Question 17. Figure 9 only has room for a single category group and a single feature group, but similar results were obtained for all of the category and feature groups in Experiment 3.

To support these qualitative analyses we ran permutation tests to evaluate the null hypothesis that responses to all questions within a given group were drawn from the same distribution. Responses to each question in a category group were used to construct an empirical distribution over features, and responses to each question in a feature group were used to construct an empirical distribution over categories. Within each group, we randomly shuffled responses between questions, and computed the average pairwise Kullback-Leibler divergence between the empirical distributions as a test statistic. Repeating this process 1000 times produces a distribution over values of the test statistic for each group, and this distribution can be compared with the value of the statistic for the actual (i.e. unshuffled) responses. For all 13 groups, the value of the test statistic for the unshuffled responses is extreme, allowing the null hypothesis to be rejected ($p < 0.002$ in all cases).

Since Figure 9 suggests that participants integrate the three statements in each question, we expect that the Bayesian model should outperform the baseline model across most questions in our experiment. The questions shown in Figure 8 are the three that best discriminate between these two models. Note that the baseline model often chooses a category and feature that are incompatible. For example, in Figure 8a the two best responses according to the baseline model are the pairs (*is a fish, beaver*) and (*is pink, beaver*), which are

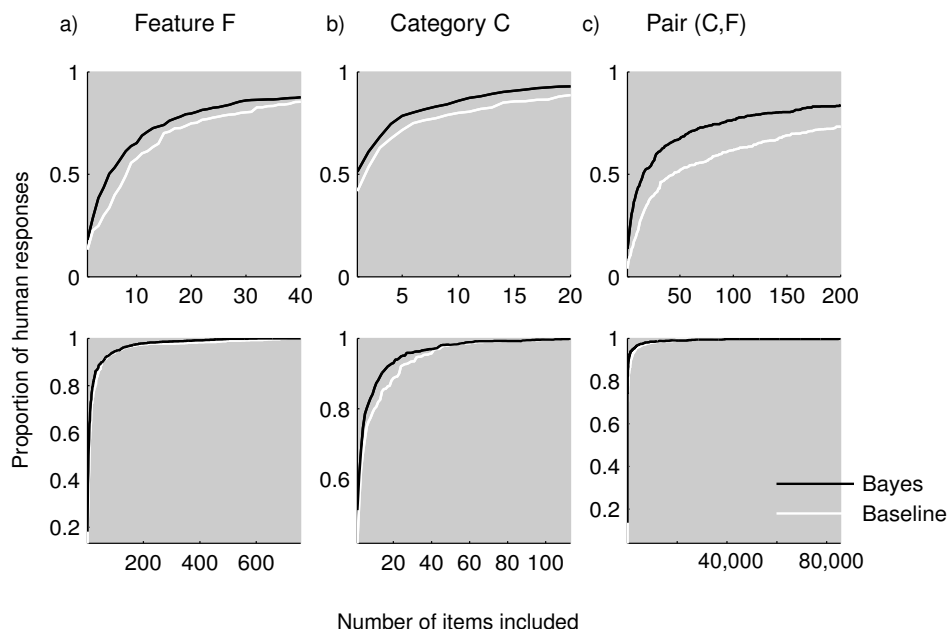


Figure 11. Rank analysis for Experiment 3. Each plot shows the proportion of human responses (features, categories, or category-feature pairs) included when we take the best n responses according to a model. The results in the first two columns are based on the marginal distributions over features and categories respectively. The results in the final column are based on the distribution over pairs induced by each model.

both incompatible with the statement that the hidden category has the hidden feature. The Bayesian model successfully identifies the most common human response to this question.

The relatively poor performance of the baseline model can be attributed to its failure to integrate all of the available information. Figure 9 supports this idea by showing model predictions within a category group and a feature group. In response to any given question, both models generate a probability distribution over all possible category-feature pairs. The predictions in Figure 9 show the marginal distribution over features (9a) or the marginal distribution over categories (9b). Figure 9a.iii shows that the baseline model makes identical inferences about feature F for all questions within the same category group, and 9.b.iii shows the corresponding result for all questions within the same feature group. Figures 9a.ii and 9b.ii show that the Bayesian model provides a better account of inferences within each group. Note, for example, that the Bayesian model successfully predicts the top human choice across the first three questions in the *has fur* feature group.

The correlations achieved by the Bayesian and baseline models across the full set of questions are shown in Figure 10. The three plots show how well the two models predict the marginal distribution on features, the marginal distribution on categories, and the distribution on category-feature pairs provided by our participants. The majority of data points fall above the line, indicating that the Bayesian model outperforms the baseline model on most questions.

