Chapter 5: Semantics Without Categorization

Timothy T. Rogers
University of Wisconsin-Madison

James L. McClelland
Stanford University

Correspondence may be directed to:

Timothy T. Rogers
Department of Psychology
1202 W Johnson Street
UW-Madison
Madison, WI, 53706

James L. McClelland
Department of Psychology and
Center for Mind, Brain and Computation
Stanford University
Stanford, CA, 94305

Human beings have a remarkable ability to attribute meaning to the objects and events around them. Without much conscious effort, we are able to recognize the items in our environment as familiar "kinds" of things, and to attribute to them properties that have not been observed directly. We know, for instance, that the banana on the kitchen counter has a skin that easily peels off, and that beneath the peel we will find a soft yellow-white interior. We know that the banana is meant to be eaten, and can anticipate what it will taste like. Such inferences spring readily to mind whether we observe the banana itself or, as with this paragraph, simply read or hear statements referring to bananas. The cognitive faculty that supports these abilities is sometimes referred to as "semantic memory," and a key goal of much research in cognitive psychology is to understand the processes that support this aspect of human cognition.

One long-standing hypothesis places categorization at the heart of human semantic abilities. The motivation for this view is that categorization can provide an efficient mechanism for storing and generalizing knowledge about the world. As Rosch (1978) put it, "…what one wishes to gain from one's categories is a great deal of information about the environment while conserving finite resources as much as possible." Thus categorization-based theories propose that knowledge about the world is stored in a set of discrete category representations, each encoding or providing access to information about the properties that characterize members of the class. New items are assigned to stored categories through a process that is sensitive to the similarity between the item and the stored representations; and once the item has been categorized, it is attributed the properties known to typify the category. There are a great many different hypotheses about / models of the processes by which items are assigned to categories and

subsequently are attributed properties (Anderson, 1991; Ashby & Alfonso-Reese, 1995; Kruschke, 1992; Love, Medin, & Gureckis, 2004; Nosofsky, 1984; Pothos & Chater, 2002), but these share commitment to the idea that categorization is the engine that drives storage and generalization of knowledge about the world. Indeed, the idea that semantic abilities are supported by categorization processes is so pervasive that it is seldom treated as a hypothesis. In the preceding quotation, for instance, Rosch inquires only what one wants from one's categories—as though the question of whether our semantic memory system actually employs category representations is itself beyond question.

Still, the idea that categorization is the core mechanism supporting semantic abilities brings with it a series of challenges and puzzles that have yet to be solved. We briefly summarize some of the challenges that have motivated our work; a more extensive discussion of these issues is presented in the first chapter of *Semantic Cognition: A Parallel Distributed Processing Approach* (Rogers & McClelland, 2004), henceforth *SC.*

Multiple category representation. As has long been known, objects in the world usually belong simultaneously to many different categories (Barsalou, 1993; Collins & Quillian, 1969; Murphy & Lassaline, 1997; Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976). Lassie, for instance, belongs to the categories *collie, dog, pet, animal* and *movie star*. Chickens belong to the categories *bird*, *animal*, *poultry* and *livestock*; bulldozers are both *vehicles* and *construction equipment*; and so on. Moreover, the different categories to which an item belongs can license different conclusions about the item's unobserved properties. If the chicken is categorized as an *animal*, this might license the conclusion that it can't fly, since most animals are flightless. If classified as a

bird, the reverse conclusion is warranted (most birds can fly); and if classified as a *chicken*, the original conclusion holds (chickens can't fly). The notion that semantic knowledge resides in or is accessible from a set of stored category representations thus raises the question of how the different competing categories are "selected" as the appropriate ones for governing generalization in a given context. This in turn seems to require involvement of some additional representational structure or processing mechanism that governs the selection of / interaction among category representations. For instance, the "spreading activation" theories of the 1970's proposed that different category representations were connected in a graph structure that facilitated the "flow of activation" between categories that are "linked" in memory (Collins & Loftus, 1975; Collins & Quillian, 1969).

Category coherence. The problem of multiple category representation highlights a second more general challenge for prototype-like theories where semantic knowledge is thought to be stored in summary representations of categories. Specifically, how does the semantic system "know" for which groupings of items it should form a stored category representation (Murphy & Medin, 1985)? Some groupings of items seem to form coherent sets that are useful for governing generalization (e.g. "birds," "dogs," "cars"), whereas other groupings seem less coherent and useful (e.g. "things that are either blue or orange", "things that have corners", "things that could fit in the trunk of my car"). How does the system "know" that it should store a summary representation of the class of dogs, but should not store a summary representation of the class of things that are blue and orange?

Primacy of different category structures in development, maturity, and dissolution. The general question of how the semantic system "knows" which categories to store is further constrained by empirical evidence regarding the acquisition of conceptual distinctions over infancy and early childhood, the primacy of certain kinds of category structures in healthy adult cognition, and the dissolution of conceptual knowledge in some forms of dementia. Together, these sources of evidence generally support two seemingly contradictory conclusions about which kinds of categories are easiest to learn and most robustly represented in memory.

First, a large body of research generally suggests that infants and children differentiate quite gross conceptual distinctions (such as the distinction between animals and manmade objects) earlier in life than more fine-grained distinctions (such as the distinction between birds and fish; Keil, 1979; Mandler, 2000; Mandler & McDonough, 1996; Pauen, 2002, 2002). A complementary body of work in neuropsychology has shown that the progressive dissolution of conceptual knowledge observed the degenerative syndrome semantic dementia (SD) follows a reverse path: patients with SD first lose the ability to differentiate quite fine-grained conceptual distinctions (e.g. they cannot tell a robin from a canary, but know both are birds), then gradually lose the distinctions among increasingly more general categories as the disease progresses (Patterson & Hodges, 2000; Patterson, Nestor, & Rogers, 2007; Rogers et al., 2004; Rogers & Patterson, 2007). Together, these literatures suggest that more general or global conceptual distinctions are both the first to be acquired and the most robust in the face of global semantic impairment.

Seemingly in contrast to this conclusion, however, are long-standing observations about children's lexical development and the categorization behaviour of healthy adults, both of which seem to suggest that categories at an intermediate level of specificity are primal in both acquisition and adult performance. For instance, a long tradition of research convincingly demonstrates that children learn to name objects at an intermediate or *basic* level of specificity (e.g. "bird," "fish"), prior to learning more general (e.g. "animal," "vehicle") or specific (e.g. "robin," "trout") names (Brown, 1958; Mervis & Crisafi, 1982; Mervis, 1987); and seminal work by Rosch and others (Jolicoeur, Gluck, & Kosslyn, 1984; Mervis & Rosch, 1981; Murphy & Smith, 1982; Rosch et al., 1976) demonstrated that adults (i) usually produce basic-level names in free-naming tasks and (ii) are faster and more accurate to categorize typical objects at the basic level relative to more general or more specific levels. This body of research seems to suggest that basic-level categories are acquired earlier and are more robustly represented in the semantic system than are more general or more specific representations.

Thus to the general question "How does the semantic system know which categories should be stored in memory," categorization-based approaches face several further questions pertaining to acquisition and dissolution. Why are more general category distinctions acquired earlier by preverbal infants? Why are intermediate-level names learned earliest in development, and why are they so robust in healthy adults? Why are the more general distinctions also more robust to semantic impairment?

Domain-specific patterns of inductive projection. Finally, classic research in the study of children's inductive inferences has found that different kinds of properties tend to generalize from a reference item in quite different ways, at least in older children and

adults (Carey, 1985; Gelman & Williams, 1998; Gelman & Coley, 1990; Gelman & Wellman, 1991; Jones, Smith, & Landau, 1991; Macario, 1991; Massey & Gelman, 1988). For instance, when told that a robin "has an omentum inside," older children and adults generalize the property strongly to the class of birds, but if told that a robin "likes to live in an omentum," they generalize the property much more narrowly—for instance, only to small birds that build nests in trees. One puzzle for categorization-based theories is how to explain these different patterns of inductive projection for different kinds of properties. If generalization is governed by an item's similarity to stored category representations, how can new learning about a given item generalize in quite different ways, depending upon the "kind" of information? A further challenge is that patterns of inductive projection for a given property can also vary substantially depending upon the item in question. Thus, for instance, children weight shared shape more heavily than shared colour when generalizing new knowledge in the domain of toys, but show the reverse pattern in the domain of foods(Macario, 1991). Such patterns pose a chicken-and-egg problem for categorization-based theories: a new item cannot be categorized until one knows how to weight its properties, but because these weights depend upon the category, they cannot be determined until the item has been categorized(Gelman & Williams, 1998).

One response to these challenges, of course, is the further development of categorization-based approaches to semantic knowledge. The current volume attests that there are many interesting new developments in this vein. Our own work, however, is motivated by the observation that categorization is not the only efficient mechanism for storing and generalizing knowledge about the world. A tradition of work extending back

to the distributed memory model of McClelland and Rumelhart (1986) and the semantic memory models of Rumelhart (Rumelhart, 1990; Rumelhart & Todd, 1993) and Hinton (1981; 1986) provides an alternative mechanism that, like categorization, is efficient in that it does not require addition of a new representational element for every new class or learning event, and that also promotes semantic generalization—all without requiring any internal process of categorization. The internal representations employed in our approach do not correspond to explicit, discrete representations of different classes of objects, and the processes that govern generalization of knowledge to new items do not involve the assignment of the item to one of a finite number of discrete equivalence classes. Consequently, questions about which categories are stored in memory, or how the system determines which category to use when generalizing, are moot in this framework. Furthermore, our framework suggests potential mechanisms that account for the variety of phenomena summarized above, and also provides some clues as to how the semantic system may be organized in the brain. Thus we believe that our approach—semantics without categorization—can resolve many of the puzzles faced by categorization-based approaches to semantic knowledge.

Assumptions of the approach.

Our approach adopts the basic assumptions of the connectionist or parallel distributed processing (PDP) approach to cognition generally (Rumelhart, McClelland, & Group, 1986):

(1) Cognitive phenomena arise from the propagation of activation amongst simple, neuron-like processing units, each of which adopts, at any point in time, an activation state analogous to the mean firing rate of a population of neurons. "Input"

units explicitly encode the state of the environment via direct “sensory” inputs; “output” units explicitly encode representations of potential responses; and “hidden” units mediate the flow of activation between inputs and outputs.

