Are theories necessary to constrain concepts?

Timothy T. Rogers
Center for the Neural Basis of Cognition
Department of Psychology
Carnegie Mellon University

James L. McClelland
Center for the Neural Basis of Cognition
Department of Psychology
Carnegie Mellon University and
Department of Neuroscience University of Pittsburgh

Poster presented at the 2000 meeting of the Cognitive Neuroscience Society

Abstract

In traditional theories of semantic memory, performance of semantic tasks relies upon a mediating process of categorization. However, categorization-based theories do not capture the complex and flexible ways in which people use their conceptual knowledge to perform natural semantic tasks imposed on them by the environment. For example, both children and adults understand that a given property may be important for categorizing some kinds of objects, but not others; that different kinds of properties generalize across different groups of objects; and that insides can be more important for determining category membership than outsides. Consequently, some researchers propose to describe conceptual knowledge in terms of naive theories about causal mechanisms. In the current work, we present simulations using a simple connectionist network that learns the mappings between objects and their properties in different contexts. We show that the evolution of representations throughout learning in our model constrains the ease with which particular object properties can be learned, and how they will generalize. The configuration of weights at any point during development may provide the kinds of 'enabling constraints' on acquisition that some researchers attribute to naive theories. Many of the phenomena that arise in the theory-theory tradition may be understood within this framework. Knowledge about how object properties vary across contexts is stored in connection weights that are learned from experience. This knowledge plays the role that naive theories play in the theory-theory framework.

Introduction

Theories of conceptual knowledge that emphasize the role of learning and experience in acquisition have come under fire in recent years. Such theories are generally thought to be too underconstrained to adequately explain conceptual development, without additional explanatory constructs, such as implicit theories. Among the phenomena that would seem to support this view are the following:

- *Illusory correlations:* children and adults may create or enhance some object-property correlations, while ignoring others.
- *Feature centrality:* A given feature may be important for some categories of objects, but not others.
- *Flexible generalization:* Children and adults can generalize their knowledge in ways that challenge simple similarity-based mechanisms.
- *Expertise:* Different kinds of experts may acquire different representations of objects in the same domain.

We have been investigating the capacity of the parallel distributed processing (PDP) framework to provide a general theory of semantic memory. Our approach builds upon earlier work by Hinton (1981) and Rumelhart (Rumelhart, Smolensky, McClelland, & Hinton, 1986; Rumelhart & Todd, 1993). Under the PDP theory, semantic memory is encoded in the weights of a connectionist network, which must learn the mappings between objects and their properties in different contexts. Domain-general learning mechanisms sensitive to the structure of the environment lead the system to gradually acquire correct mappings; and in so doing, to discover abstract, distributed representations of objects that capture their deep similarity relations in the context of a particular task. Thus, under this view, the development of conceptual knowledge is largely driven by experience. Learned similarities among the system's internal representations provide a mechanism for knowledge generalization and induction. However, because knowledge about a given object and a particular task both provide graded constraints on the system's internal states, different kinds of knowledge may generalizae across different groups of objects.

We believe this framework provides a powerful set of tools for understanding human performance in semantic tasks. However, the PDP theory clearly relies to a great extent on mechanisms of learning to explain conceptual development. How might it explain the empirical observations that seem to undermine learning-based theories?

A simple implementation of the theory, adapted from Rumelhart and Todd (1993), is shown in Figure 1. Input units appear on the left, and activation propagates from the left to the right. Where connections are indicated, every unit in the pool on the left is connected to every unit in the pool to the right. Each unit in the *Item* layer corresponds to an individual object in the environment. Each unit in the *Context* layer represents contextual constraints on the kind of information to be retrieved. Thus, the input pair *canary can* corresponds to a situation in which the network is shown a picture of a canary, and asked what it can do. The network is trained to turn on all those units that represent correct completions of the input query. In the example shown, the correct units to activate are *grow, move, fly,* and *sing*. To find a set of weights that allow the model to perform correctly, it is trained with backpropagation. As small changes to the weights accumulate, the network gradually acquires distributed internal representations of the various items that capture their semantic relations.

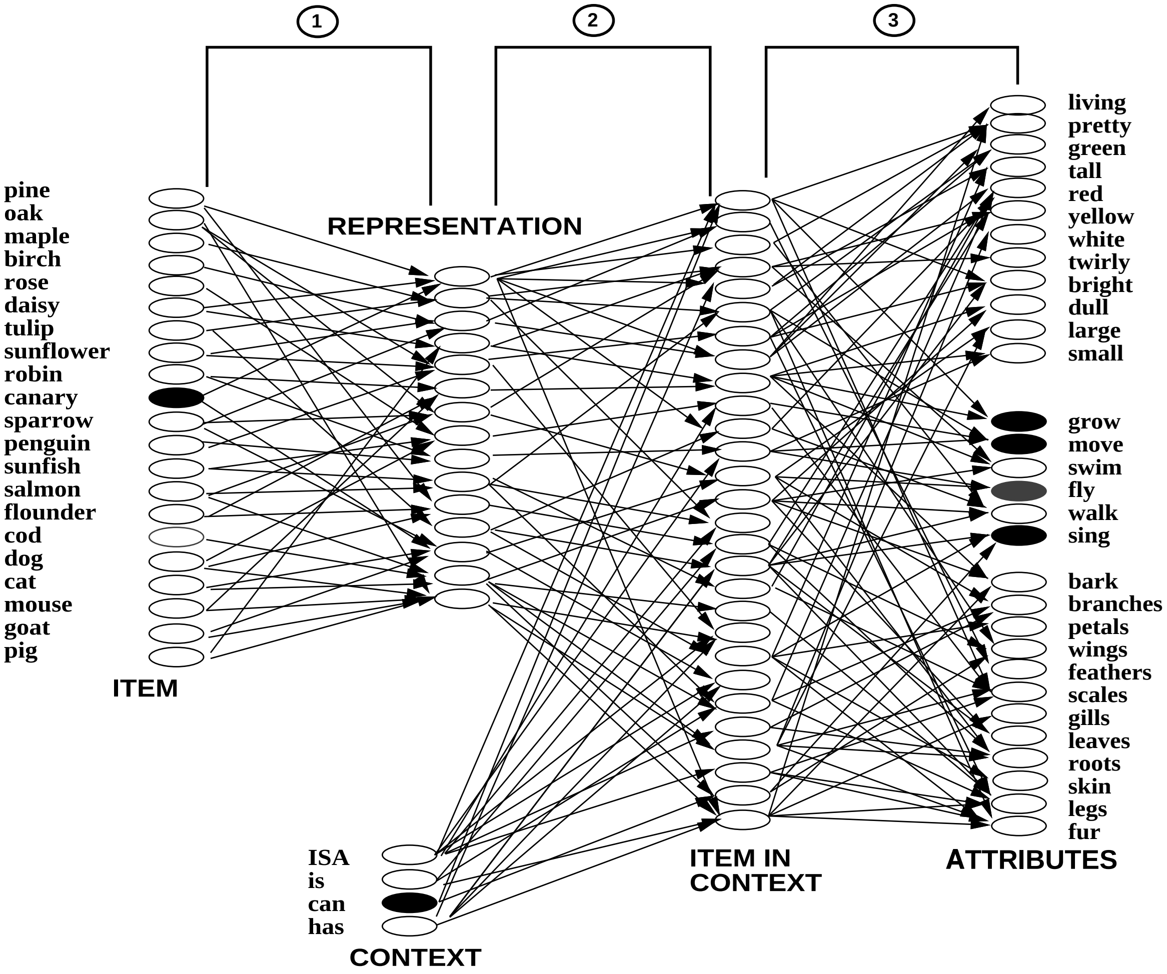


Figure 1. A simple feed-forward implementation of the theory, based on the model proposed by Rumelhart and Todd (1993).