The rank analysis in Figure 11 provides further evidence that the Bayesian model performs better than the baseline model. Figure 11a shows how accurately the models

identify the hidden feature F , and Figure 11b shows how accurately the models identify the hidden category C . These curves are again computed using marginal distributions over features (Figure 11a) or categories (Figure 11b). In both cases the Bayesian model tends to assign a higher rank to the items chosen by humans than the baseline model does. The curves in Figure 11c show success at identifying category-feature pairs, and again the Bayesian model outperforms the baseline model. Even though the number of possible category-feature pairs exceeds 80,000, Figure 11c shows that the top three responses of the Bayesian model account for more than 20% of the pairs chosen by humans, and that the top ten responses account for more than 40% of human responses.

Although the Bayesian model performs better than the baseline model, Figure 10c shows that there is still room to improve. Predicting which of the 80,000 possible pairs will be chosen for any given question is a very tough challenge, and Figure 10c shows that the Bayesian model achieves relatively low correlations on many questions. Our results, however, do suggest that humans can integrate multiple sources of information when identifying categories and features, and future studies can aim to model this ability more closely.

Discussion

We presented a probabilistic model of identification and described experiments that explore category identification, feature identification, and joint category and feature identification. The results indicate that people can solve all of these problems, and suggest that our model accounts better for human inferences than several alternatives.

Our model captures two general principles that help to explain how humans solve identification problems. First, prior knowledge matters. Knowledge about which categories have which features is obviously important, and is captured in our framework by the distributions $p(c|f)$ and $p(f|c)$. Knowledge that some categories and features are more familiar than others is also important, and is captured in our framework by the prior distributions $p(c)$ and $p(f)$. Our results for Experiments 1 and 2 suggest that the prior ($p(c)$ or $p(f)$) plays an important role, and show that our Bayesian model performs better than an alternative (the Likelihood model) which is very similar but does not incorporate a prior distribution.

The second general principle is that statistical inference helps to explain how people integrate multiple sources of information. Given, for example, that Cs are white and have fur, a reasoner must combine both statements in order to guess that category C is *polar bear*. Previous formal models have used several strategies to combine multiple observations, and different authors suggest that scores for individual observations can be combined using a sum (Lombardi & Sartori, 2007) or a product (Medin & Schaffer, 1978), or by taking the maximum (Osherson et al., 1990). A Bayesian approach avoids arbitrary choices between functions like these, and provides an account of information integration that is both principled and extremely general. Once we have specified how the individual observations are generated, statistical inference immediately specifies how multiple observations should be combined. The same basic approach can be applied across many inductive contexts. For example, a Bayesian approach can incorporate both positive observations (animal C is white) and negative observations (animal C is not white) (Kemp & Tenenbaum, 2009), and can adjust its predictions depending on whether the observations are sampled randomly or chosen by a knowledgeable teacher (Xu & Tenenbaum, 2007). To the best of our knowledge, no other account of information integration can offer all of these advantages.

Prior knowledge and the semantic repository

To model our three experiments we used a semantic repository that specifies four distributions: $p(c)$, $p(f)$, $p(c|f)$, and $p(f|c)$. As already described, the distribution $p(c)$ was defined using familiarity ratings, and the remaining three distributions were defined using a feature-listing matrix L and a truth matrix T . Our model depends critically on the knowledge captured by these distributions, and it is important to think carefully about the status of this knowledge.

At first glance the method used to generate the feature-listing matrix L may seem almost identical to our feature-identification task (Experiment 2). In the original feature-listing task, participants were given a category and asked to list features of that category. In our second experiment, participants were given one or more categories then asked to guess a feature common to all of these categories. Since these tasks are closely related, it is natural to wonder whether any account of feature identification based on feature-listing data is unavoidably circular. In other words, perhaps our model uses the results of one experiment (the feature-listing task) to explain another (our feature identification task) without contributing anything important of its own.

The value added by our model is an account of how multiple sources of information should be combined. Although Experiments 1 and 2 both included a handful of problems where only a single piece of information was available, the majority of problems specified two or more pieces of information, and participants were required to combine all of this information in order to guess the hidden category or feature. Information integration is also the issue of real interest in Experiment 3, where participants must combine information about a hidden category and a hidden feature in order to identify them both. As our comparison between the two relevance models suggests, there are many ways in which multiple pieces of information might be combined, and choosing to integrate information one way rather than another (e.g. with a sum rather than a product) can have important consequences. Our primary contribution is an account of information integration that is both principled and flexible enough to handle problems like the questions in Experiment 3.

Although our model takes the semantic repository for granted, future modeling efforts can aim to explain the origins of this repository. The repository used in this paper relies on a very simple representation—a weighted matrix—but richer representations will be needed to capture human semantic knowledge in full. When participants are asked whether hamsters breathe air, for example, they are unlikely to rely on a pre-existing association between this category (*hamster*) and this feature (*breathes air*). Instead, they probably make an inference using representations that are much more sophisticated than a collection of weights or associations, and that may include taxonomic hierarchies (Collins & Quillian, 1969), logical theories (Kemp, Goodman, & Tenenbaum, 2008) and other relational structures. The feature listing and truth matrices (L and T) collected as part of the Leuven data may approximate some aspects of the *content* of human semantic knowledge, but understanding the *form* of this knowledge is an important challenge for future work.