(2) Propagation of activation is constrained by weighted synapse-like connections between units. At any given time, the activation of a unit depends on (a) the states of the other units from which it received incoming connections, (b) the values of the weights on the incoming connections, and (c) a transfer function that is typically nonlinear and differentiable.

(3) Input and output representations are directly assigned by the theorist and are intended to capture aspects of structure existing in the environment as encoded by sensory and motor systems. Internal representations of the environment take the form of distributed patterns of activation across subsets of hidden units, and information processing involves the updating of unit activations over time in response to some direct input from the environment, which has the effect of transforming input representations into internal representations that are useful for generating appropriate output patterns.

(4) The internal representations generated by inputs depend upon the architecture, that is, the gross pattern of connectivity among units in the network, as well as the particular values of the weights on these connections. Learning—defined as a change in processing prompted by previous experience—arises from experience-dependent changes to the values of the weights on these connections.

In addition to these general principles of parallel distributed processing, our approach adopts further assumptions specific to the domain of semantic cognition:

1) The function of the semantic system is to generate context- and item-appropriate inferences about the properties of objects (either perceived directly or referred to in speech) that are not directly observed in the current situation.

2) The semantic system adopts a “convergent” architecture in which all different kinds of information, regardless of the semantic domain or the modality of input or output, are processed through the same set of units and weights. This architecture permits the model to exploit high-order patterns of covariation in the sets of visual, tactile, auditory, haptic, functional, and linguistic properties that characterize objects.

3) Changes to weights in the system are generated by a process of predictive error-driven learning. The key idea is that, upon encountering a given object in a particular situation, the semantic system generates implicit predictions about what will happen next, which are either confirmed or disconfirmed by subsequent experience. Learning involves adjustment of weights throughout the system so as to reduce the discrepancy between observed and expected outcomes over a broad set of experiences with different objects and situations.

4) Weights are initially very small and random, so that all different kinds of inputs generate very similar internal representations.

5) Acquisition of new semantic knowledge proceeds slowly and gradually. To promote generalization, our model assumes that semantic representations are

distributed and overlapping. To prevent catastrophic interference in such a system, learning must proceed in a slow and interleaved manner, so that new learning does not "over-write" useful weight configurations built up over past experience.

6) The semantic system interacts with a fast-learning episodic memory system (Mcclelland, Mcnaughton, & Oreilly, 1995). Of course, human beings are capable of rapidly acquiring new facts about objects and generalizing them appropriately. Our theory proposes that this ability is supported by a fast-learning system in the medial temporal lobes, consistent with a long history of research in the neural basis of episodic memory (Mcclelland et al., 1995). The fast-learning system employs sparse representations so that even very similar objects and events are represented with non-overlapping patterns. Consequently, the fast-learning system can learn rapidly without catastrophic interference, but cannot generalize well. We assume that the fast-learning MTL system interacts with the slow-learning semantic system to support the immediate generalization of newly-learned information. Our model implements the slow-learning semantic system only.

7) Internal representations are shaped by the situation or context. We assume that the internal representations that guide semantic cognition are shaped, not only by the current object of interest, but also by learned representations of the situation or task-context in which the item is experienced. So, for instance, the same object—say, a chicken—can evoke different internal representations, depending on whether the current situation demands retrieval of the item's name, function, expected behaviour, shape, and so on. In the models reported here, the *Relation* inputs provide information about the context that influences the *Hidden* unit activations directly; however in

other work, we have investigated models that must learn distributed internal representations that capture similarity structure across relation contexts (Rogers & McClelland, 2008).

These are the core assumptions adopted by our approach. In the next section we describe a simple feed-forward model that conforms to these assumptions, and illustrate how it offers leverage on the challenges faced by categorization-based approaches raised earlier.

A simple model implementation.

The model architecture is based on that described by Rumelhart (1990) and shown in Figure 5.1. It consists of five layers of units connected in a feed-forward manner as indicated in the illustration. Units in the *Item* input layer directly encode localist representations of individual items that may be encountered in the environment. In the original model employed by Rumelhart, these included 8 items taken from the classic hierarchical spreading-activation model of Collins and Quillian (1969). In the simulations we will discuss, we extended this simple corpus to include eight plants (four different flowers and four different trees) and 13 animals (4 birds, 4 fish, and 5 mammals). Units in the *Relation* input layer directly encode localist representations of different relational contexts in which the various items might be encountered. The relation context constrains which of the item's various properties are immediately relevant to the current situation. For instance, the "can" context indicates a situation in which the system must anticipate or report the item's expected behaviour; the "has" context indicates a situation in which the system must anticipate or report its component parts; the "is" situation indicates

contexts in which the item's visual appearance is relevant; and so on. Of course, in the real world there exist many more potential items and potential contexts. In this model, the items and the relation contexts both represent a simple elaboration of the subordinate concepts and relation terms employed by Collins and Quillian (1969) and by Rumelhart (1990).

--Figure 5.1 about here --

Whereas the activations of *Item* and *Relation* units are directly set by the environment, all other unit activations are determined by the inputs they receive from the units to which they are connected. The net input to a receiving unit is the inner product of the activations of the units from which it receives projections and the values of the weights on those projections, plus a fixed bias constant of -2 that serves to turn units off in the absence of input. Unit activations are set by passing the net input through a sigmoidal activation function:

$$a_i = 1/(1 + e^{-net_i})$$

…where a_i is the activation of unit i and net_i is the net input to unit i. This function is bounded between 0 and 1 and increases smoothly and monotonically, but nonlinearly, in this range. The sigmoid activation function is important for learning in multi-layer networks because it is nonlinear and differentiable, and so allows gradient-descent learning to form internal representations capable of supporting essentially any input-output mapping (Rumelhart, Durbin, Golden, & Chauvin, 1995; Rumelhart, Hinton, & Williams, 1986). It is extensively used in connectionist modelling partly because it has

a comparatively simple derivative for use in gradient-descent learning, and partly because it approximates the expected firing rate for a population of integrate-and-fire neurons given the same net input (Movellan & McClelland, 1993).

Units in the *Item* layer project forward to the layer labelled *Representation*. Activation of a single unit in the *Item* layer thus provokes a distributed pattern of activation across the *Representation* units. This pattern depends upon the values of the weights projecting from input to *Representation* layers, and these weights are shaped by learning and experience—so that the patterns produced in the trained model are distributed learned internal representations of the inputs.

These distributed representations in turn send connections forward to the *Hidden* layer, which also receives inputs from the *Relation* input units. The distributed patterns that arise here are therefore influenced by both the internal representation of the current item and by the current relation—so that the same item can give rise to quite different internal representations, depending upon the particular relation context in which it is encountered. Thus, whereas the *Representation* layer encodes a context-independent internal representation of the current item that is the same across all relation contexts, the *Hidden* layer encodes a context-dependent representation that varies across different contexts.

A final set of weights projects forward from the *Hidden* layer to the *Attribute* layer. Units in this layer correspond to explicit properties that can be attributed to objects, including their perceptual properties, their names and other verbal statements about them, their behaviours, or the motor responses one might generate when interacting with them. In general we view these as directly capturing properties that can in principle be directly

experienced from the environment, though all the properties may not be present in any given situation. For instance, there is an *Attribute* unit corresponding to the property "can move," which may be directly available from the environment whenever an item is observed to be moving, but which can also be inferred by the knowledgeable observer even if the item is currently stationary.

The model is "queried" by presenting inputs to the *Item* and *Relation* units, corresponding to the observation of a given item in a particular situation, and computing activations for all units in a forward pass, based on the values of the interconnecting weights and the sigmoidal activation function of the units. When the configuration of weights is such that the model activates all and only the appropriate responses for all the various possible queries—for instance, when it activates the units "grow," "move," "fly," and "sing" for the canary—the model can be said to "know" the domain.

To find such a configuration of weights, the model is trained with the backpropagation learning algorithm (Rumelhart, Hinton, & Williams, 1986). In backpropagation learning, the output generated by the network for a given input is compared to the desired output, and the difference is converted to a measure of error. The derivative of this error with respect to each weight in the network is then computed, and all weights are adjusted by a small amount in the direction that reduces the error for the given training item. In our simulations, we employed the sum-squared error: for each output unit, the squared difference between the target and the actual output is computed, and this is summed across output units to provide a total error signal to guide gradient-descent learning for each pattern. Though error propagation is thought by some to be a biologically implausible mechanism for learning, it is possible for error-like signals to be

carried in unit activation states, and hence to drive learning, in networks with bidirectional connectivity (Hinton & McClelland, 1988; O'Reilly, 1996). We take backpropagation to be a simple way of approximating this kind of learning in a feed-forward network, and as a simple and direct instantiation of the general assumption stated earlier that learning takes place via a process of predictive error-driven learning.

In keeping with the assumption that learning in the semantic system is slow and gradual, the weight changes for any given event are small and incremental. As a consequence, changes that improve prediction for a single item- pair, but hurt performance for other item-relation pairs, are soon reversed by subsequent learning, whereas changes that improve prediction for many different item-relation pairs accrete over time, allowing the model to discover a set of weights that "work" well for all of the items in the training environment simultaneously. This slow and gradual learning process thus provides a mechanism for developmental change in knowledge acquisition. The internal representations and outputs generated by any given input depend upon the configuration of weights at the time of testing. These weights—and consequently the internal representations and responses—evolve in interesting ways over the course of training with a fixed environment, providing a means of understanding patterns of behaviour in different age groups and across different tasks.

After training, the patterns that arise over the *Representation* layer capture similarity structure apparent in the output patterns describing each individual item. Items that have many properties in common across the different relation contexts are represented with similar patterns in this layer, whereas those with few properties in common are represented with quite different patterns. Because semantically related items

tend to have many properties in common, these internal patterns come to capture the semantic or conceptual similarity relations among items—so they can serve as a basis for semantic generalization.

The *Representation* units, because they receive inputs from *Item* but not *Relation* inputs, must find representational structure that works well across all different relation contexts. This structure will, however, not be useful for governing generalization in every individual relation context, because different "kinds" of properties can capture quite different similarity relations among the items. The representations that evolve on the *Hidden* units are constrained by input from the *Relation* units, and so, as elaborated below, can "reshape" the deeper similarity structure encoded in the *Representation* units. Feed-forward networks with different internal architectures—for instance, networks that connect inputs directly to outputs, or connect via a single hidden layer—are not obliged to simultaneously find both context-neutral and context-specific levels of representation, and so do not show many of the interesting behaviours of the Rumelhart network (see Chapter 9 of Rogers and McClelland,2004, for further discussion). The five-layer architecture of this model is therefore very important to its functioning.