The first layer of weights maps each individual input unit to a distributed pattern of activity across the units in the layer labeled *Representation*. Initially, all the weights in the network are small and random, and the patterns of activity corresponding to various items are all similar. As the network's weights change to improve its performance, these internal representations gradually differentiate. Figure 2 shows a multidimensional scaling of the network's internal representations of all 21 items, at ten different points during training. The proximity of points in the diagram indicates the degree to which their internal representations are similar. Each line corresponds to a single item, and traces the trajectory of that item's representation throughout learning. The figure shows that initially, all representations are similar to one another. The model first differentiates items into global categories (plants and animals), and only later differentiates finer-grained categories. To the extent that two items have similar representations, the network is pressured to generalize its knowledge from one to the other.

1

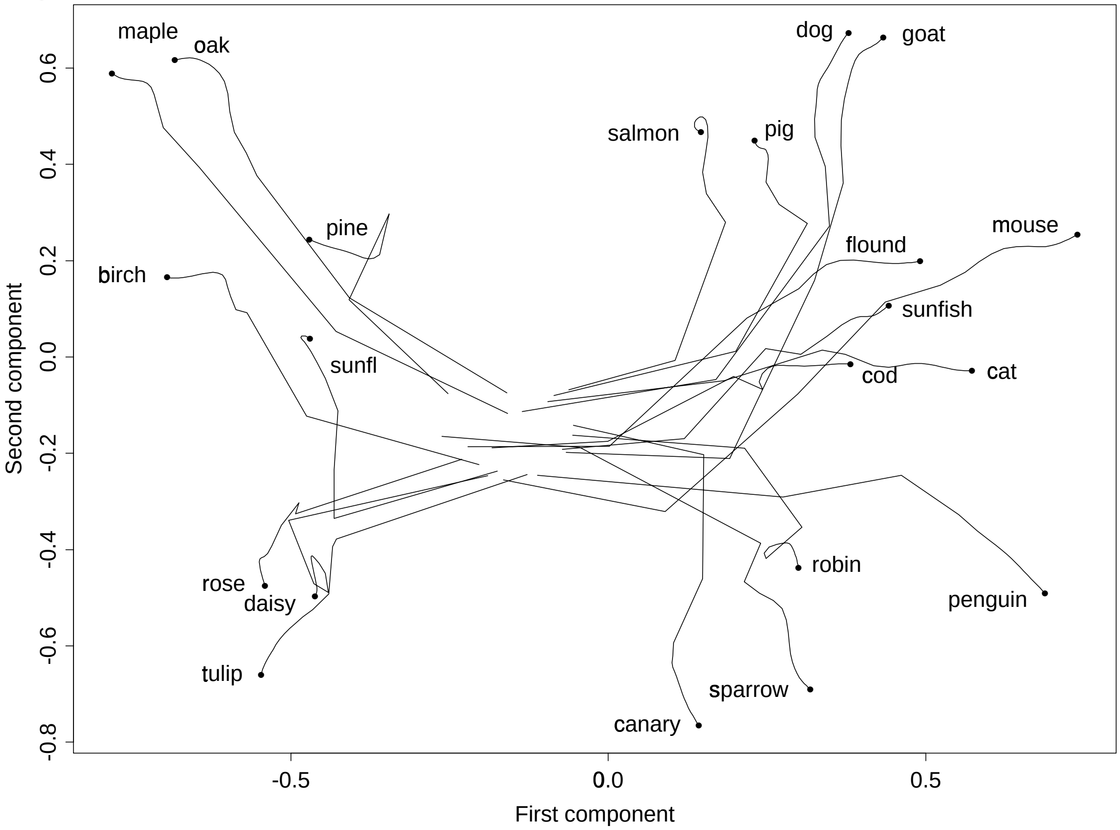


Figure 2. A multidimensional scaling of the model's internal representations of objects, at several points during training. See text for explanation.

The second layer of weights, including those projecting from the *Context* layer, capture the similarities among various items in a particular context. The weights projecting from the *Representation* layer and the *Context* layer provide graded constraints on the pattern of activity generated across units in the layer labeled *Item in context*. In this layer, the same object may be represented differently, depending upon the context. Figure 3 shows a multidimenstional scaling of the representations of sixteen items in the *Representation* layer (middle); and in the *Item in context* layer, with either the *is* (top) or the *can* (bottom) context units active. In the *is* context, representations of objects are fairly spread out. In the *can* context, all the plants are collapsed to an essentially identical representation, because they all share the same behavior—the only thing they can do is grow.

The third layer of weights captures the mappings between representations of items in context, and the actual properties that may be inferred. Thus, as shown in Figure 4, the canary activates a completely different set of output units in the can and is contexts.

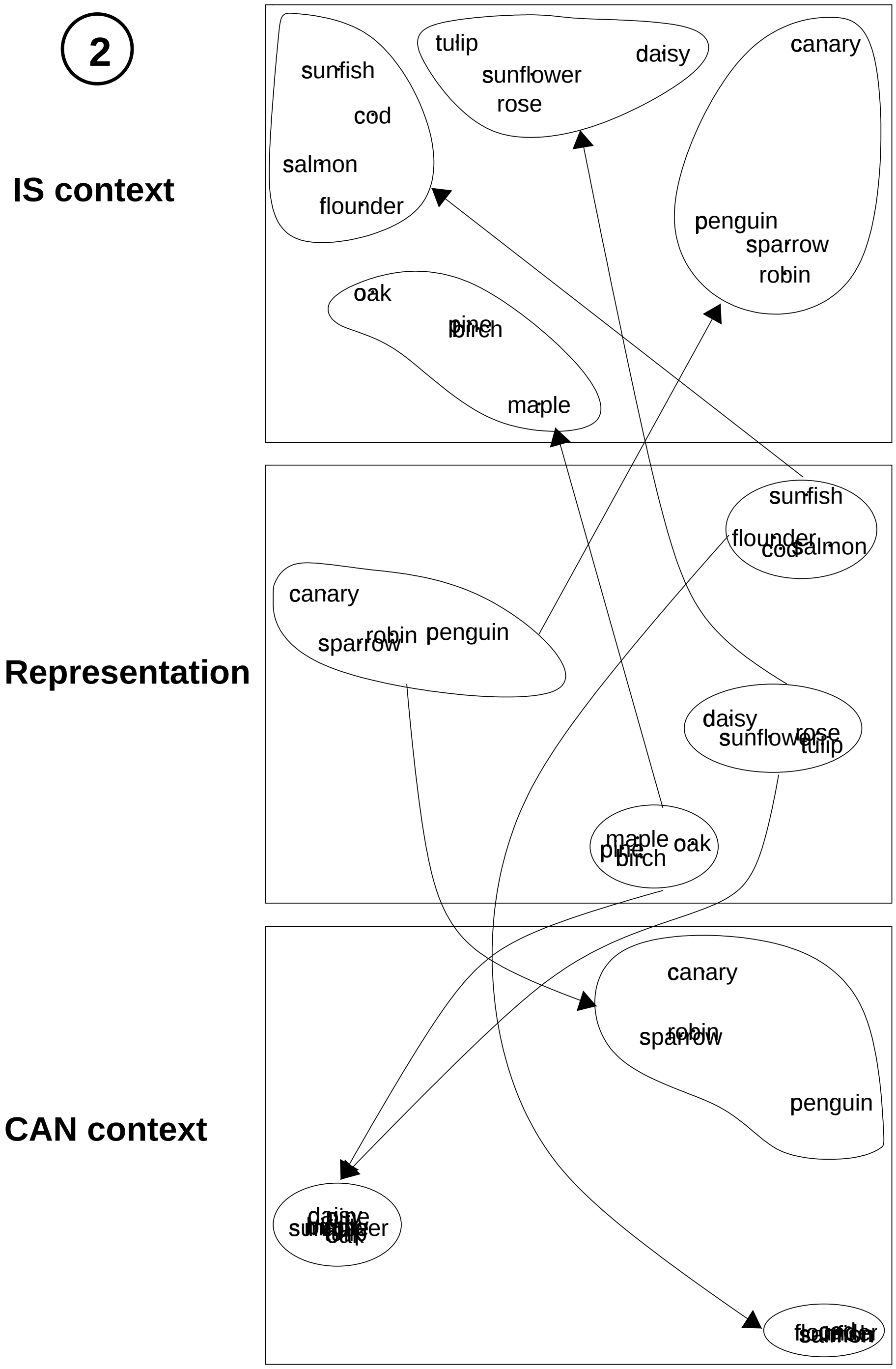


Figure 3. A multidimensional scaling of the model's internal representations of objects independent of context (middle), and across the *Item in context* layer, with the *is* (top) and the *can* (bottom) context units active in the input. The illustration shows that different similarity relations among the same items are represented in different contexts.