Understanding the structure of semantic representations must go hand in hand with understanding how these representations are used. Future work can explore how people use their semantic representations to respond to feature listing tasks. One relevant principle is the idea that people tend to choose maximally informative features, and this principle

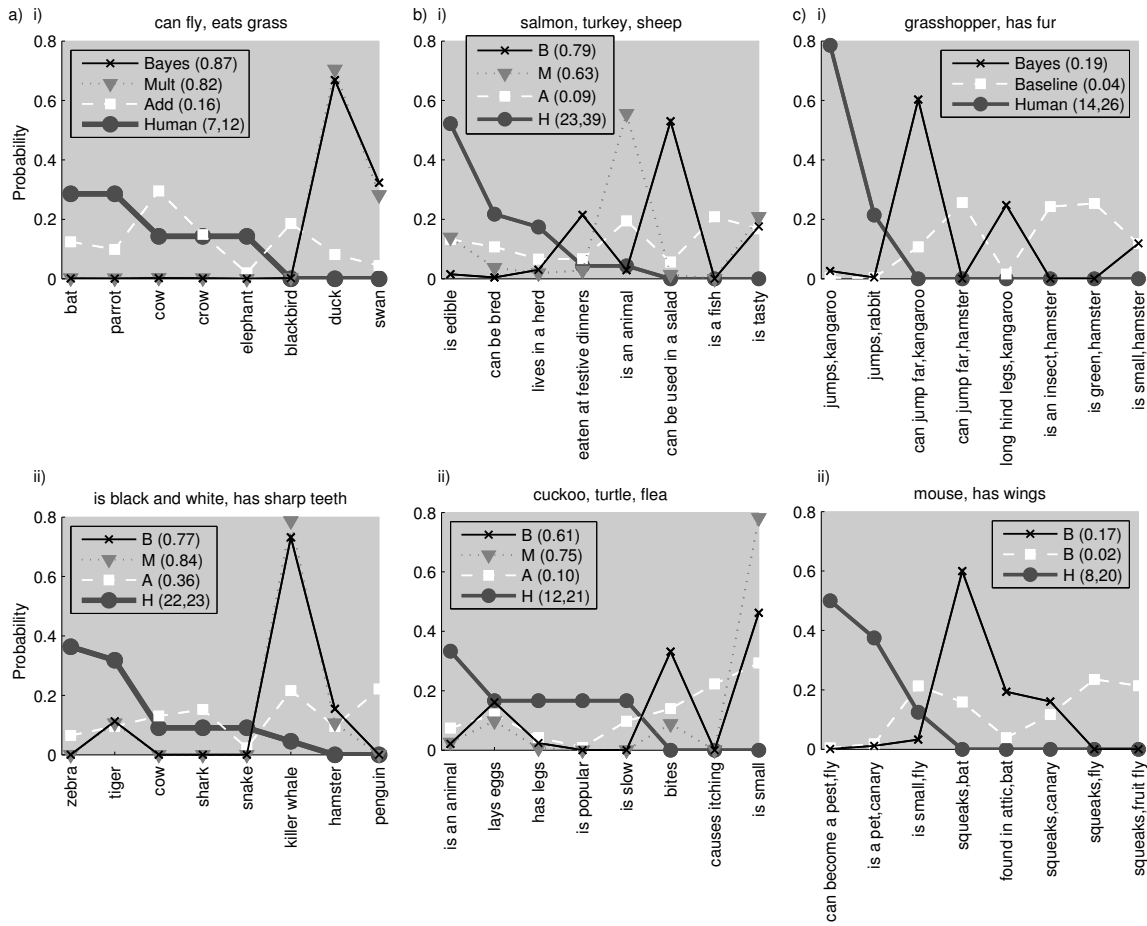


Figure 12. Cases where the Bayesian model accounts least well for human inferences.

may help to explain why people tend to generate features that are relatively rare (Navarro and Perfors, this issue). Much work remains to be done, but ultimately it may be possible to develop unified models of semantic cognition that explain how semantic knowledge is represented and how computations over these representations support both feature listing and identification.

Limitations of the model

Although our Bayesian model performs well with respect to the relevance models it does not account for the human data in all cases. Figure 12 shows results for the two questions in each experiment that produced the greatest departure between model predictions and human responses. These questions correspond to the points with the smallest y coordinates in Figures 4a, 4b and 10c.