Recommended implementation.

Simulations reported in Rogers and McClelland (2004) were conducted using the pdp++ software package. The latest version of this software, now called Emergent, can be found on the Web at http://grey.colorado.edu/emergent/index.php/Main_Page. Use of this software requires, however, somewhat specialized knowledge and training. For those without this expertise, we therefore recommend implementing the model in Matlab using PDPTool.

PDPTool is a fully-functional re-implementation, in Matlab, of the original software released with the PDP Handbook (McClelland & Rumelhart, 1988), coupled with an intuitive graphical user interface. It takes the form of a library of object-types, functions, and graphical display objects that can be loaded in the Matlab environment or can be run as a stand-alone application. Currently the library includes objects and functions that implement feed-forward, simple recurrent, and fully recurrent backpropagation, interactive activation and competition networks, and competitive learning. The library, including instructions for installing the software, a user's manual, a short tutorial, and an online version of the PDP Handbook, can be downloaded from the Web at http://www.stanford.edu/group/pdplab/resources.html A brief article describing the aims and utility of PDPTool was published in the Academic Edition of the Matlab Digest in October 2009.

The PDPTool release comes with a set of pre-built network, template, and environment files, some for use with the Handbook and tutorials, and others providing implementations of classic PDP networks from the literature. These example files include an implementation of the original 8-item Rumelhart network. A picture of the PDPTool display for this network is shown in Figure 5.2.

–Figure 5.2 about here –

We further note that, in order to replicate simulation results reported in Rogers and McClelland (2004), including those reviewed below, the user will need to build and parameterize networks as described in the source material. The original work develops a

series of increasingly elaborate model implementations. The architecture, training patterns, and model parameters for each implementation are, however, documented in the Appendices to the book and these should be sufficient for the reader to replicate the original work in PDPTool.

Addressing the core phenomena.

In this section we revisit the core phenomena motivating the approach and briefly indicate how the model addresses these.

Multiple category representation. In our framework, the internal representations that govern how knowledge generalizes are not discrete category representations, but are patterns of activation across continuous-valued units in the *Representation* and *Hidden* layers of the network. Each pattern can be viewed as a point in a continuous high-dimensional space, with each individual unit encoding one dimension. In this sense our model is similar to exemplar-based models, except that i) the dimensions of the space do not correspond to interpretable semantic features and ii) there is no "database" of stored exemplars. Instead each exemplar leaves its "trace" on the values of the interconnecting weights. Categories have no real existence in the model's internal representations, but correspond roughly to densely-occupied regions of the representation space. For instance, the four individual fish are all represented as somewhat similar to one another, and as quite different from the birds and mammals, and so form a fairly tight cluster that corresponds roughly to the category *fish*. More general categories correspond to more inclusive clusters in the space, and more specific categories to less inclusive clusters. Because there are no direct internal category representations in the model, there is no

problem of deciding which categories are “stored” by the semantic system, or of adjudicating which category structures should be used to govern generalization for a given task.

Nevertheless, the model can categorize in its overt behaviour. Each output unit can be viewed as a probabilistic categorical response to a given input. For instance, the output unit corresponding to the name “bird” can be viewed as indicating the likelihood that a given item belongs to the class of things that are labelled “bird.” Such judgments depend upon the configuration of the weights that project from the *Hidden* units to the *bird* name unit, which will strongly activate this unit for a subvolume of the representation space defined by the units in this layer. Whenever the model’s internal representation occupies a point in this subspace, it will generate the overt judgment that the represented item is called a “bird.” The precise location and extent of this volume depends upon the values of the weights projecting from *Hidden* to *Output* layers, which are acquired through the learning rule. Thus the model, like people, can generate explicit category judgments in its outputs, but it does not employ an internal categorization process to do so. It is also the case that explicit categorization labels may or may not align well with the organization of the model’s internal representations. Though labels like “salmon,” “fish,” and “animal” may align well with clusters at different levels of granularity, nothing prevents the model from learning other labels that cross-cut this structure. For instance, a name like “pet” might apply to the canary, dog, and cat, but not to the other birds or mammals in the corpus.

Category coherence. Because the model does not store overt category representations, there is no question about why some categories (like “dog”) are stored

and others (like “blue-and-orange things”) are not. Beyond this, however, the model suggests one reason why some groupings of items seem to provide good candidates for naming and for inductive generalization whereas others do not. Specifically, the model’s internal representations are strongly shaped by the covariance structure of the properties of objects in the environment. Sets of items that tend to share many properties in common with one another get represented as quite similar to one another. Consequently learning about the properties of one such item tends to generalize strongly to all other similarly-represented items. This in turn means that the properties common to most items within such a cluster get learned very rapidly, whereas the properties that individuate items within a cluster will be somewhat more difficult to learn. On this view, coherent categories—sets of items that provide good vehicles for induction and are likely to receive names in the language—are those groups of items sharing properties that themselves reliably covary with many other properties. We note that this conception of coherence does not necessarily just reflect raw overall similarity amongst items: Sets of items that share many properties will not be represented as similar if the properties they happen to share do not themselves covary strongly with many other properties in the set (see Chapter 3 of the Semantic Cognition book).

Primacy of superordinate structure in development and dementia. Although the fully-trained model finds internal representations that capture the similarity structure of the training patterns, the network does not discover this organization of internal representations all at once. Instead, the representations undergo a progressive nonlinear process of differentiation—first discriminating items from grossly different conceptual domains (e.g. plants and animals) without any apparent fine-grained structure; then

capturing intermediate distinctions (e.g. birds vs. fish) without further subordinate organization; and finally pulling apart individual items. Figure 5.3 shows a multidimensional scaling diagram of the model's internal representations taken at evenly-spaced points throughout training. Each line shows the trajectory of a single item's representation throughout learning. Initially, all items are represented as similar to one another, because the network is initialized with small, random values. Very quickly, however, the plants become differentiated from the animals, while within these coarse categories we see very little differentiation. Some time later, the network begins to differentiate more intermediate clusters (birds and fish, flowers and trees) but with little differentiation of the individual items. In the last phase, the individual items begin to pull apart from one another. A full explanation of the reasons for this phenomenon is beyond the scope of this paper but was provided in Chapter 2 of the Semantic Cognition book. This progressive coarse-to-fine discrimination of internal representations mirrors the pattern of conceptual development observed in carefully controlled studies of pre-verbal infants in the work of Mandler (2000), Pauen (2002; 2002) and others.

-- Figure 5.3 about here --

Figure 5.3 suggests why more general or superordinate-level information tends to be more robust in progressive semantic syndromes like semantic dementia. In its fully-trained state, the network has learned to map out from its internal representations (coded in *Representation* and *Hidden* layers) to explicit representations of overt responses. For properties that are only true of very specific concepts—for instance, the name "canary,"

or the property “can sing” (true only of the canary)—the network has learned to activate the corresponding unit only from the internal state corresponding to the represented item. Thus there is a relatively narrow subvolume of the representation state space from which the network will activate the units corresponding to very specific properties. For properties that characterize more general classes—properties like has wings (true of all birds) or can move (true of all animals)—the network has learned to activate the corresponding output units for a wider set of items, all of whom are represented as somewhat similar to one another. The consequence is that, for these items, there is a broader subvolume of the representation state space from which the model has learned to generate the corresponding response. When the system degrades with disease, the internal representations generated by a given input become distorted—some of the connections that encode the representation are no longer present. For very specific properties, small distortions can move the representation out of the relatively narrow subvolume from which the model can generate the appropriate response. For more general properties, small distortions to the correct pattern will not move the representation out of this subvolume, and the model will continue to generate the appropriate response. Thus, the more general the category to which the property applies, the more robust knowledge of the property will be to semantic impairment.

-- Figure 5.4 about here --

The left panel of Figure 5.4 shows the network’s ability to activate subordinate, basic, and superordinate-level names for items in its environment as the patterns of

activation in the *Representation* layer are subject to increasing amounts of noise (simulating the degradation of semantic representations arising from progressive brain damage). Whereas basic-level names are initially the most strongly active labels, activation of these units declines more sharply than the activation of the more superordinate name units, producing a "cross-over" effect where general-level responses are more active than basic-level responses when the network is subject to moderate to severe semantic impairment. The right side of Figure 5.4 shows analogous data from patients with semantic dementia performing a category-verification task (Rogers & Patterson, 2007).

Primacy of the basic level in lexical acquisition and adult categorization and naming. If more general concepts are the first to be differentiated in acquisition and are the most robust to semantic impairment, why do categories at the more intermediate or "basic" level appear to be "privileged" in lexical acquisition and in adult categorization? Our approach suggests an answer to this seeming paradox, which stems from the observation that the coarse-to-fine differentiation of concepts is observed fairly early in development, before children have begun to speak. By the time children have begun to name objects, they are also able to differentiate concepts at the basic level (Mandler, 2000; Mandler & Bauer, 1988; Pauen, 2002). Once the semantic system has begun to differentiate basic clusters within some more general domain, it is actually at a disadvantage in learning superordinate relative to basic-level names.

To see this, consider teaching the Rumelhart model to name at a point in its development such that the birds have been differentiated from the fish, but the individual

birds and fish are still quite similar to one another. For each item, there are 3 different names that might be appropriate—for instance, for the canary, the network could name it as an “animal,” a “bird,” or a “canary.” Suppose further that all of the different names occurred equally frequently in the environment—say, 4 times per epoch of training. Which names would the network learn first?

First consider the name “animal.” When the model learns that the canary is an animal, this response will tend to generalize strongly to the robin (which is represented as quite similar) but less strongly to the salmon and the sunfish (which are now somewhat distinct). Similarly when the network learns that the salmon is an animal, this response will generalize strongly to the sunfish but less strongly to the canary. So each time the name appears for one of the 4 animals, only half of the items that share the name benefit strongly from the learning.

When learning to call the canary a “bird,” the same pattern is observed: the response generalizes strongly to the robin and not to either of the fish. This time, however, the name bird is only *ever* applied to one of the 2 individual birds. If it occurs 4 times per epoch, then it occurs twice with the canary and twice with the robin. So *every* time the name appears it benefits *all* of the items to which it applies. On these grounds, this intermediate-level name should be learned more rapidly than the more general name.