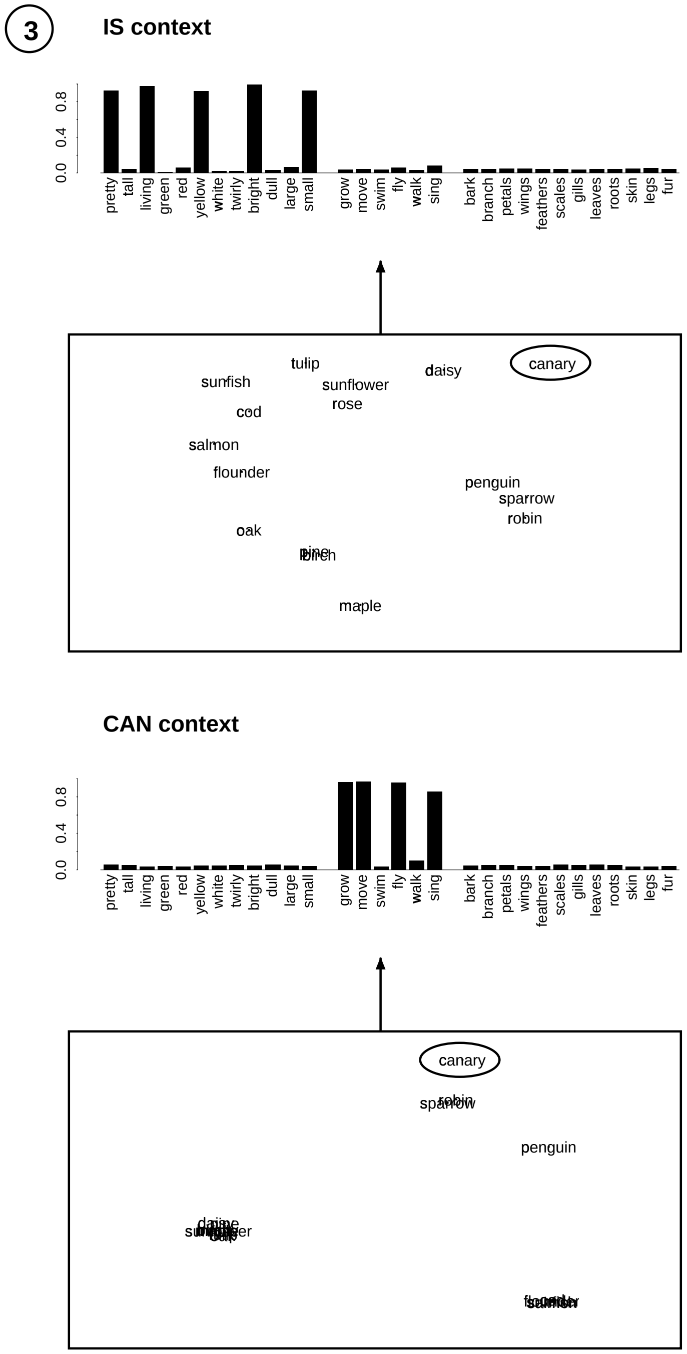


Figure 4. The final set of weights capture the mappings between representations of an object in a given context, and explicit object properties. The top illustration shows the set of output units that activate for the item *canary*, when it is probed in the *is* and *has* contexts. Obviously, the same item activates a completely different set of properties in the two different contexts. Note that, because different similarity relations are represented in the two contexts, the network will tend to generalize *has* and *is* properties across different groups of objects.

Illusory correlations

Several studies have shown that subjects may persist in the belief that particular objects and properties have occurred together frequently, even in the face of empirical evidence to the contrary; and they may discount or ignore the co-occurrence of object-property pairs during learning, on the basis of past experience (e.g. Keil, 1991). For example, Massey and Gelman (1988) showed preschoolers static pictures of various unfamiliar objects, living and nonliving, and asked them to decide whether each object would be able to move up and down a hill on its own. Some of the nonliving things were shaped like animate objects (e.g. a figurine), whereas some of the living things were extremely atypical animals (e.g. an echidna). The ontological status of each item (i.e. living or nonliving) was not revealed to the child, but had to be inferred from the appearance of the object. After making their decision, the authors asked children to explain their choices. The following protocols demonstrate that children often referred to properties of objects that were clearly not present in the picture, and ignored the presence of other properties that were not consistent with their decision:

> M.B. (3 yrs 7 mos)
>
> **Insect-eye Figurine**
>
> DOWN BY SELF? No. WHY NOT? Because it's metal and doesn't have any shoes or feet.
>
> ARE THOSE FEET? (E POINTS TO THE FEET.) No.
>
> **Echidna**
>
> DOWN BY SELF? Yeah. WHY? It has feet and no shoes.
>
> CAN YOU POINT TO THE FEET? I can't see them.

Illusory correlations seem to present a problem for learning-based theories of conceptual representation: if children acquire their semantic knowledge simply by learning which objects have which properties, why should they make these claims, which are clearly in violation of their perceptual experience? The model suggests one answer: early in learning, it assigns similar representations to all the plants, including the pine tree. Because most of the plants have leaves, this property is learned quickly and generalizes to the pine tree by virtue of representational similarities. Even though the network is taught repeatedly that the pine tree does not have leaves, such learning is overwhelmed by similarity-based generalization. As the pine differentiates itself from the other plants, these forces diminish, and the model ultimately learns the correct response. These properties are shown in Figure 5: at epoch 1500, when the model has been trained 150 times that the pine does not have leaves, it nevertheless strongly activates the *has leaves* unit when probed with the *pine*. The same influences prevent the network from activating the property *can sing* for the canary until it is sufficiently distinguished from related objects, all of which can not sing.