The six examples in Figure 12 illustrate several reasons why the model sometimes fails. Some of these errors appear to result from misalignments between the English features provided by our participants and the Dutch features in the Leuven data. For example, Figure 12b.i shows a case where the top human choice (*edible*) receives low probability

according to the Bayesian model. The reason is that *edible* does not appear among the features of sheep in the listing matrix L , although *its meat is eaten* appears with relatively high weight. Figure 12c.i shows a case where the best responses according to the Bayesian model and our participants are identical except that the model response includes the feature *can jump far* and the human response includes the feature *jumps*. Some discrepancies of this kind may reflect subtle differences in meaning between English and Dutch words. Others may simply indicate that coding responses to feature-listing tasks is challenging, and that the decisions made by our coders may not always match the decisions made by the team that compiled the Leuven data.

Other model errors may result from the way in which we converted the Leuven data to a truth matrix T . Recall that four participants provided acceptability judgments for all category-feature pairs, and that we considered a pair to be true if any of the four participants considered it to be acceptable. We reasoned that any pair considered acceptable by one of the four raters was also likely to be considered acceptable by at least some of the participants in our experiment. It is likely, however that some fraction of the entries in T result from idiosyncratic decisions on the part of a single participant. Only one of the four acceptability-raters indicated that ducks eat grass and only one indicated that cuckoos bite, and these two ratings end up making a significant contribution to the model predictions in Figures 12a.i and Figure 12b.ii. Future analyses could overcome this problem by collecting acceptability ratings from a larger group of participants and relying on a more stringent truth criterion—for example, a category-feature pair might be considered true only if two or more raters agree. Future analyses could also allow raters to express degrees of acceptability, which may be useful for distinguishing between pairs that are clearly true (cuckoos breathe air) and pairs that are questionable (cuckoos bite).

The errors discussed so far appear to depend primarily on methodological limitations, some of which can be addressed by future studies. In some cases, however, the failures of our model are more revealing. Figure 12a.ii shows a case where *killer whale* is arguably the best response by most sensible criteria. The two features (*black and white* and *sharp teeth*) were both generated for *killer whale* in the Leuven feature listing task, and no other category (including *zebra* and *tiger*) elicited both of these responses. The feature listing data therefore suggest that people understand that *killer whale* is strongly linked with these features, but for some reason it is difficult to retrieve *killer whale* when given the features and asked to generate an appropriate category. This result may reflect a genuine cognitive limitation—a limitation that may be diagnostic of the structure of mental representations. For example, the result may suggest that semantic memory is addressed more easily by categories (e.g. *killer whale*) than by features (e.g. *black and white*).

Figure 12a.i shows a second case where a limitation of our model may expose an important direction for future work. Failing to predict that *bat* and *parrot* are the top two responses does not seem like a serious failing, since each response was generated by only two participants. The real limitation of the model is that it fails to predict that the question in Figure 12a.i is more difficult than any of the remaining questions in our experiment. Notice, for example, that participants generated 29 or more responses for each question in Figure 3, but only 12 for the question in 12a.ii. Our model will happily provide some answer to any question that is posed, but future models can attempt to explain why people find some questions much harder than others, and why they sometimes fail to generate any response

at all.

Identification, categorization, and generalization

Although we have focused on category and feature identification, previous authors have discussed related problems including object identification, categorization and generalization. This section discusses some of the similarities and differences between these problems and provides a partial characterization of the full set of inductive problems that psychologists should ultimately aim to address. Kemp and Jern (2009) describe a more comprehensive taxonomy of inductive problems that is closely related to the account sketched here, but here we focus on the problems that are related most closely to identification.

Identification.

Consider first the problem of object identification. Suppose that you work in a wildlife park and have named many of the individual animals that you regularly see, including all members of a small herd of zebra. We will refer to each individual as an *object*. If you see a certain object (e.g. Henry the zebra) on two different occasions, we will say that you have encountered two *tokens* of that object. *Object identification* is the problem of deciding which object corresponds to a given object token (first row of Table 2). The problem of object identification is typically solved by inferring that an object token corresponds to a previously encountered object (e.g. Henry), but we will say that you have solved this problem if you correctly infer that an object token corresponds to an object that you have never previously encountered.

Although “identification” is sometimes used to refer exclusively to object identification, we suggest that there are several kinds of identification problems, including both category and feature identification. Category identification (second row of Table 2) can be defined as a problem where a category token is provided and the task is to infer the corresponding category. Feature identification can be defined as the problem of inferring which feature corresponds to a given feature token (third row of Table 2).

The problem of identification has been discussed by scholars from several disciplines, including psychology (James, 1890; Nosofsky, 1986; Rips, Blok, & Newman, 2006), machine learning (McCallum & Wellner, 2005; Milch et al., 2005), statistics (Bunge & Fitzpatrick, 1993), and philosophy (Millikan, 2000). Most prior modeling and empirical work has focused on the problem of object identification. Consider, for instance, the classic learning paradigm where participants are repeatedly shown objects that vary along a small number of dimensions (e.g. circles that vary in size and color). Each object can be given a unique identifier, and participants can be asked to identify each object each time it appears (Shepard, 1957; Nosofsky, 1986). Researchers working within this paradigm have described many empirical phenomena and attempted to explain them using computational models.