What about specific names? In our scenario, the word “canary” appears 4 times per epoch, always with the *canary* item. As before, this response will have a strong tendency to generalize to the *robin*. In this case, however, the generalization is detrimental—the name “canary” does not apply to the robin. So, when the model encounters the robin, it must reverse the weight changes, to turn off the name unit

corresponding to “canary” and turn on the unit corresponding to “robin.” This similarity-based interference will prevent the network from rapidly learning this more specific name. Only when the robin and canary are sufficiently differentiated from one another will the system easily learn to generate different subordinate names for them.

In other words, when word-frequency is controlled, then at any given point in development, the network will best be able to learn those words that demarcate items within a relatively tight cluster. More general names will be learned more slowly because they apply to items that are dispersed in the space and so do not promote strong cross-item generalization; more specific names will be learned more slowly because they apply to items with similar representations and so suffer from strong cross-item interference. Since children learn to name only after they have differentiated intermediate concepts, they are most likely to learn intermediate-level names. And because basic-level clusters continue to be “tight” and well-separated into adulthood, adults are likely to show similar advantages in basic-level categorization and naming.

Basic-level advantages arise in our model for largely the same reasons proposed by Rosch and others (Murphy & Brownell, 1985; Rosch et al., 1976; Tanaka & Taylor, 1991) . A key difference is that, on the PDP theory, the representational similarity structure that promotes basic-level advantages is not present initially, but only emerges after coarser conceptual distinctions have been acquired. Thus the primacy of the basic level in lexical acquisition and adult categorization co-exists with the coarse-to-fine differentiation of concepts in pre-verbal infants and the preservation of general level information in semantic impairment. For instance, in the simulation results shown in Figure 5.4a, the undamaged model activates basic-level names more strongly than either

superordinate or subordinate names—but nevertheless, superordinate-level information is more robust when the network is damaged and, though not shown here, the internal representations still differentiate in a coarse-to-fine manner in the same simulation (Rogers & McClelland, 2004).

Domain-specific patterns of inductive projection. Finally, the basic architecture of the network in Figure 5.1 suggests one answer to the puzzle of domain-specific inductive projection. Recall that, in classic studies of inductive projection—where children are taught a new property of a familiar item, and are then asked what other objects likely share the property—the pattern of inductive projection can differ depending upon the kind of property in question. For instance, if the property is understood to be a biological trait, it may generalize to one subset of items, but if understood to be a physical trait, it may generalize to a very different subset (Carey, 1985; Gelman & Markman, 1986). If categorization is the process that supports inductive projection, then how can new learning about a given item show very different patterns of generalization?

In our framework, the internal representations that constrain generalization of new learning are shaped, not only by the item in question, but also by information about the current task context. Specifically, the patterns of activation arising across *Hidden* units in the model depend partly upon inputs from the *Representation* units, but also on inputs from the *Context* units. Weights projecting out from the *Relation* units are, like the weights projecting out from the *Item* units, shaped by learning—so the network must learn how to treat items in different relation contexts, just as it learns how to represent the items themselves.

A natural consequence of this architecture is that, when the properties associated with two different relations capture quite different aspects of similarity among a set of items, then the similarity structure arising across *Hidden* units can be quite different, depending on the relation context. To see this, consider Figure 5, which shows a multidimensional scaling of the model's internal representations of 16 different items in different contexts. The middle plot shows the similarity of the 16 items as encoded in the *Representation* layer of a trained model. The top panel shows the same 16 items as encoded by the *Hidden* layer when the *is* context is activated in the input, whereas the bottom panel shows *Hidden* layer patterns for these items when the *can* context is activated in the input. In the network's environment, all of the plants can do only one thing: grow. In the *can* context, then, the model has learned to represent all 8 plants as nearly identical to one another. The birds and the fish, because they can do different things, remain well differentiated from one another. The *is* context, in contrast, all of the various items have quite different and somewhat random properties (mostly colours and idiosyncratic visual traits; note that the *is* relation is separate from the class inclusion relation denoted by *isa* in both our model and Collins and Quillian's original work). Consequently the 16 individual items are all fairly well differentiated from one another, although the deeper similarity structure present in the base representation can still be observed (for instance, the fish are all more similar to one another than they are to the birds).

-- Figure 5.5 about here --

It seems that the inputs from the *Relation* layer can "re-shape" the base similarity structure encoded by *Representation* units, to better capture structure suited to the context at hand. Because this is so, the model will show quite different patterns of inductive projection for the exact same reference item, depending upon the relation context in which the property is learned. For instance, Figure 5.6 shows what happens when the model is taught a new fact about the maple tree: that it *can queem*, *has a queem*, or *is queem*. Depending upon the relation, the model generalizes this new fact in quite different ways: to all of the plants if "queem" is a kind of behaviour (the *can* context), to just the trees if "queem" is a part (the *has* context), and to an idiosyncratic set of items if "queem" is a superficial appearance property (the *is* context).

-- Figure 5.6 about here --

In summary, an important aspect of our theory is the idea that internal semantic representations capture knowledge, not just about the item in question, but also about context in which the task is being performed. This context can capture information about the particular kind of information that the system is being asked to retrieve—consequently the kinds of generalization behaviours exhibited by the system can vary depending upon this information.

Relation to other approaches.

Our theory addresses a series of empirical and theoretical issues that have proven challenging for some categorization-based approaches. We are not aware of other

computational approaches which have tackled precisely the same set of motivating phenomena as has our work. In particular, our focus on core issues in child development and in neuropsychology makes it difficult to compare our approach to others in this volume, since these approaches are, by and large, targeted at explaining detailed observations about adult categorization behaviour, mostly with reference to new category-learning experiments. With regard to the nature of semantic representations and the mechanisms that support semantic generalization, however, our approach does share some commonalities and some differences with other computational approaches to categorization.

Exemplar theories / non-parametric density estimation. Our framework shares some characteristics with exemplar-based approaches to categorization (Kruschke, 1992; Medin & Shaffer, 1978; Nosofsky, 1986), some of which may be viewed as similar in many ways to nonparametric density estimation (Griffiths, Sanborn, Canini, & Navarro, 2008). In these approaches, as in our work, categorization behaviour is viewed, not as reflecting the internal mechanism that governs knowledge generalization, but as a probabilistic response generated from a continuous internal representation space. The approaches differ in their conception of the nature of the underlying representations and the processes that govern generalization.

In exemplar theories, the elements are a vast set of stored discrete representations of previous learning events, each typically construed as a point in a high-dimensional feature space. Learning involves adding new representations to this set with each learning episode. From these representations, a continuous probability density function is estimated for the full space, and generalization is then governed by maximum likelihood

estimates given the observed features of a new item and the estimated probability density. In our theory, the dimensions of the representation space do not correspond to interpretable semantic features, nothing is added with new learning, and there is no store of discrete events in memory. Instead what is "stored" is a single matrix of connection weights, and each learning event leaves a "trace" through its influence on the values of these weights.

Exemplar theories have mainly focused on learning to assign items to a single set of mutually exclusive categories, and it is not clear to what extent the best known theories (such as Nosofsky's Generalized Context Model, Chapter 2, or Kruschke's ALCOVE model, Chapter 6) fare when required to learn assignment of items to multiple different categorization schemes (Palmeri, 1999; Verheyen, Ameel, Rogers, & Storms, 2008). Our approach assumes that there is no single categorization scheme that is always employed for all items, but that the semantic system is capable of categorizing items according to a variety of different schemes.

Finally, some exemplar theories have focused on understanding how people weight the different properties of objects when categorizing them (Kruschke, 1992; Nosofsky, 1986). The PDP theory also suggests a mechanism for feature-weighting—specifically, it suggests that sets of properties that covary coherently with many other properties will receive greater weight in determining an item's internal representation, and consequently will strongly shape the similarity structure that governs generalization in the system. Understanding the similarities and differences between this approach to feature weighting and that offered by other computational approaches remains a goal for future research.

Prototype theories / parametric density estimation. In some respects, our approach may seem to be more similar to parametric approaches to density estimation (Ashby & Alfonso-Reese, 1995). In these approaches, the probability density in some feature space is computed, not by retaining a full record of all previous events, but by fitting some set of *n* parameterizable distributions to the observed data. For instance, such an approach might assume that each cluster in a multidimensional feature space has been generated from some Gaussian distribution; that the probability distribution for the full space can be approximated by a mixture of such Gaussians; and that "categories" correspond either to individual Gaussians or to some set of Gaussians in the mixture. On this view, learning serves to help determine how many "clusters" (Gaussians) there are, and what the parameters of the distributions are (i.e. the location of their modes in the feature space, the height of the mode and the variance of the distribution). Prototype theories (Chapter 3) can be viewed as a special case of parametric density estimation in which there exists one distribution ("prototype") for each known category.

Our approach is similar to parametric density estimation in that knowledge is stored in a fixed set of parameterizable elements—namely the weights—so that learning does not "add" new elements into a memory store. We believe this analogy to be somewhat misleading, however, because in most parametric approaches to density estimation, the basic representational elements are the distributions that need to be parameterized. That is, there is typically a one-to-one or a many-to-one correspondence between the individual distributions and the categories stored in memory. This is not the case in the PDP model—there is no sense in which either the weights or units in our model correspond directly to some explicit category. The internal representations

generated for various inputs always depend on the full set of weights, and each weight contributes to the representation of all categories and items.

Structured probabilistic models. Under structured probabilistic approaches(Kemp & Tenenbaum, 2008; Tenenbaum & Griffiths, 2001; Tenenbaum, Griffiths, & Kemp, 2006), semantic cognition is viewed as the inductive problem of deciding which of a vast number of conceptual structures is most likely to have generated the observed properties of a set of items in a domain. The relationship between such approaches and the PDP approach is the subject of much current debate beyond the scope of this chapter; the interested reader can find recent commentary in (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, submitted; Mcclelland et al., submitted; Rogers & McClelland, 2008). Here we simply wish to note some key commonalities and differences between the approaches.

Briefly, both views emphasize that the function of the semantic system is to promote inductive inference; both propose that inductive inference is probabilistic; and both situate the problem of learning within an optimization framework. The key differences lie in basic assumptions about representation and mechanism. Structured probabilistic approaches assume that the semantic system approximates optimal probabilistic inference—so that any method and any representational assumption that allows the theorist to compute the true probability distribution over some set of outcomes may be employed to understand how the system is working. In practice the representations often involve discrete category representations embedded in graph structures, with probability distributions computed over all possible categories and all possible graphs. The approach also requires specification of initial biases for all categories and graph structures, and so presupposes considerable initial knowledge.