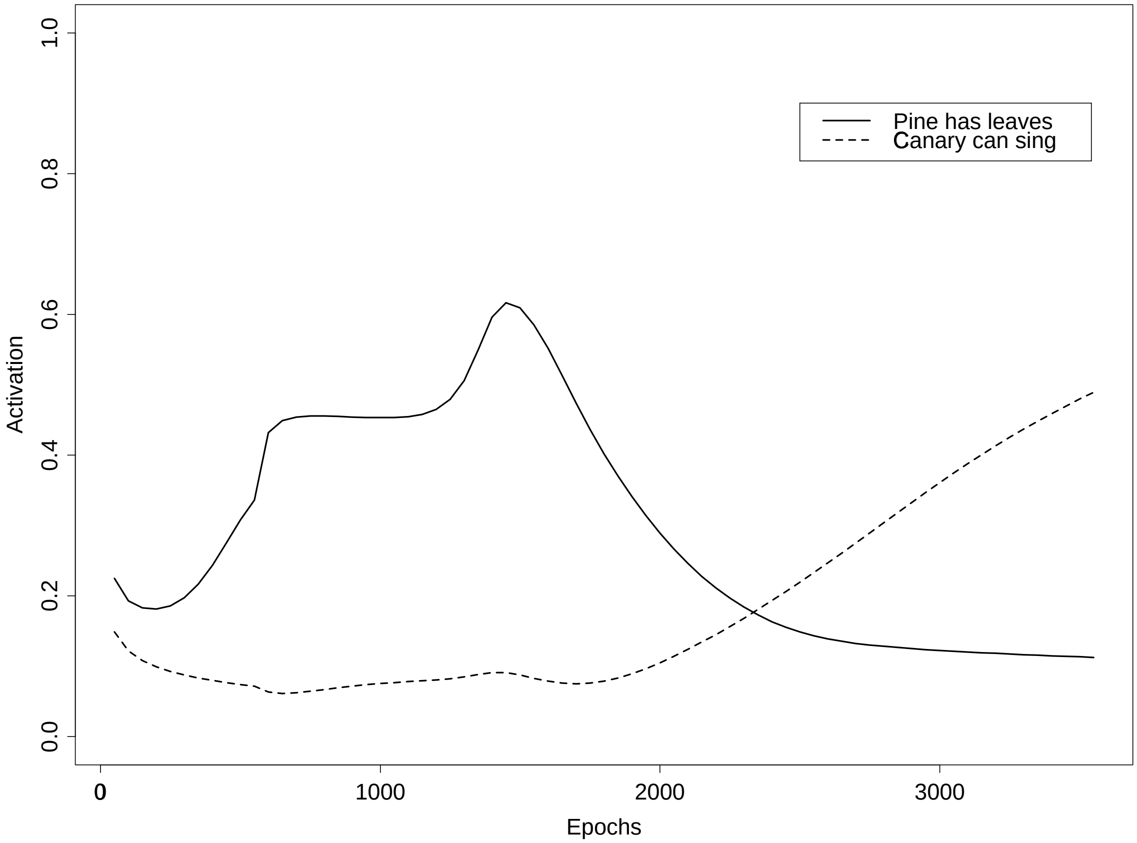


Figure 5. The response of two output units (*has leaves* and *can sing*) when the network is probed for its knowledge of the *pine* and *canary*, respectively. See text for interpretation.

Feature centrality

A second problem for learning-based theories of concept acquisition is that the relevance of a given feature can vary from category to category. For example, color appears to be more important for discriminating among foods than toys. To demonstrate this, Macario (1991) showed children a collection of wax objects that varied in shape and color. The children were introduced to a toy alligator, and were taught a new fact about one of the nonsense objects. In one condition, children were told that the toy alligator liked to eat the object; in the other, they were told the alligator liked to play with the object. They were then asked to guess which other nonsense objects the alligator might like to eat or play with. When lead to believe the objects were a kind of food, children inferred that other objects of the same color would be good to eat. When they thought the objects were kinds of toys, they ignored color information and inferred that objects of the same shape would be fun to play with. Without any foreknowledge of the categories food and toy, how can the semantic system figure out which properties are important and which irrelevant?

In the model's environment, properties were assigned to items in such a way that size, but

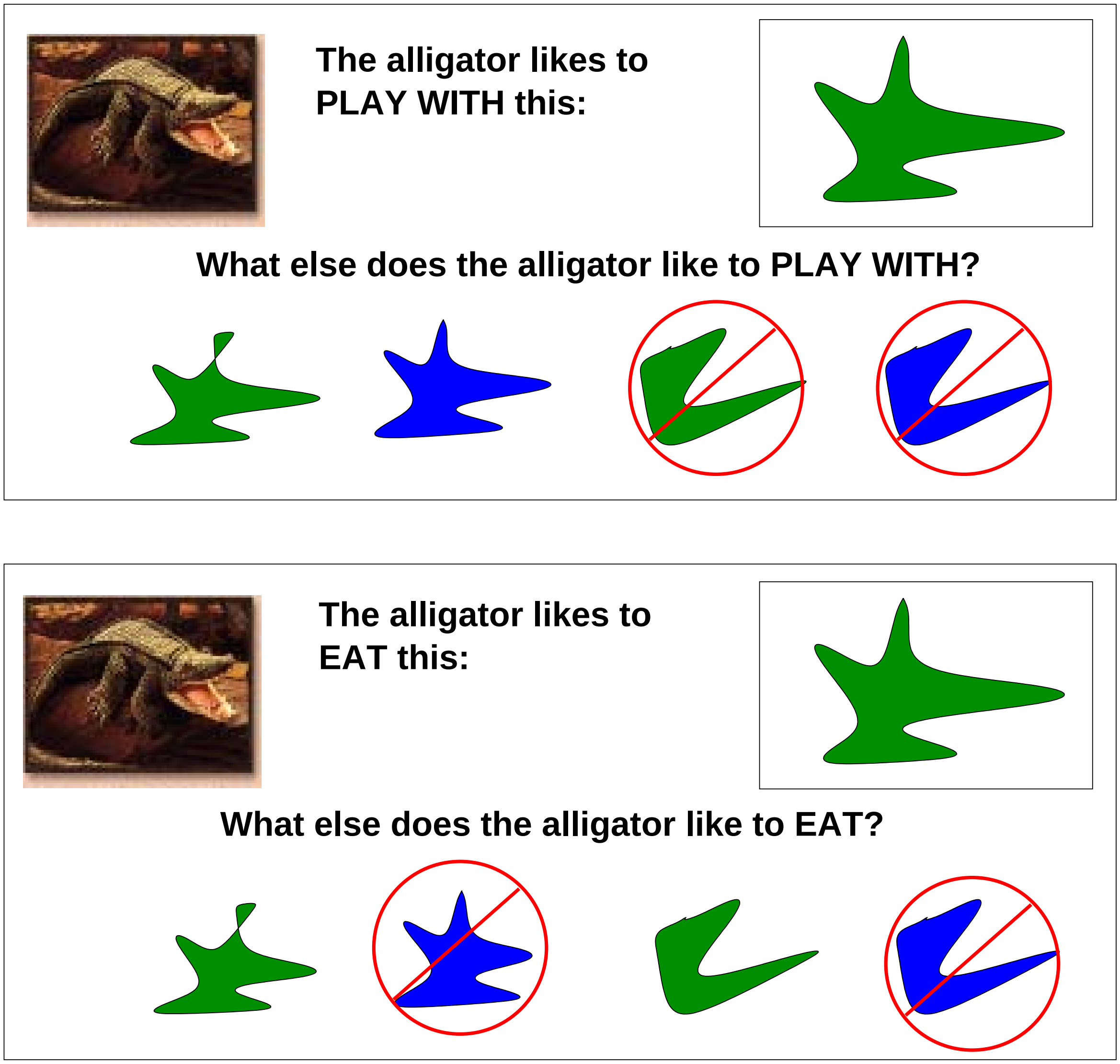

The alligator likes to PLAY WITH this:
What else does the alligator like to PLAY WITH?
The alligator likes to EAT this:
What else does the alligator like to EAT?

Figure 6. A schematic of Macario's (1991) experiment, showing that color seems to be more important than shape for determining similarity among food objects; but that the reverse is true for toys. See text for interpretation.