Category and feature identification have received less attention than object identification, and our names for these two inductive problems are non-standard. Both problems, however, have received some attention in the literature on word learning and reading comprehension. Werner and Kaplan (1952) explored how children use verbal context to learn word meanings, and developed a “Word Context test” where participants must infer the meaning of a novel word after hearing it used in several contexts. For example, after hearing that “you can make a corplum smooth with sandpaper” and that “painter used a corplum





Problem	Input	Output	Example Input	Example Output
Object identification	object token	object		 = Henry
Category identification	category token	category	Cs have stripes	C = <i>zebra</i>
Feature identification	feature token	feature	Zebras have F	F = <i>stripes</i>
Object categorization	object	category	Henry	<i>zebra</i>
Category categorization	category	category	<i>zebra</i>	<i>natural kind</i>
Feature categorization	feature	feature	<i>stripes</i>	<i>perceptual feature</i>
Object generalization	(object ₁ , feature ₁)	(object ₁ , feature ₂)	Henry has lost his teeth	Henry is very old
Feature generalization	(object ₁ , feature ₁)	(object ₂ , feature ₁)	Henry has sesamoid bones	Hilda has sesamoid bones
Category generalization	(category ₁ , feature ₁)	(category ₁ , feature ₂)	Zebras live in groups	Zebras communicate with each other
Feature generalization	(category ₁ , feature ₁)	(category ₂ , feature ₁)	Zebras have sesamoid bones	Horses have sesamoid bones
Object token generalization	(object token ₁ , feature ₁)	(object token ₁ , feature ₂)	 has lost his teeth	 is very old
Feature token generalization	(object ₁ , feature token ₁)	(object ₂ , feature token ₁)	Henry has F	Hilda has F
Category token generalization	(category token ₁ , feature ₁)	(category token ₁ , feature ₂)	Cs live in groups	Cs communicate with each other
Feature token generalization	(category ₁ , feature token ₁)	(category ₂ , feature token ₁)	Zebras have F	Horses have F

Table 2: Fourteen inductive problems. Identification is the problem of deciding which object, category or feature corresponds to an observed object, category or feature token. Categorization is the problem of organizing objects, categories or features into categories. Generalization is the problem of making inferences about unobserved features of objects, object tokens, categories, or category tokens.

to mix his paints” you might infer that a corplum is a stick. The same basic problem has also been addressed using the cloze procedure (Taylor, 1953), where some of the words in a passage are replaced with empty slots and a reader must identify the words that fill these slots. Computational approaches based on latent semantic analysis (Landauer & Dumais, 1997) and topic models (Griffiths, Steyvers, & Tenenbaum, 2007) can be used to address these problems. As Steyvers suggests in this issue, corpus statistics and feature listing tasks appear to capture somewhat different kinds of information, and ultimately it may be useful to develop accounts of category and feature identification that combine these two sources of data. Here, however, we explored an approach that relies on feature-listing data alone.

Categorization.

Although “category identification” is rarely discussed by psychologists, categorization has received a great deal of attention. Categorization can be defined as the problem of organizing items (typically objects) into categories (typically object categories). One example is the problem of deciding whether Henry is a zebra or a horse (fourth row of Table 2). By default we have used “category” to refer to a category of objects, but note that many other kinds of category are possible. For example, object categories such as *zebra* and *horse* are both members of the category *natural kinds*, and features such as *stripes* and *spots* are both members of the category *perceptual features* (fifth and six rows of Table 2).

Consider now the difference between category identification (second row of Table 2) and object categorization (fourth row of Table 2). Both problems may seem closely related—note, for instance, that the output in both cases is the category *zebra*. The relationship between the two seems even closer when we consider categorization problems where the input is linguistic rather than visual. For example, suppose that a friend tells you that “Henry has stripes” then asks you to infer which category Henry belongs to. This categorization problem may appear equivalent to the category identification problem in Row 2 of Table 2, but note that there is a subtle difference between the two problems. The feature provided is an object-feature in one case (“Henry has stripes”) but a category-feature in the other (“Cs have stripes”). We will argue that object-features and category-features should be distinguished, and it follows that the problems of categorization and category identification should also be distinguished.

The most obvious difference between object-features and category-features is that some object-features cannot apply to categories and some category-features cannot apply to objects. Consider a categorization problem where you learn that “Henry has lost his teeth.” A feature of this kind can be sensibly applied to an object (e.g. Henry) but not to a category (e.g. *zebra*). On the other hand, consider a category identification problem where you learn that “Cs are extinct.” As linguists and philosophers have emphasized, being extinct is a feature that can sensibly be applied to a category but not to an individual object such as Henry (Carlson, 2009). Features like *is extinct* might initially seem like exotic special cases, but features of this kind are directly relevant to the work described here. The Leuven feature-listing data include the feature *is extinct* in addition to features like *exists in different sizes*, *can have different colors*, and *often run over by cars*. These last three features can sensibly be attributed to objects—for instance, “Fred, who has a very tough spine, is often run over by cars.” We suspect, however, that the participants who generated these features were aiming for a category-level interpretation (“snakes are often

run over by cars”). As these examples suggest, the features in the Leuven database and in our own experiments are better viewed as category features rather than object features, and our work is aimed specifically at the problem of category identification rather than categorization.