Learning involves using observed data to select the most probable set of categories and graph structures governing some behaviour, and then updating prior beliefs about the likelihood of these different structures.

On the PDP approach, representations must take the form of distributed patterns of activity across processing units—hence there is no analogy to the graphical structures often employed in structured probabilistic approaches. Also, there are no discrete internal category representations, no explicit "hypotheses" about categories and structures, and no explicit biases attached to different hypotheses and category structures. The internal representations and outputs generated by objects and relations depend upon the particular configuration of weights instantiated in the network at a given point in time. This configuration itself can be viewed as a point in a fully continuous multidimensional weight space. The weight-space determines the full set of input-output mappings that can be expressed by the network, and in this sense is similar to the "hypothesis space" (the set of all possible concepts and graph structures) presupposed by the probabilistic framework. There are, however, several key differences. First, unlike the structured probabilistic approach, there is no enumerated set of possible structure types (or constructors for such structure types). Second, there is never a sense of comparative evaluation of alternative possible structure types—at any given point in time, the system has a single configuration of weights, capturing a single input-output mapping. Third, learning occurs quite differently in the connectionist framework: each new experience results in a slight adjustment of the connection weights, gradually leading, through an ongoing developmental process, to elaboration of a structured knowledge representation. While the structured probabilistic approach can be applied in such an "on-line" method, it

is typical to view each new experience as adding to the corpus of stored experiences from which the best possible structure will be inferred using arbitrary computational methods. Thus, the connectionist framework imposes strong constraints on the learning process not considered within structured probabilistic approaches.

Conclusion

In summary, we believe it is fruitful to investigate approaches to semantic cognition that do not invoke an internal categorization mechanism as the primary vehicle for knowledge storage and generalization. The PDP approach to semantic cognition provides an alternative set of mechanisms for efficient storage and generalization of semantic information. Although our framework shares some characteristics with other computational approaches, it also differs in several key respects. In particular, the applicability of the framework to phenomena in cognitive development and disordered semantic cognition allows the theory to address a comparatively broad range of empirical findings.

Figure Captions

Figure 5.1. A simple model implementation of the theory adapted from Rumelhart and Todd (1993), used to learn all the propositions true of the specific concepts (pine, oak, etc.) in the classic hierarchical spreading activation model of Collins and Quillian (1969). Input units are shown 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 item in the environment. Each unit in the *Relation* 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. All simulations discussed were conducted with variants of this model.

Figure 5.2. Graphical display of the model implemented as a part of the standard release of the PDPTool Matlab application. Input units appear on the left, and activation flows rightward toward the attribute unit on the right. The rightmost column of units displays the target values used to train the model. The Figure shows activation of unit states when the trained model is queried with the input *canary can*. The model correctly activates all correct responses in the output, including the attributes *grow*, *move*, *fly* and *sing*.

Figure 5.3. Progressive differentiation of concepts and preservation of superordinate information in the model. The labelled endpoints show a multidimensional scaling diagram of the model's internal representations of 8 items after it has correctly learned all properties of all items. The lines show a multidimensional scaling of the trajectory of these representations over the course of learning. Though representations begin very similar to one another, animals and plants quickly differentiate from one another, followed by more intermediate categories, and finally the individual items are pulled apart. This "coarse-to-fine" differentiation of concepts mirrors patterns of conceptual

differentiation observed in child development. The shading provides a conceptual illustration of the reason why superordinate information tends to be preserved: Properties shared by animals, for instance, are true of all birds and fish, and so tend to be activated by points within a broad region of the representation space (light gray shading in upper-left). Properties shared only by the birds, in contrast, are activated by points in a comparatively narrower volume of the space (dark shading around canary and robin representations), whereas properties true of individual items are only activated for a very narrow volume of space (the white "bubbles" around each individual item representation).

Figure 5.4. Left: Activation of correct name units for names at general (e.g. "animal"), basic (e.g. "fish"), and specific (e.g. "salmon") levels in a variant of the model trained with 21 items, when the model's internal representations are subject to increasing amounts of noise. Whereas the basic-level name is initially the most active unit, this activation declines more rapidly than the more general name unit, so that the network performs better when naming at more general levels. Right: Smoothed data from 8 patients with SD performing a category-verification task in which they must decide whether a picture matches a name at either the general, basic, or specific level. Like the model, performance is initially better for basic-level names, but declines with increasing semantic impairment, so that the more impaired patients show an advantage for categorizing at the most general level. Panels reprinted with permission from Rogers and McClelland (2004) Figure 5.4, page 196, and from Rogers and Patterson (2007), Figure 2, page 455.

Figure 5.5. Multidimensional scaling showing the similarities represented by the model for objects in different relation contexts. The middle plot shows the similarities among object representations in the *Representation* layer. The top graph shows the similarities among the same objects in the *Hidden* layer, when the *is* relation unit is activated. The bottom graph shows the similarities across these same units when the *can* relation unit is activated. The *is* relation context exaggerates differences among related objects; for example, relative to the similarities in the *Representation* layer, the trees are fairly well spread out in the *is* context. Moreover, similarities in object appearances are preserved in these representations; for example, the canary is as close to the flowers as to the other birds in the *is* context, by virtue of being pretty. By contrast, the *can* context collapses differences among the plants, because in the network's world, all plants can do only one thing: grow.

Figure 5.6. Barplot showing that activation of the nonsense property "queem" in an extended version of the model when the network is queried with various inputs, after it has learned that the maple "can queem," "has a queem," or "is queem." If the network learns the new property after 500 epochs of training, the property generalizes across the entire superordinate category, regardless of the relation context. When the network is taught the novel property after 2500 epochs of training, it shows different patterns of generalization depending on whether "queem" is understood to be a behavior, a part, or a physical attribute.

References

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review, 98*(3), 409-426.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of Mathematical Psychology, 39*(2), 216-233.

Barsalou, L. (1993). Flexibility, structure, and linguistic vagary in concepts: Manifestations of a compositional system of perceptual symbols. In A. F. C. a. S. E. Gathercole (Ed.), *Theories of memory*.Hillsdale, NJ: Lawrence Erlbaum Associates.

Brown, R. (1958). How shall a thing be called? *Psychological Review, 65*, 14-21.

Carey, S. (1985). *Conceptual change in childhood*.Cambridge, MA: MIT Press.

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82*, 407-428.

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior, 8*, 240-247.

Gelman, R., & Williams, E. M. (1998). Enabling constraints for cognitive development and learning: A domain-specific epigenetic theory. In D. K. a. R. Siegler (Ed.), *Handbook of Child Psychology, Volume II: Cognition, perception and development* (Vol. 2, pp. 575-630). New York: John Wiley and Sons.

Gelman, S., & Coley. (1990). The importance of knowing a dodo is a bird: Categories and inferences in 2-year-old children. *Developmental Psychology, 26*, 796-804.

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

Gelman, S. A., & Wellman, H. M. (1991). Insides and essences: Early understandings of the nonobvious. *Cognition, 38*, 213--244.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (submitted). Probabilistic models of cognition: Exploring the laws of thought.

Griffiths, T. L., Sanborn, A. N., Canini, D. J., & Navarro, D. J. (2008). Categorization as nonparametric Bayesian density estimation. In N. Chater & M. Oaksford (Eds.), *The probabilistic mind: Prospects for a Bayesian cognitive science*.Oxford, UK: Oxford University Press.

Hinton, G. E. (1981). Implementing semantic networks in parallel hardware. In G. E. H. a. J. A. Anderson (Ed.), *Parallel Models of Associative Memory* (pp. 161-187). Hillsdale, NJ: Erlbaum.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Cognitive Science Society* (pp. 1-12). Hillsdale, NJ: LEA.

Hinton, G. E., & McClelland, J. L. (1988). Learning representations by recirculation. In D. Z. Anderson (Ed.), *Neural Information Processing Systems* (pp. 358-366). New York: American Institute of Physics.

Jolicoeur, P., Gluck, M., & Kosslyn, S. M. (1984). Pictures and names: Making the connection. *Cognitive Psychology, 19*, 31-53.

Jones, S. S., Smith, L. B., & Landau, B. (1991). Object properties and knowledge in early lexical learning. *Child Development, 62*(3), 499-516.

Kemp, C., & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences, 105*(31), 10687-10692.

Kruschke, J. K. (1992). ALCOVE: An exemplar-based connectionist model of category learning. *Psychological Review, 99*(1), 22-44.

Love, B. C., Medin, D. L., & Gureckis, T. M. (2004). SUSTAIN: A network model of category learning. *Psychological Review, 111*(2), 3009-3332.

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

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognition and Development, 1*, 3-36.

Mandler, J. M., & Bauer, P. J. (1988). The cradle of categorization: Is the basic level basic? *Cognitive Development, 3*, 247-264.

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.

Mcclelland, J. L., Botvinick, M. B., Noelle, D., Rogers, T. T., Seidenberg, M., & Smith, L. (submitted). Letting structure emerge.

Mcclelland, J. L., Mcnaughton, B. L., & Oreilly, R. C. (1995). Why There Are Complementary Learning-Systems in the Hippocampus and Neocortex - Insights from the Successes and Failures of Connectionist Models of Learning and Memory. *Psychological Review, 102*(3), 419-457.

McClelland, J. L., & Rumelhart, D. E. (1986). A distributed model of human learning and memory. In J. L. McClelland, D. E. Rumelhart & t. P. R. Group (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition* (Vol. 2, pp. 170-215). Cambridge, MA: MIT Press.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises*.Cambridge, MA: MIT Press.

Medin, D. L., & Shaffer, M. M. (1978). Context theory of classification learning. *Psychological Review, 85*, 207-238.

Mervis, C. A., & Crisafi, M. A. (1982). Order of acquisition of subordinate-, basic-, and superordinate-level categories. *Child Development, 53*(1), 258-266.

Mervis, C. B. (1987). Child basic object categories and early lexical development. In U. Neisser (Ed.), *Concepts and conceptual development: Ecological and intellectual factors in categorization*.Cambridge, England: Cambridge University Press.

Mervis, C. B., & Rosch, E. (1981). Categorization of natural objects. *Annual Review of Psychology, 32*, 89-115.

Movellan, J. R., & McClelland, J. L. (1993). Learning continuous probability distributions with symmetric diffusion networks. *Cognitive Science, 17*, 463-496.