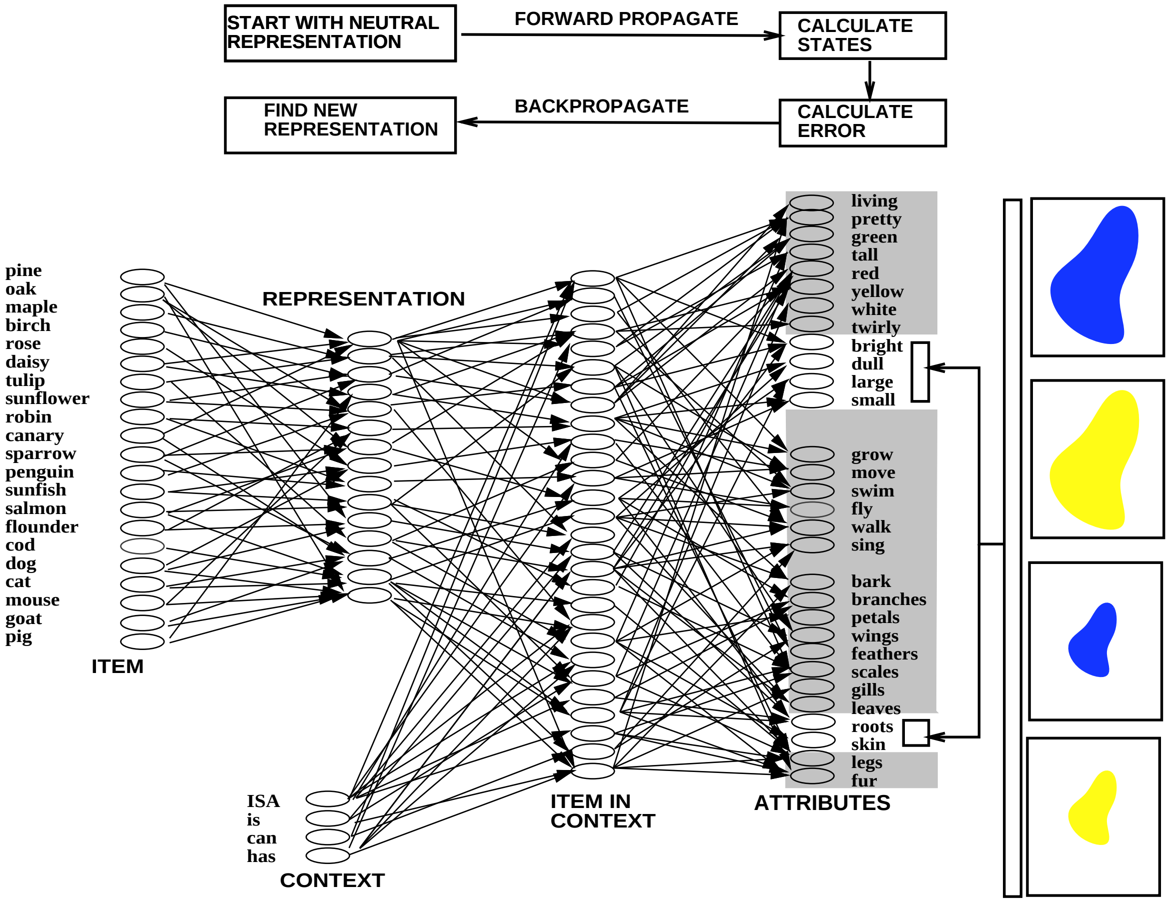


Figure 7. The figure illustrates how the model can represent novel objects with familiar properties—see text for details.

not brightness, was important for discriminating between the trees and flowers; and brightness, but not size, was important for discriminating between the birds and fish. Thus, among the animals, all the birds were bright and all the fish were dull, but a given bird or fish could be either large or small. The reverse was true for plants: trees were large and flowers were small, but a given tree or flower could be either bright or dull.

To simulate Macario's (1991) experiment, we presented the model with novel nonsense objects varying in size (large or small) and brightness, and allowed it to find appropriate representations for each, by backpropagating activation from the appropriate output units. For example, to find an appropriate representation for a large, bright object, we would choose a neutral starting pattern across the *Representation* units, forward-propagate activity to the output units, and calculate the error signal across the units large and bright. We would then use this error signal to adjust the pattern of activity across *Representation* units.

In the first simulation, we used this procedure to find appropriate representations for four objects varying in size and brightness, all of which shared a property common to plants (*has roots*). In the second, the network found representations for the same four items, but was taught that they all shared a property with animals (*has skin*).

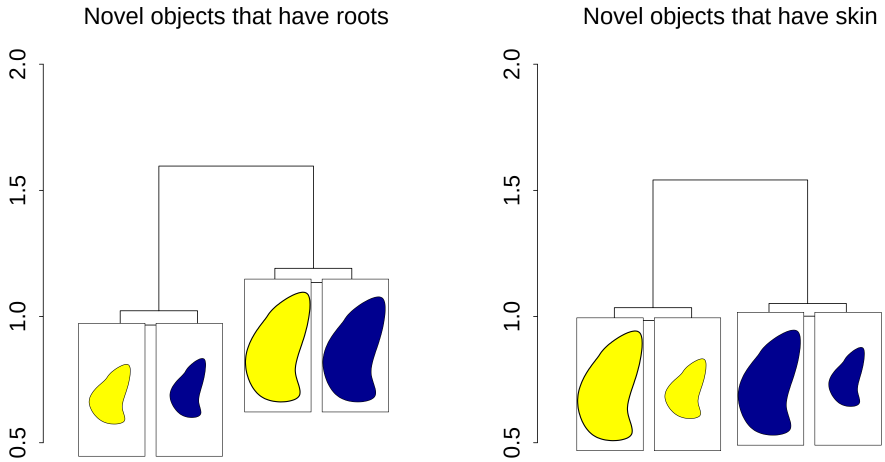


Figure 8. Hierarchical cluster analysis showing the similarities among representations discovered by the network for four novel objects varying in size and brightness. When the four objects shared a property with plants, they were grouped by the model on the basis of size. When they shared a property with animals, they were grouped on the basis of brightness.

Figure 8 shows a hierarchical cluster analysis of the internal representations derived by the network, given that the four nonsense objects shared a property with plants or with animals. When they shared a property with plants, the network grouped them on the basis of size; when they shared a property with animals, it grouped them on the basis of color. Thus the network seems to “know” that color is important for discriminating animals; but size is important for discriminating plants. This knowledge is encoded in the weights that allowed the model to find appropriate representations for the objects, and was learned from the statistical structure of its environment. In the network’s world, all trees are big and all flowers are small; but a given tree or flower can be either bright or dull. The opposite is true for animals. Hence, the relative contribution of size or brightness to the representations derived by the network depends on the extent to which the stimulus context—the other features that make up the stimulus—are animal- or plant-like.

Flexible generalization

Children and adults may also generalize newly-acquired knowledge in quite sophisticated ways. For example, children at quite young ages know that different kinds of properties generalize across different groups of objects. Such flexibility is sometimes taken to undermine similarity-based theories of concept knowledge. The model illustrates how the integrated influences of object and context knowledge may conspire to support such sophisticated patterns of generalization.

Figure 9 shows how we teach the network a new fact about a familiar object. We first add a new unit to represent the novel predicate (e.g. *queem*), and a new set of weights connecting this unit to the *Item in context* layer. Next, we activate the appropriate *Item* and *Context* inputs, and train the network to activate the new output unit. In this case, to form an association between the model's internal representation and the new property without disrupting its knowledge of the domain, we allow it to adjust only the weights projecting to the new property unit. Once the network has learned, we may probe it with other inputs and contexts, to see how its newly acquired knowledge generalizes.

Figure 10 shows how the network generalizes a newly learned property in different contexts, at two different points during learning. After 500 epochs of training, the network generalizes the new property to all items in the same superordinate category, regardless of the context in which the new fact was learned. Later in learning, however, the model generalizes the newly learned property differently, depending on whether it is an is, can, or has property. Thus, the model seems to know that different properties generalize across different groups of objects. Moreover, this capacity shows a developmental progression.

The reason the network shows this behavior is that different contexts bring out different similarity relations among items, as shown above in Figure 3. The constraints provided the *Context* weights evolve gradually, as do the network's internal representations of objects. Together, these conspire to produce different patterns of generalization throughout development.

Young children's induction behavior is sophisticated enough that they may generalize newly learned facts in ways that are entirely appropriate, but which violate taxonomic constraints. In one study, children were shown a picture of a brontosaurus and a picture of a rhinocerous (Gelman & Markman, 1986). They were taught the names of the animals, and a new fact about each. Half the children were taught a biological fact (e.g. the dinosaur has cold blood, but the rhinocerous has warm blood), and half were taught a physical property (e.g. the dinosaur weighs one hundred tons, but the rhinocerous weighs one ton). The children were then shown a picture of a triceratops, which they were told was another kind of dinosaur; however, the triceratops more closely resembled the rhinocerous. Children who learned the biological fact extended it to the triceratops on the basis of category membership; that is, they were more likely to say that the triceratops was cold-blooded like the brontosaurus. By contrast, children who were taught the physical property were less likely to use the category as the basis for generalization: most either guessed that the triceratops was one ton, like the hippo, or were uncertain.