Even in cases where a feature (e.g. *stripes*) can be applied to both objects and categories, the feature appears to carry different meanings in these two cases. The statement that “Henry has stripes” provides direct information about the physical appearance of Henry, but the statement that “zebras have stripes” does not indicate that the category *zebra* has a certain appearance. Instead, it suggests that members of this category are striped by default, although there may be exceptions (e.g. albino zebras or zebras that have undergone plastic surgery). We have already seen that category features (e.g. *is extinct*) can be different from object features, but the *stripes* example suggests in addition that the predication relationship between categories and category features is different from the corresponding relationship between objects and object features. For both of these reasons, categorization and category identification are best treated as distinct problems.

Although it is useful to distinguish between categorization and category identification, there are tasks that combine aspects of both problems. Suppose, for example, that you are introduced to an animal and told that “Henry is a C.” Although formulated as a category identification problem, this task is equivalent to an object categorization problem where you are simply presented with Henry and asked to infer his category. A category identification problem will typically specify category features (e.g. “Cs have stripes”) rather than examples of category members (“Henry is a C”), but some problems combine both kinds of information. For example, suppose that Hilda looks like a white horse and you are told that “Hilda is a C, but Cs are normally striped.” In solving this problem you will use both object features (e.g. the fact that Hilda has a mane) and category features (Cs are normally striped) to infer, for instance, that $C = \textit{zebra}$ and that Hilda is an albino. Since this task draws on both object features and category features, it follows that it is not equivalent to either a pure categorization problem or a pure category identification problem. Psychologists should ultimately aim to develop unified accounts of induction that can handle joint problems of this kind, but here we have focused on pure identification problems.

Generalization.

Although identification can be studied as an inductive problem in its own right, this problem can also arise as a component of other inductive tasks. Consider, for example, two tasks that we refer to as category token generalization and feature token generalization. In the first task a reasoner is given one or more features of a category token then asked to predict which other features the hidden category might have. If you learn, for example, that Cs have wings, you might be able to guess whether Cs are able to fly (compare row 13 of Table 2). Feature token generalization is a related problem where a reasoner is told that several categories share a hidden feature then asked to predict which other categories will have this feature. For example, if you learn that zebras have feature F, you might be able to predict whether horses also have this feature (row 14 of Table 2).

Our identification model can handle generalization problems involving tokens by inferring the likely identity of the hidden item and exploiting its pre-existing knowledge about

this item. For example, given that Cs have wings, our model might infer that category C could be eagle and might therefore conclude that Cs can fly. Our model, however, does not address generalization problems where there is no uncertainty about the identity of the category or feature involved. For example, given that *Melipotes carolae* was recently discovered in New Guinea and that *Melipotes carolae* has wings, you might infer that *Melipotes carolae* can fly (compare row 9 of Table 2). Similarly, given that zebras have sesamoid bones, you might infer that horses also have this feature (row 10 of Table 2). Since both problems refer to categories or features that are clearly novel, there is no identity uncertainty to resolve and our model does not apply.

In general, a language learner will not know whether a novel label is a new name for a familiar category or feature or whether this label picks out a category or feature that is genuinely new. For example, before learning about the discovery in New Guinea you might be unsure whether *Melipotes carolae* is the scientific name for a familiar category or the name of a novel category. Young children face a similar kind of uncertainty—for example, a word like “cutlery” could refer to a category that a child has already noticed or to a new category that needs to be learned. Maratsos (2001) acknowledges both possibilities by distinguishing between two inductive problems: “category recognition,” or the identification of pre-existing categories, and “category assembly,” or the construction of new categories. In terms of this distinction, our work has focused on category recognition rather than category assembly, but future work can aim to develop models that address both problems.

Conclusion

Humans draw on semantic knowledge to address many kinds of inductive problems, including problems that require inferences about categories and their features. This paper considered the problem of identification. We presented a probabilistic account of identification that helps to explain how identification is guided by prior knowledge, and how humans integrate multiple sources of information when solving identification problems. Our experiments relied on three simple laboratory tasks, but our general approach may help to explain how first- and second-language learners acquire novel labels for pre-existing concepts.

Although this paper focused on the problem of identification, previous probabilistic approaches have addressed other inductive problems including categorization and generalization. Future work should aim to develop a unified framework that handles all of these problems. Humans solve many inductive problems that are related but distinct, and a probabilistic approach can help to clarify the relationships between these problems and to identify common principles that support solutions to all of them.