Murphy, G. L., & Brownell, H. H. (1985). Category differentiation in object recognition: Typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11*(1), 70-84.

Murphy, G. L., & Lassaline, M. E. (1997). Hierarchical structure in concepts and the basic level of categorization. In K. L. a. D. Shanks (Ed.), *Knowledge, concepts and categories* (pp. 93-131). Hove, East Sussex, UK: Psychology Press.

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

Murphy, G. L., & Smith, E. E. (1982). Basic level superiority in picture categorization. *Journal of Verbal Learning and Verbal Behavior, 21*, 1-20.

Nosofsky, R. (1984). Choice, similarity, and the context theory of classification. *Journal of experimental psychology: Learning, memory, and cognition, 10*, 104-110.

Nosofsky, R. M. (1986). Attention, similiarity and the identification-categorization relationship. *Journal of Experimental Psychology: Learning, Memory and Cognition, 115*(1), 39-57.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8*, 895--938.

Palmeri, T. J. (1999). Learning categories at different hierarchical levels: A comparison of category learning models. *Psychonomic Bulletin and Review, 6*, 495-503.

Patterson, K., & Hodges, J. (2000). Semantic dementia: one window on the structure and organisation of semantic memory. In J. Cermak (Ed.), *Handbook of Neuropsychology vol.2, Memory and its Disorders* (pp. 313-333). Amsterdam: Elsevier Science.

Patterson, K., Nestor, P. J., & Rogers, T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience, 8*, 976-987.

Pauen, S. (2002). Evidence for knowledge-based category discrimination in infancy. *Child Development, 73*(4), 1016-1033.

Pauen, S. (2002). The global-to-basic shift in infants' categorical thinking: First evidence from a longitudinal study. *International Journal of Behavioural Development, 26*(6), 492-499.

Pothos, E. M., & Chater, N. (2002). A simplicity principle in unsupervised human categorization. *Cognitive Science, 26*, 303-343.

Rogers, T. T., Lambon Ralph, M. A., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R., et al. (2004). The structure and deterioration of semantic memory: a computational and neuropsychological investigation. *Psychological Review, 111*(1), 205-235.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*.Cambridge, MA: MIT Press.

Rogers, T. T., & McClelland, J. L. (2008). A simple model from a powerful framework that spans levels of analysis. *Behavioral and Brain Sciences, 31*, 729-749.

Rogers, T. T., & Patterson, K. (2007). Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General, 136*(3), 451-469.

Rosch, E. (1978). Principles of categorization. In E. R. a. B. Lloyd (Ed.), *Cognition and Categorization*.Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Z. a. J. L. D. a. C. Lau (Ed.), *An Introduction to Neural and Electronic Networks* (pp. 405-420). San Diego, CA: Academic Press.

Rumelhart, D. E., Durbin, R., Golden, R., & Chauvin, Y. (1995). Backpropagation: The basic theory. In Y. C. a. D. E. Rumelhart (Ed.), *Back-Propagation: Theory, Architectures, and Applications* (pp. 1-34). Hillsdale, NJ: Erlbaum.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. R. a. J. L. M. a. t. P. R. Group (Ed.), *Parallel Distributed Processing: {Explorations} in the Microstructure of Cognition* (Vol. 1, pp. 318--362). Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*(9), 533-536.

Rumelhart, D. E., McClelland, J. L., & Group, t. P. R. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume I: Foundations & Volume II: Psychological and Biological Models*.Cambridge, MA: MIT Press.

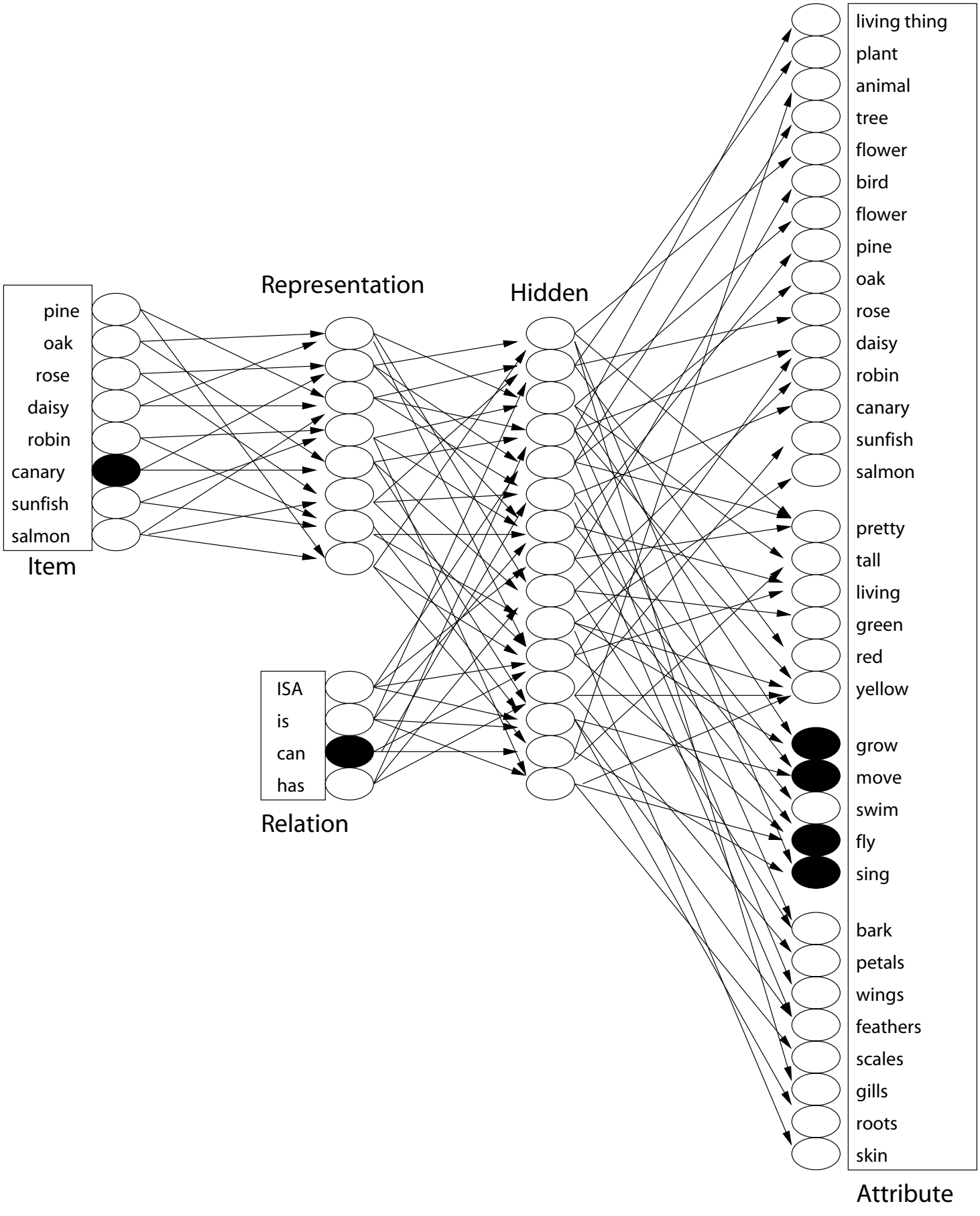
Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience* (pp. 3-30). Cambridge, MA: MIT Press.

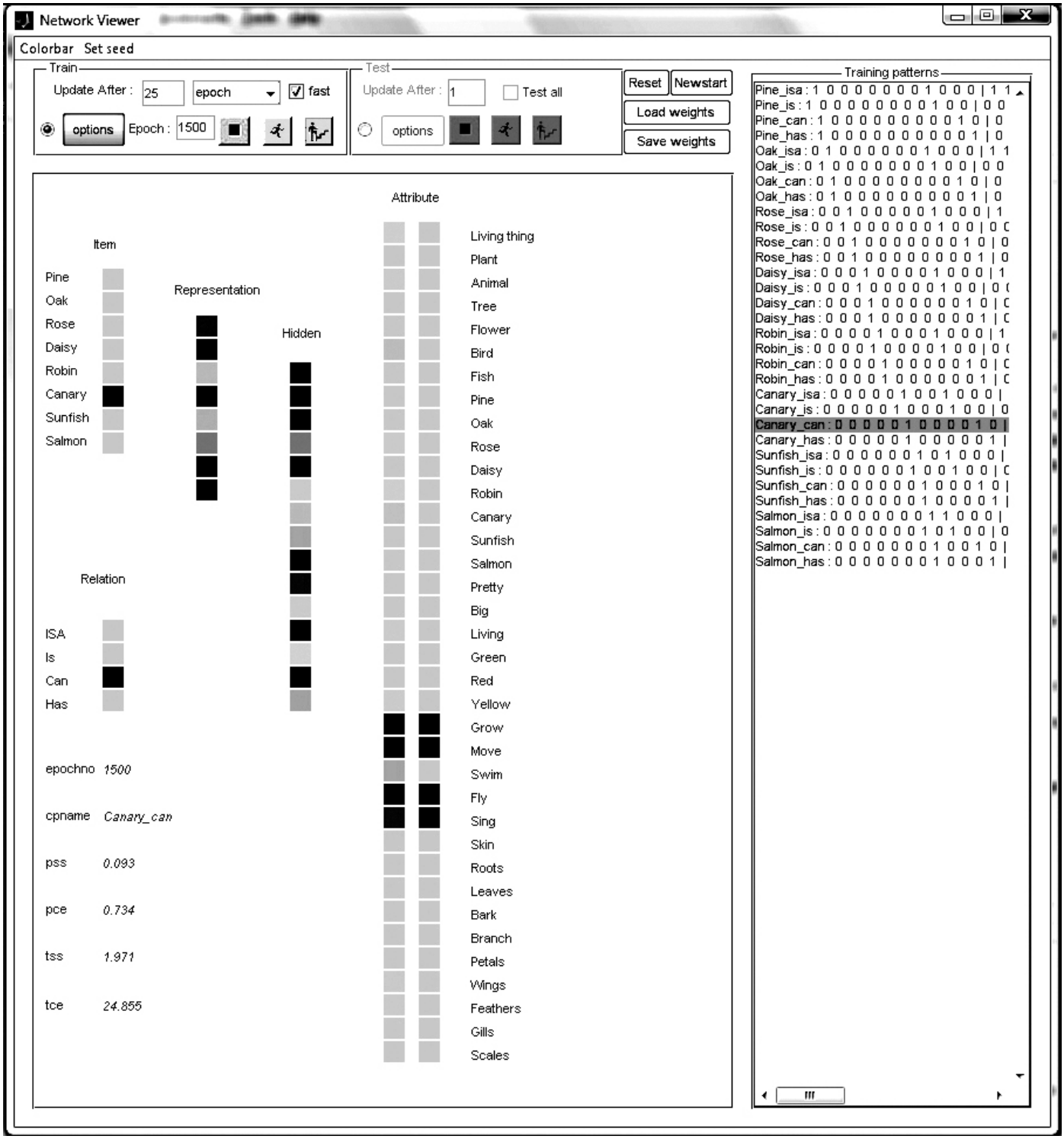
Tanaka, J., & Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology, 23*, 457-482.