To demonstrate the flexibility that can be captured by the model, we simulated Gelman and Markman's (1986) experiment, using a combination of the techniques described above. We first showed the model a novel brontosaurus, with the visible properties *is large, is bright, has skin,* and *has legs*; and a novel rhinocerous, with the properties *is small, is dull, has skin,* and *has legs*. Using the procedure described above, we allowed the network to derive appropriate representations for these items. We then taught the model three new facts about each: a name ("dinosaur" or

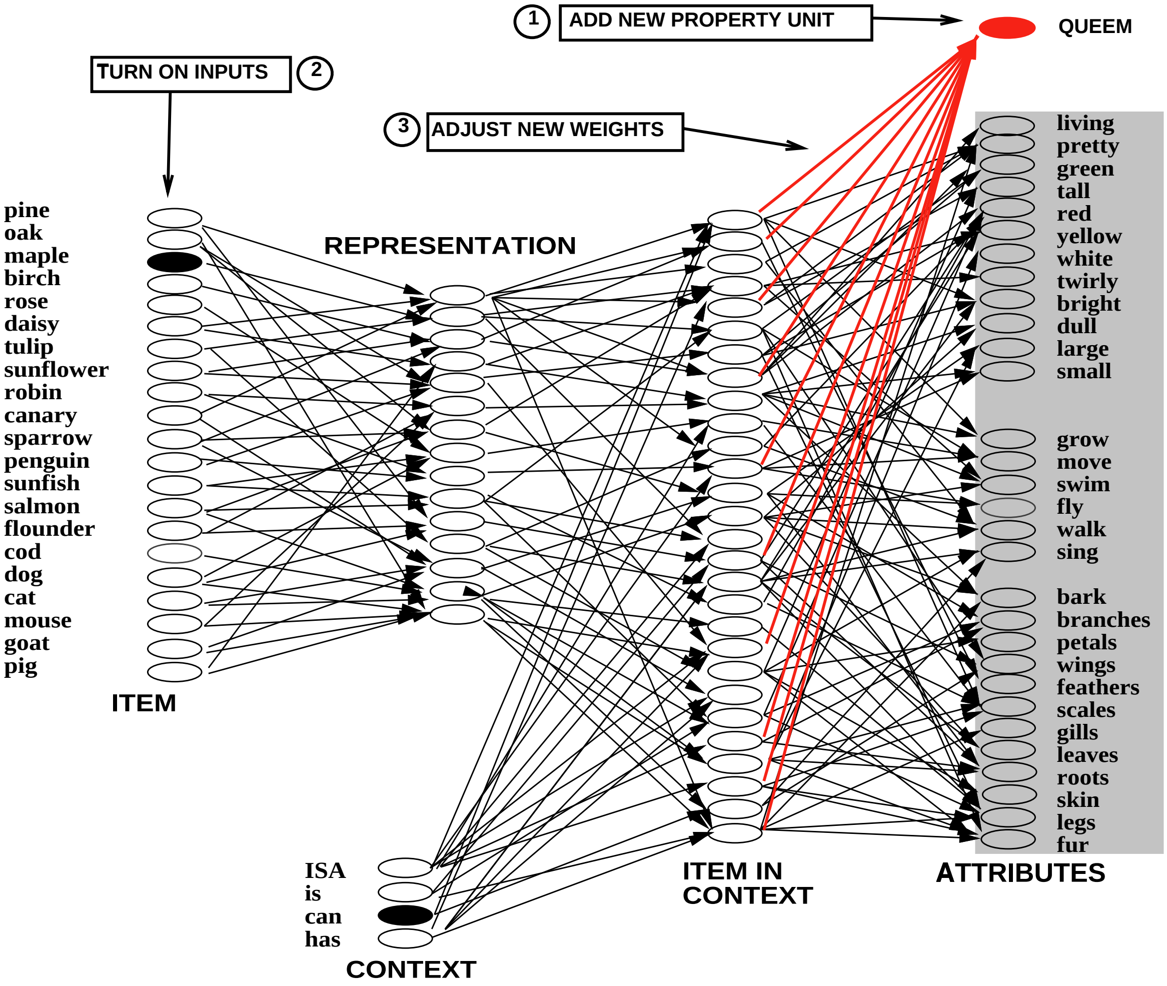


Figure 9. To teach the model a new fact about a familiar object, we simply add a new output unit to represent the novel property, and train the model to activate it for the appropriate object and context input. See text for details.

“rhinocerous”), a biological fact (*has warm blood* or *has cold blood*), and a physical fact (*is one ton* or *is one hundred tons*). We added a new output unit corresponding to each property, and trained the model to activate the correct units when probed with the brontosaurus or rhinocerous representations in the appropriate contexts. Finally, we “showed” the network a third novel animal (the triceratops), which had the same visible properties as the rhinocerous (*small, dull, skin, legs*); but we also taught the network that it was called a “dinosaur” like the brontosaurus. We allowed the network to derive an appropriate representation for the triceratops, and then probed it with the *has* and *is* contexts to see how it would generalize the newly-learned biological and physical properties.

The results are shown in Figure 12. The network inferred that the triceratops is one ton like the rhinocerous, on the basis of is physical similarity; but that it has cold blood like the brontosaurus, on the basis of having the same name.

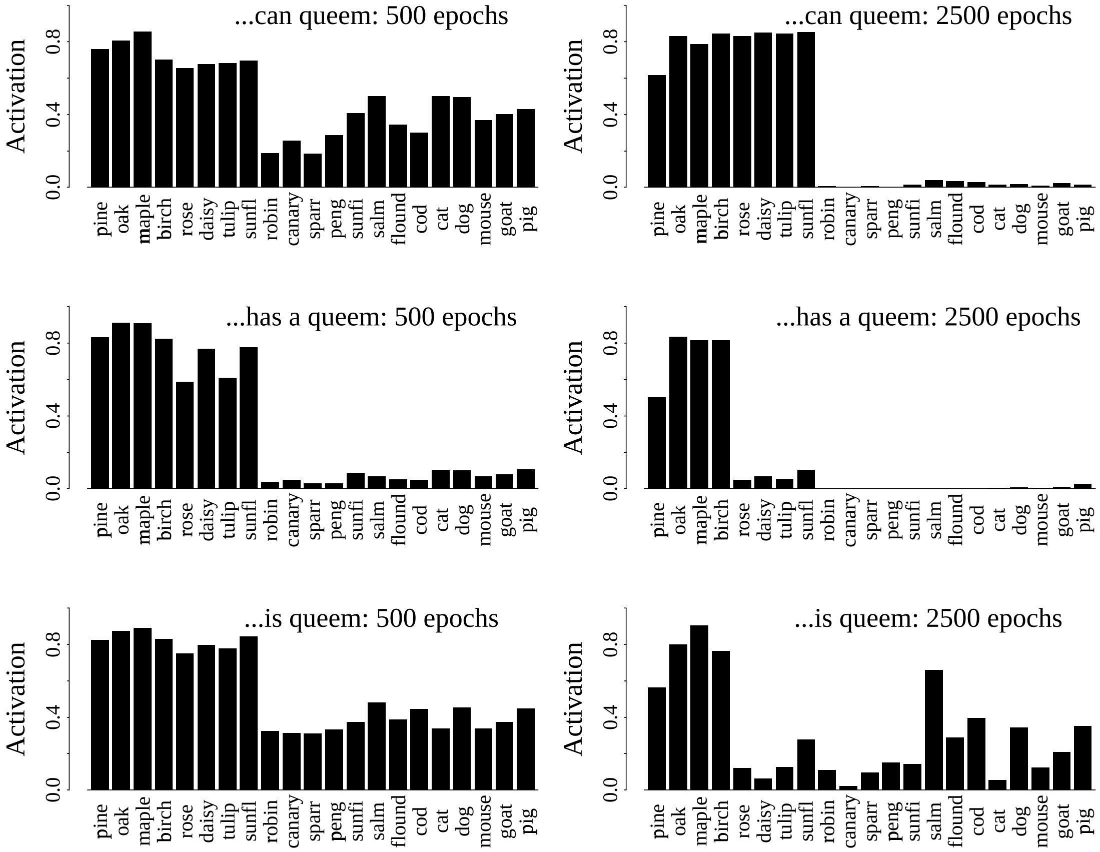


Figure 10. The figure shows the activation of the novel output unit *queem* in response to various items, after the model has been trained to activate the unit in response to the input *maple* paired with each of the *is*, *has*, and *can* contexts. Early learning, the property generalizes across the entire superordinate category, regardless of the context in which it was learned. Later, it generalizes across different groups of objects, depending on the context.