References

- Anderson, J. R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3), 409–429.
- Anderson, N. H. (1981). *Foundations of information integration theory*. New York: Academic Press.
- Bloom, P. (2000). *How children learn the meanings of words*. Cambridge, MA: MIT Press.

- Bunge, J., & Fitzpatrick, M. (1993). Estimating the number of species: a review. *Journal of the American Statistical Association*, *88*(421), 364–373.
- Carlson, G. (2009). Generics and concepts. In F. J. Pelletier (Ed.), *Kinds, things and stuff: mass terms and generics*. New York, NY: Oxford University Press.
- Chomsky, N. (1991). Linguistics and cognitive science: Problems and mysteries. In A. Kasher (Ed.), *The Chomskyan turn* (pp. 3–25). Cambridge, MA: Basil Blackwell.
- Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240–247.
- De Deyne, S., Verheyen, S., Ameel, E., Vanpaemel, W., Dry, M. J., Voorspoels, W., et al. (2008). Exemplar by feature applicability matrices and other Dutch normative data for semantic concepts. *Behavior Research Methods*, *40*(4), 1030–1048.
- Fodor, J. A. (1975). *The language of thought*. Cambridge: Harvard University Press.
- Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review*, *114*(2), 211–244.
- Heit, E. (1998). A Bayesian analysis of some forms of inductive reasoning. In M. Oaksford & N. Chater (Eds.), *Rational models of cognition* (p. 248–274). Oxford: Oxford University Press.
- James, W. (1890). *Principles of psychology*. New York: Holt.
- Kemp, C., Goodman, N. D., & Tenenbaum, J. B. (2008). Learning and using relational theories. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in Neural Information Processing Systems 20* (pp. 753–760). Cambridge, MA: MIT Press.
- Kemp, C., & Jern, A. (2009). A taxonomy of inductive problems. In *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 255–260). Austin, TX: Cognitive Science Society.
- Kemp, C., & Tenenbaum, J. B. (2009). Structured statistical models of inductive reasoning. *Psychological Review*, *116*(1), 20–58.
- Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, *104*, 211–240.
- Lombardi, L., & Sartori, G. (2007). Models of relevant cue integration in name retrieval. *Journal of Memory and Language*, *57*, 101–125.
- Maratsos, M. (2001). How fast does a child learn a word? *Behavioral and Brain Sciences*, *24*(6), 1111–1112.
- McCallum, A., & Wellner, B. (2005). Conditional models of identity uncertainty with application to noun coreference. In L. K. Saul, Y. Weiss, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems 17* (pp. 905–912). Cambridge, MA: MIT Press.
- McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, and Computers*, *37*(4), 547–559.
- Medin, D. L., & Schaffer, M. M. (1978). Context theory of classification learning. *Psychological Review*, *85*, 207–238.
- Mervis, C. B. (1987). Child-basic object categories and early lexical development. In U. Neisser (Ed.), *Concepts and conceptual development* (pp. 201–233). Cambridge: Cambridge University Press.

- Milch, B., Marthi, B., Russell, S., Sontag, D., Ong, D. L., & Kolobov, A. (2005). BLOG: Probabilistic models with unknown objects. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence* (pp. 1352–1359).
- Millikan, R. G. (2000). *On clear and confused ideas: an essay about substance concepts*. New York: Cambridge University Press.
- Mitchell, T. M. (1997). *Machine learning*. New York: McGraw Hill.
- Navarro, D. J., & Perfors, A. F. (this issue). Similarity, feature discovery and the size principle. *Acta Psychologica*.
- Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*, 39–57.
- Oden, G. C., & Massaro, D. W. (1978). Integration of featural information in speech perception. *Psychological Review*, *85*(3), 172–191.
- Osherson, D. N., Smith, E. E., Wilkie, O., Lopez, A., & Shafir, E. (1990). Category-based induction. *Psychological Review*, *97*(2), 185–200.
- Rips, L. J., Blok, S., & Newman, G. (2006). Tracing the identity of objects. *Psychological Review*, *113*(1), 1–30.
- Robertson, S. (2004). Understanding inverse document frequency: on theoretical arguments for IDF. *Journal of Documentation*, *60*(5), 503–520.
- Sartori, G., & Lombardi, L. (2004). Semantic relevance and semantic disorders. *Journal of Cognitive Neuroscience*, *16*(3), 439–452.
- Shepard, R. N. (1957). Stimulus and response generalization: a stochastic model relating generalization to distance in psychological space. *Psychometrika*, *22*, 325–345.
- Shepard, R. N. (1987). Towards a universal law of generalization for psychological science. *Science*, *237*, 1317–1323.
- Steyvers, M. (this issue). Incorporating features into statistical topic models. *Acta Psychologica*.
- Taylor, W. L. (1953). “Cloze procedure:” a new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.
- Werner, H., & Kaplan, E. (1952). The acquisition of word meanings: a developmental study. *Mono-graphs of the Society for Research in Child Development*, *15*(1), 1–119.
- Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*(2), 245–272.