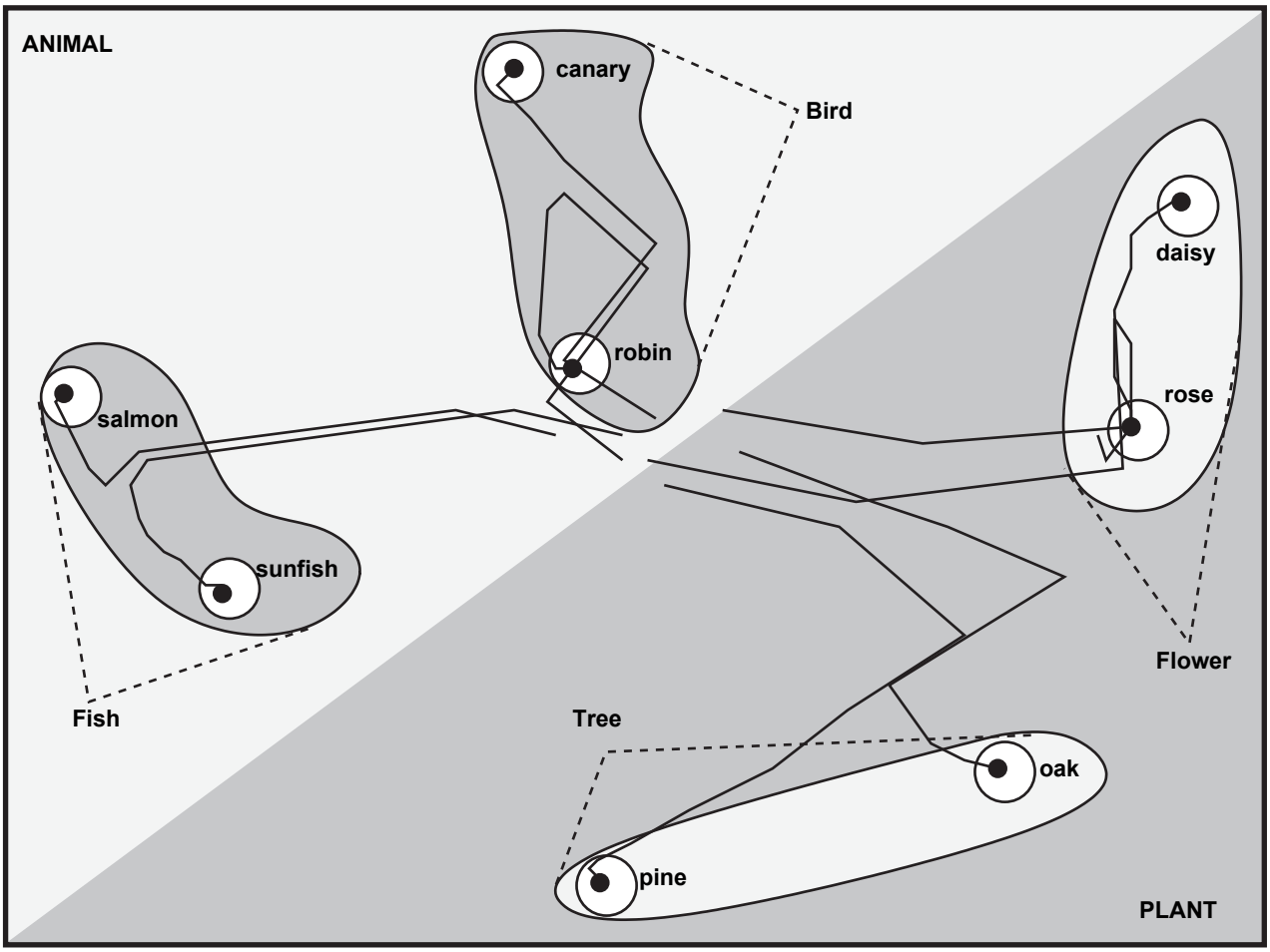
Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and Brain Sciences, 24*, 629-640.

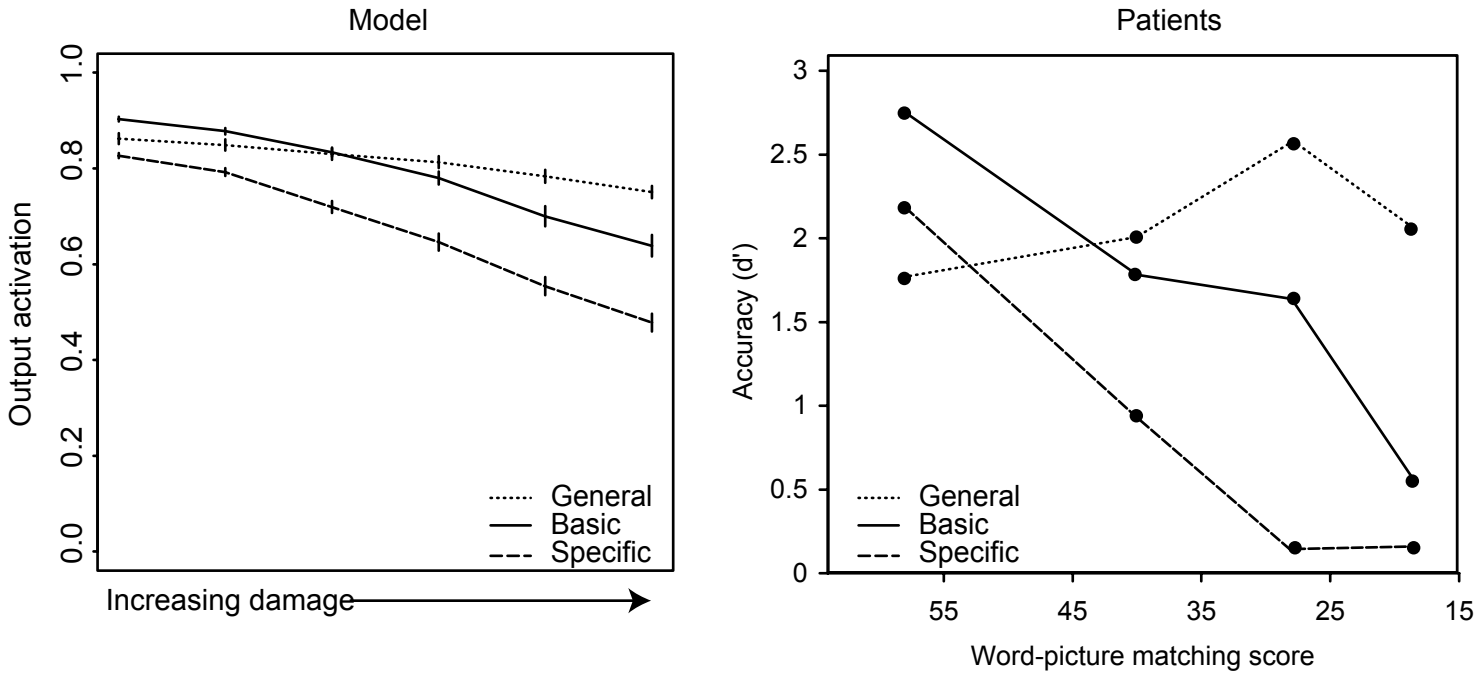
Tenenbaum, J. B., Griffiths, T. L., & Kemp, C. (2006). Theory-based Bayesian models of inductive learning and reasoning. *Trends in Cognitive Sciences, 10*(7), 309-318.

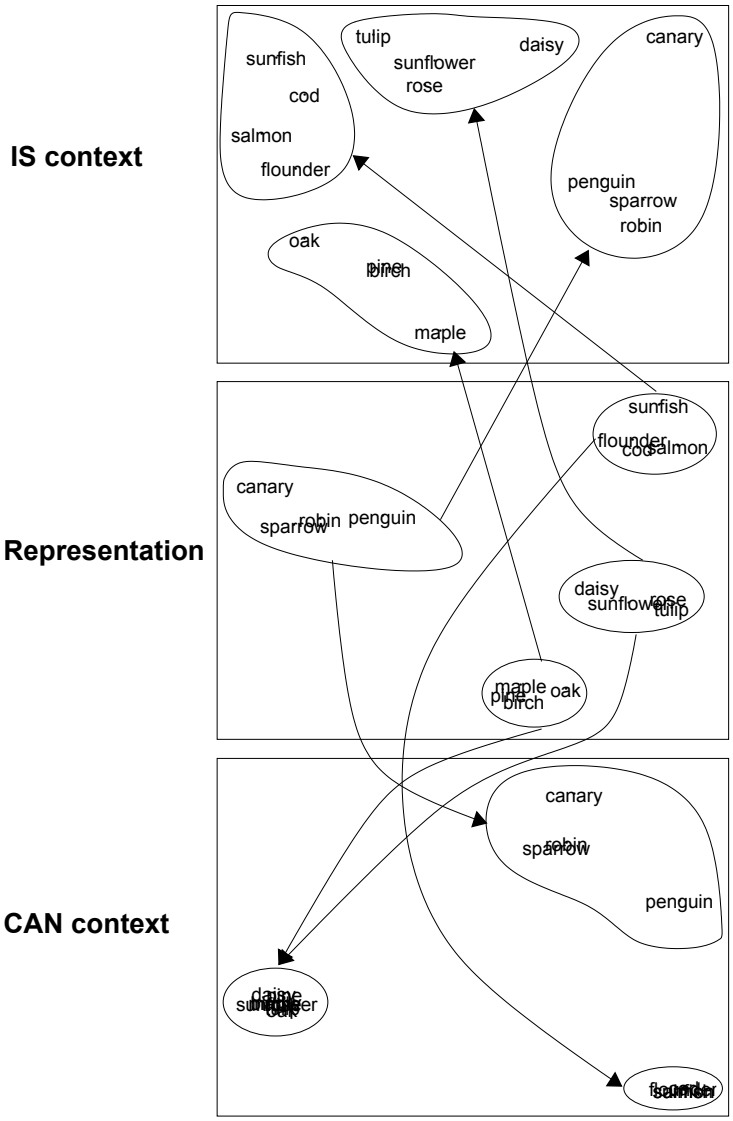
Verheyen, S., Ameel, E., Rogers, T. T., & Storms, G. (2008). *Learning a hierarchical organization of categories.* Paper presented at the Proceedings of the Cognitive Science Society, Amsterdam, the Netherlands.