Expertise

Finally, recent studies of expertise have also undermined learning-based theories of conceptual knowledge representation. Medin, Lynch, and Coley (1997) have reported that judgments of similarity can vary among experts, depending upon the kind of expertise they have acquired. They performed sorting experiments with three kinds of tree experts: landscapers, biologists, and parks maintenance workers. Presumably the three groups were equally familiar with the same species of trees. Despite this, there were interesting differences in their sorting behavior. Biologists tended to stick fairly close to the scientific taxonomy, regardless of the surface characteristics of various trees, whereas lanfscapers and maintenance workers were more likel;y to deviate from the scientific taxonomy for uncharacteristic trees. Moreover, landscapers and maintenance workers consistently grouped together a category of “weed” trees, whose members shared no essential or biological characteristics, but were grouped together presumably because they demand similar treatment in

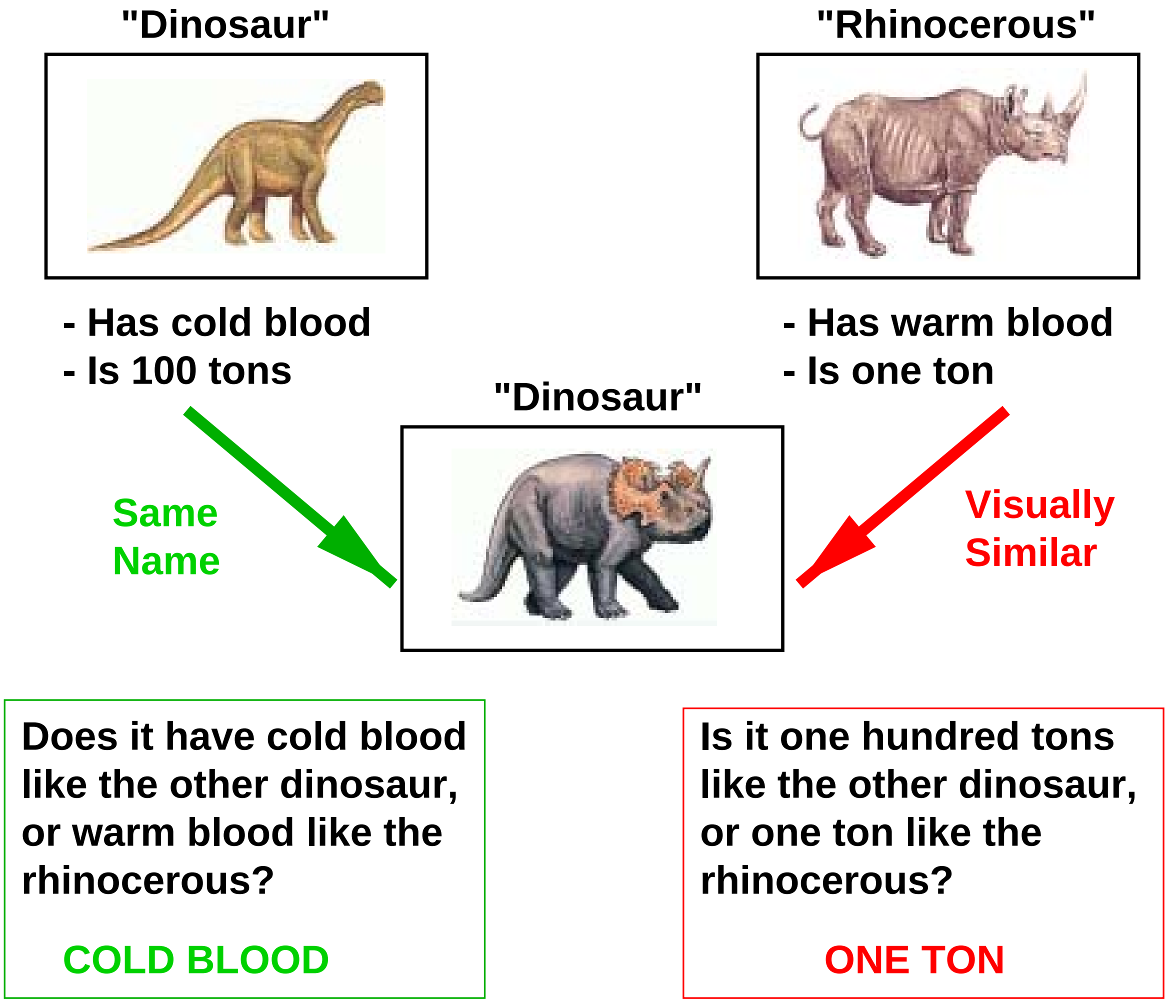

"Dinosaur"
"Rhinocerous"
- Has cold blood
- Is 100 tons
- Has warm blood
- Is one ton
"Dinosaur"
Same Name
Visually Similar
Does it have cold blood like the other dinosaur, or warm blood like the rhinocerous?
COLD BLOOD
Is it one hundred tons like the other dinosaur, or one ton like the rhinocerous?
ONE TON

Figure 11. Schematic of Gelman and Markman's (1986) experiment. Young children use category-label information to generalize biological properties (such as warm or cold blood), but use visual similarity to generalize physical properties (such as weight). See text for details.

the day-to-day tasks of the maintenaqnce workers and landscapers. These results suggest that the processes by which we construct semantic representations involve more than the simple accumulation of perceived properties. Instead, the manner in which we use and interact with objects may also inform the structures we acquire to guide our performance in semantic tasks.

The same is true in our simple network. To demonstrate this, we trained two models on the same set of patterns; however, we varied the frequency with which the items were presented in a particular context. To create a "scientific" network, concerned with the behaviors of objects, we trained the model most frequently in the *can* context. To create an "artistic" network, concerned with the appearance of objects, we trained a second model most frequently in the *is* context. Both models ultimately learned to complete all patterns in all contexts; however, the internal representations of objects they acquired in the context-free *Representation* layer differed substanitally, as shown by the hierarchical cluster analysis in Figure 13.