Appendix
Experimental stimuli

Tables A1, A2 and A3 show the stimuli used in our three experiments. In Table A3 the first row corresponds to the question:

cow, feature F
animal C, feature F
animal C, has antlers

Since the Leuven data include many idiosyncratic features, we chose not to generate the stimuli by sampling from the distribution assumed by our model. We instead selected the questions for each experiment by hand, taking care to ensure that each question had at least one sensible answer, and that many questions had several good answers.

stings		
is edible		
has wings		
climbs trees		
has fur		
eats grass		
frightens people		
has a long neck		
has sharp teeth		
is a pet		
has a long neck		has a beak
is grey		has sharp teeth
is white		has feathers
jumps		has six legs
is slow		lives in Africa
is yellow		has sharp teeth
is green		is a predator
is a pet		sings
is smart	used as a laboratory animal	
is nocturnal	has sharp teeth	
is edible	jumps	
is yellow and black	has fur	
lives in cold areas	has a beak	
can fly	has fur	
is a pet	has two legs	
is black and white	has sharp teeth	
is hunted	is yellow and black	
can fly	eats grass	
has two legs	is nocturnal	
has four legs	is pink	
has a long neck	has wings	can't fly
is green	has a rough skin	has a long tail
eats mice	is nocturnal	has four legs
is edible	is hunted	eats grass
is a pet	has many colors	lives in a cage
jumps	is small	bites
has long ears	is grey	has a long tail
is black and white	can swim	lays eggs
eats plants	has fur	is brown
is dangerous	is yellow	has a tail
has many colors	has a tail	has two legs
is long	is poisonous	has a tongue
is black	is unhygienic	buzzes
lives in water	has two legs	has feathers
has whiskers	has sharp teeth	lives in Africa
can be ridden	is grey	lives in Africa
is white	mammal	eats grass
is striped	has four legs	runs fast
has hooves	is pink	is round
has claws	has feathers	is nocturnal

Table A1: Stimuli used in Experiment 1 (category identification).

cow		
ant		
salmon		
canary		
mouse		
butterfly		
ant	fly	
eagle	penguin	
giraffe	swan	
ostrich	butterfly	
squid	beaver	
turkey	woodpecker	
duck	sparrow	
crocodile	python	
goldfish	hamster	
salmon	trout	
cockroach	ladybug	
grasshopper	kangaroo	
bee	dove	moth
cuckoo	turtle	flea
donkey	rhinoceros	shark
frog	grasshopper	turtle
dolphin	mouse	bat
alligator	cobra	piranha
ant	blackbird	killer whale
zebra	lizard	peacock
cod	squid	penguin
deer	hedgehog	squirrel
snake	goldfish	sardine
dove	polar bear	swan
alligator	sparrow	wasp
crocodile	seagull	hippopotamus
frog	lizard	spider
salmon	turkey	sheep
alligator	wolf	shark
bee	mosquito	wasp
bumblebee	dragonfly	beetle
shark	dolphin	trout
python	eel	caterpillar
pig	cow	salmon
deer	lion	fox
mosquito	bumblebee	fly
viper	wasp	crocodile
peacock	chameleon	butterfly
cricket	cockroach	turtle
tiger	falcon	deer
penguin	salmon	turtle
cow	kangaroo	zebra
cat	lizard	zebra
owl	mouse	moth

Table A2: Stimuli used in Experiment 2 (feature identification).

cow	has antlers
cow	is pink
cow	has whiskers
cow	lays eggs
salmon	has fins
salmon	lives in cold areas
salmon	chews the cud
salmon	builds dams
canary	has a mane
canary	crows
canary	can purr
canary	has a red breast
mouse	can see in the dark
mouse	eats bananas
mouse	climbs trees
mouse	swims in a bowl
butterfly	sings
butterfly	has six legs
butterfly	lives in a cage
butterfly	makes honey
penguin	is white
penguin	lays big eggs
penguin	has fins
penguin	is striped
polar bear	eaten as meat
kangaroo	eaten as meat
cod	eaten as meat
elephant	eaten as meat
mouse	has wings
ant	has wings
polar bear	has wings
canary	has wings
salmon	has sharp teeth
penguin	has sharp teeth
grasshopper	has sharp teeth
owl	has sharp teeth
chicken	is a pet
squirrel	is a pet
salmon	is a pet
tiger	is a pet
cat	frightens people
salmon	frightens people
owl	frightens people
deer	frightens people
lion	has a long neck
robin	has a long neck
zebra	has a long neck
chicken	has a long neck
swan	has fur
bee	has fur
grasshopper	has fur
turtle	has fur
penguin	is white
ant	has many colors
fly	is nocturnal
pig	chews the cud
sheep	eats fish
hippopotamus	lives in the sea
donkey	is striped
chicken	is big

Table A3: Stimuli used in Experiment 3 (joint category and feature identification).