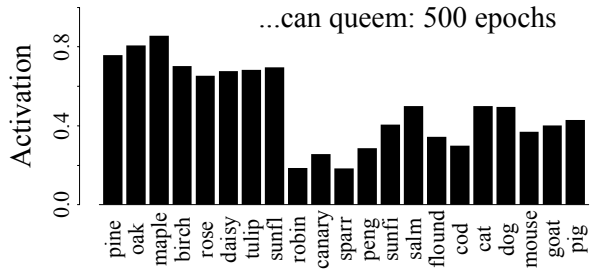

Representation
Hidden
pine
oak
rose
daisy
robin
canary
sunfish
salmon
Item
ISA
is
can
has
Relation
living thing
plant
animal
tree
flower
bird
flower
pine
oak
rose
daisy
robin
canary
sunfish
salmon
pretty
tall
living
green
red
yellow
grow
move
swim
fly
sing
bark
petals
wings
feathers
scales
gills
roots
skin
Attribute

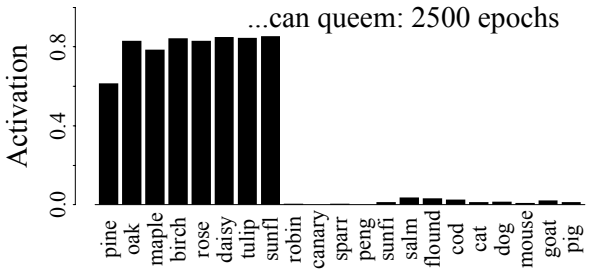

Network Viewer
Colorbar Set seed
Train
Update After : 25 epoch fast
options Epoch : 1500
Test
Update After : 1 Test all
options
Reset Newstart
Load weights
Save weights
Training patterns
Item
Pine
Oak
Rose
Daisy
Robin
Canary
Sunfish
Salmon
Representation
Hidden
Relation
ISA
Is
Can
Has
Attribute
Living thing
Plant
Animal
Tree
Flower
Bird
Fish
Pine
Oak
Rose
Daisy
Robin
Canary
Sunfish
Salmon
Pretty
Big
Living
Green
Red
Yellow
Grow
Move
Swim
Fly
Sing
Skin
Roots
Leaves
Bark
Branch
Petals
Wings
Feathers
Gills
Scales
epochno 1500
cpname Canary_can
pss 0.093
pce 0.734
tss 1.971
tce 24.855
Pine_isa : 1 0 0 0 0 0 0 0 1 0 0 0 | 1 1
Pine_is : 1 0 0 0 0 0 0 0 0 1 0 0 | 0 0
Pine_can : 1 0 0 0 0 0 0 0 0 0 1 0 | 0
Pine_has : 1 0 0 0 0 0 0 0 0 0 0 1 | 0
Oak_isa : 0 1 0 0 0 0 0 0 1 0 0 0 | 1 1
Oak_is : 0 1 0 0 0 0 0 0 0 1 0 0 | 0 0
Oak_can : 0 1 0 0 0 0 0 0 0 0 1 0 | 0
Oak_has : 0 1 0 0 0 0 0 0 0 0 0 1 | 0
Rose_isa : 0 0 1 0 0 0 0 0 1 0 0 0 | 1
Rose_is : 0 0 1 0 0 0 0 0 0 1 0 0 | 0
Rose_can : 0 0 1 0 0 0 0 0 0 0 1 0 | 0
Rose_has : 0 0 1 0 0 0 0 0 0 0 0 1 | 0
Daisy_isa : 0 0 0 1 0 0 0 0 1 0 0 0 | 1
Daisy_is : 0 0 0 1 0 0 0 0 0 1 0 0 | 0
Daisy_can : 0 0 0 1 0 0 0 0 0 0 1 0 | 0
Daisy_has : 0 0 0 1 0 0 0 0 0 0 0 1 | 0
Robin_isa : 0 0 0 0 1 0 0 0 1 0 0 0 | 1
Robin_is : 0 0 0 0 1 0 0 0 0 1 0 0 | 0
Robin_can : 0 0 0 0 1 0 0 0 0 0 1 0 | 0
Robin_has : 0 0 0 0 1 0 0 0 0 0 0 1 | 0
Canary_isa : 0 0 0 0 0 1 0 0 1 0 0 0 |
Canary_is : 0 0 0 0 0 1 0 0 0 1 0 0 | 0
Canary_can : 0 0 0 0 0 1 0 0 0 0 1 0 |
Canary_has : 0 0 0 0 0 1 0 0 0 0 0 1 |
Sunfish_isa : 0 0 0 0 0 0 1 0 1 0 0 0 |
Sunfish_is : 0 0 0 0 0 0 1 0 0 1 0 0 | 0
Sunfish_can : 0 0 0 0 0 0 1 0 0 0 1 0 |
Sunfish_has : 0 0 0 0 0 0 1 0 0 0 0 1 |
Salmon_isa : 0 0 0 0 0 0 0 1 1 0 0 0 |
Salmon_is : 0 0 0 0 0 0 0 1 0 1 0 0 | 0
Salmon_can : 0 0 0 0 0 0 0 1 0 0 1 0 |
Salmon_has : 0 0 0 0 0 0 0 1 0 0 0 1 |

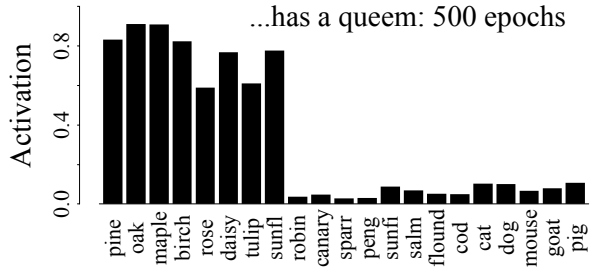

ANIMAL
canary
Bird
robin
daisy
rose
salmon
sunfish
Fish
Flower
Tree
oak
pine
PLANT

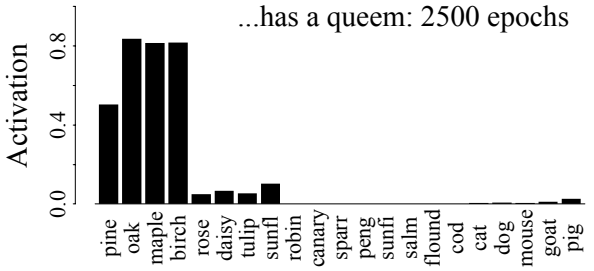

Model
Output activation
0.0
0.2
0.4
0.6
0.8
1.0
General
Basic
Specific
Increasing damage
Patients
Accuracy (d')
0
0.5
1
1.5
2
2.5
3
General
Basic
Specific
55
45
35
25
15
Word-picture matching score

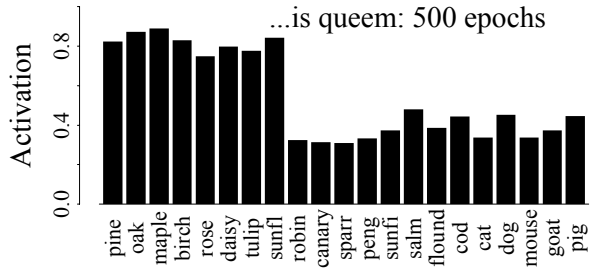

IS context
sunfish
cod
salmon
flounder
tulip
daisy
sunflower
rose
canary
penguin
sparrow
robin
oak
pine
birch
maple
Representation
sunfish
flounder
cod
salmon
canary
sparrow
robin
penguin
daisy
sunflower
rose
tulip
maple
oak
birch
CAN context
canary
sparrow
robin
penguin

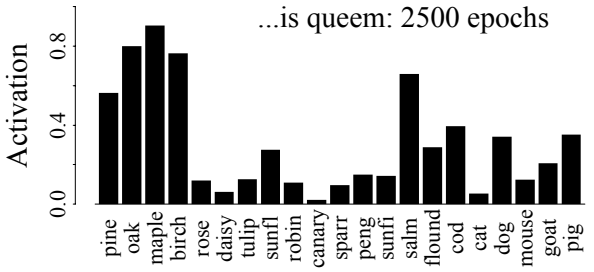

...can queem: 500 epochs
Activation
0.0
0.4
0.8
pine
oak
maple
birch
rose
daisy
tulip
sunfl
robin
canary
sparr
peng
sunfi
salm
flound
cod
cat
dog
mouse
goat
pig


...can queem: 2500 epochs
Activation
0.0
0.4
0.8
pine
oak
maple
birch
rose
daisy
tulip
sunfl
robin
canary
sparr
peng
sunfi
salm
flound
cod
cat
dog
mouse
goat
pig


...has a queem: 500 epochs
Activation
0.0
0.4
0.8
pine
oak
maple
birch
rose
daisy
tulip
sunfl
robin
canary
sparr
peng
sunfi
salm
flound
cod
cat
dog
mouse
goat
pig


...has a queem: 2500 epochs
Activation
0.0
0.4
0.8
pine
oak
maple
birch
rose
daisy
tulip
sunfl
robin
canary
sparr
peng
sunfi
salm
flound
cod
cat
dog
mouse
goat
pig


...is queem: 500 epochs
Activation
0.0
0.4
0.8
pine
oak
maple
birch
rose
daisy
tulip
sunfl
robin
canary
sparr
peng
sunfi
salm
flound
cod
cat
dog
mouse
goat
pig


...is queem: 2500 epochs
Activation
0.0
0.4
0.8
pine
oak
maple
birch
rose
daisy
tulip
sunfl
robin
canary
sparr
peng
sunfi
salm
flound
cod
cat
dog
mouse
goat
pig