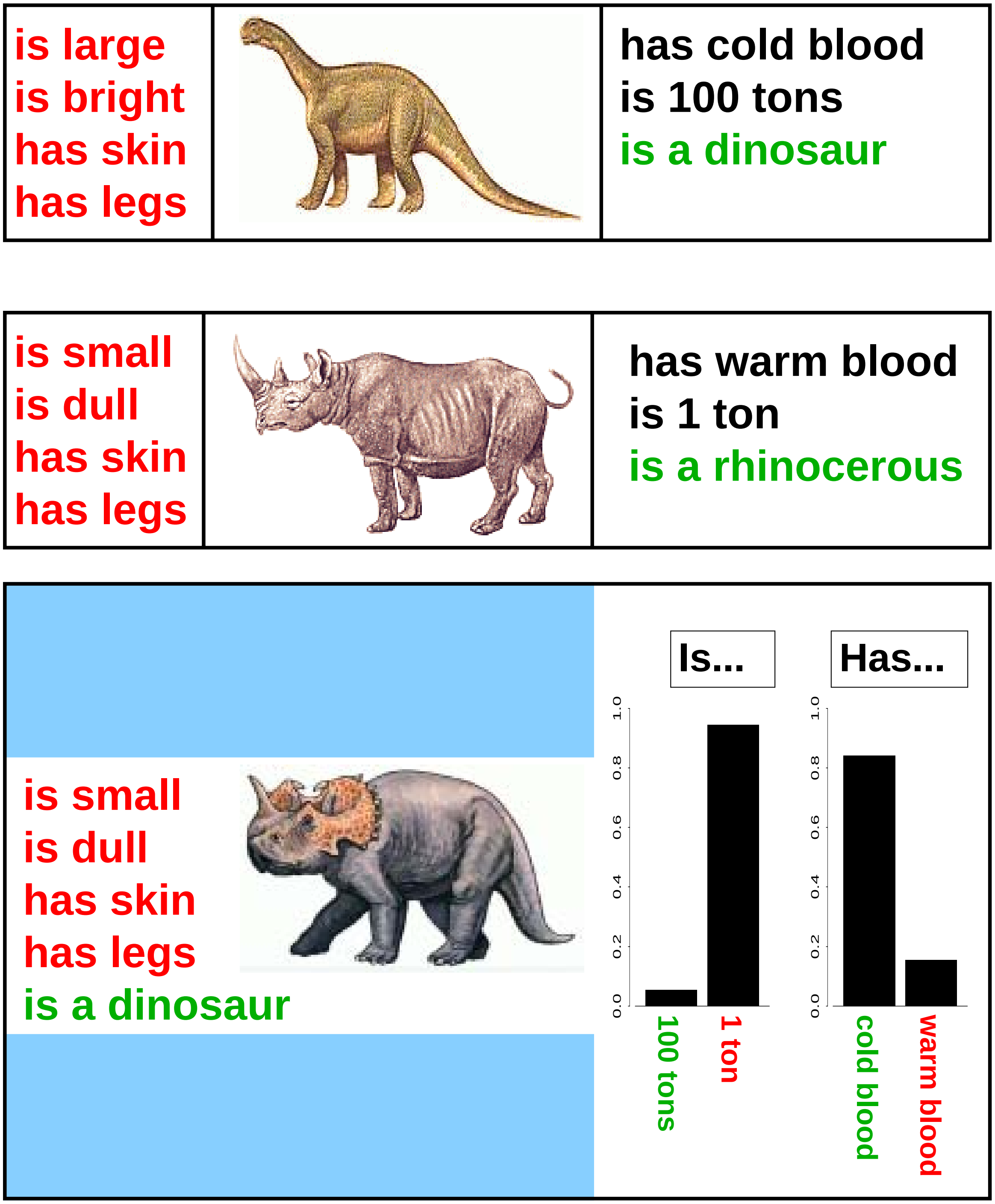
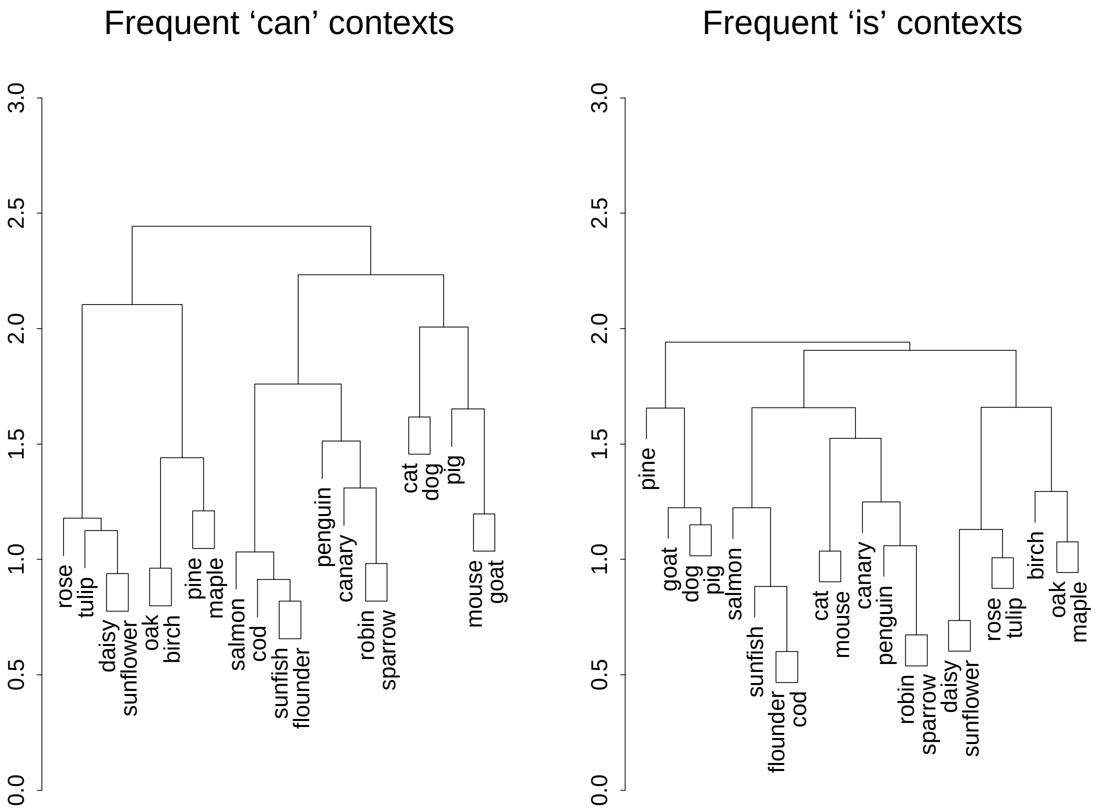


Figure 12. To simulate the experiment with the model, we use a combination of the techniques described previously. Like the children in Gelman and Markman's (1986) experiment, the network uses similarity in appearance (*is properties*) to generalize novel physical properties; and category label to generalize novel "biological" (i.e. *has*) properties. See text for details.



Frequent ‘can’ contexts

Frequent ‘is’ contexts

Figure 13. Hierarchical cluster plots of the similarities among context-independent representations in the model for the same 21 instances, when the model is trained with either *can* (left) or *has* (right) contexts most frequent.

Conclusions

Recent theoretical consensus has it that conceptual knowledge acquisition cannot be explained without reference to implicit theoretical knowledge, which constrains the manner in which the structure of the environment informs concept representations. However, we have shown that the principles embodied in the PDP framework (as illustrated by a simple feed-forward model) provide a natural explanation for many of the empirical phenomena upon which this consensus rests. To what extent, then, are theories necessary to constrain concepts? The answer to this question, of course, depends upon the meaning of the word “theory.” In the PDP framework, knowledge for all objects and properties is coded in the same set of weights; hence, such general knowledge always colors the processing of a particular item in a given context. The state of the weight matrix at any point in development constrains the ease with which new properties may be learned, and the manner in which they will generalize. Whether or not such influences count as implicit theories, then, is a question of semantics.

References

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

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. Hinton & J. A. Anderson (Eds.), *Parallel models of associative memory* (p. 161-187). Hillsdale, NJ: Erlbaum.

Keil, F. C. (1991). The emergence of theoretical beliefs as constraints on concepts. In S. Carey & R. Gelman (Eds.), *The epigenesis of mind: Essays on biology and cognition.* Hillsdale, NJ: Erlbaum.

Macario, J. F. (1991). Young children's use of color in classification: Foods and canonically colored objects. *Cognitive Development*, *6*, 17-46.

Massey, C. M., & Gelman, R. (1988). Preschooler's ability to decide whether a photographed unfamiliar object can move by itself. *Developmental Psychology*, *24*(3), 307-317.

Medin, D. L., Lynch, E. B., & Coley, J. D. (1997). Categorization and reasoning among tree experts: Do all roads lead to rome? *Cognitive Psychology*, *32*, 49-96.

Rumelhart, D. E., Smolensky, P., McClelland, J. L., & Hinton, G. E. (1986). Schemata and sequential thought processes in PDP models. In J. L. McClelland, D. E. Rumelhart, & the PDP Research Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 2: Psychological and biological models* (p. 7-57). Cambridge, MA: MIT Press.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and performance XIV: Synergies in experimental psychology, artifical intelligence, and cognitive neuroscience* (p. 3-30). Cambridge, MA: MIT Press.