BEHAVIORAL AND BRAIN SCIENCES (2008) **31**, 689–749
Printed in the United States of America
doi:10.1017/S0140525X0800589X

Précis of *Semantic Cognition: A Parallel Distributed Processing Approach*

Timothy T. Rogers
Department of Psychology, University of Wisconsin-Madison, Madison, WI 53706
ttrogers@wisc.edu **http://concepts.psych.wisc.edu**

James L. McClelland
Department of Psychology and Center for Mind, Brain, and Computation, Stanford University, Stanford, CA 94305
mcclelland@stanford.edu **http://psychology.stanford.edu/~jlm**

Abstract: In this précis of our recent book, *Semantic Cognition: A Parallel Distributed Processing Approach* (Rogers & McClelland 2004), we present a parallel distributed processing theory of the acquisition, representation, and use of human semantic knowledge. The theory proposes that semantic abilities arise from the flow of activation among simple, neuron-like processing units, as governed by the strengths of interconnecting weights; and that acquisition of new semantic information involves the gradual adjustment of weights in the system in response to experience. These simple ideas explain a wide range of empirical phenomena from studies of categorization, lexical acquisition, and disordered semantic cognition. In this précis we focus on phenomena central to the reaction against similarity-based theories that arose in the 1980s and that subsequently motivated the "theory-theory" approach to semantic knowledge. Specifically, we consider (1) how concepts differentiate in early development, (2) why some groupings of items seem to form "good" or coherent categories while others do not, (3) why different properties seem central or important to different concepts, (4) why children and adults sometimes attest to beliefs that seem to contradict their direct experience, (5) how concepts reorganize between the ages of 4 and 10, and (6) the relationship between causal knowledge and semantic knowledge. The explanations our theory offers for these phenomena are illustrated with reference to a simple feed-forward connectionist model. The relationships between this simple model, the broader theory, and more general issues in cognitive science are discussed.

Keywords: categorization; causal knowledge; concepts; connectionism; development; innateness; learning; semantics; memory; theory-theory.

When we open our eyes and look around us, we observe a host of objects – people, animals, plants, cars, buildings, and other artifacts of many different kinds – most of which are quite familiar. We have tacit expectations about the unseen properties of these objects (e.g., what we would find underneath the skin of an orange or banana) and how the objects would react or what effects they would have if we interacted with them in various ways. Would a furry animal bite if we tried to stroke it? Would a particular artifact hold a hot liquid? We can usually name these objects, describe their visible and invisible properties to others, and make inferences about them, such as whether they would likely die if deprived of oxygen, or whether they would break if dropped onto a concrete floor. Understanding the basis of these abilities – to recognize, comprehend, and make inferences about objects and events in the world, and to comprehend and produce statements about them – is the goal of research in semantic cognition. Since antiquity, philosophers have considered how we make semantic judgments, and the investigation of semantic processing was a focal point for both experimental and computational investigations in the early phases of the cognitive revolution. Yet the mechanistic basis of semantic cognition remains very much open to question.

In the 1960s and early '70s, the predominating view held that semantic knowledge was encoded in a vast set of stored propositions, and theories of the day offered explicit proposals about the organization of such propositions in memory, and about the nature of the processes employed to retrieve particular propositions from memory (e.g., Collins & Loftus 1975; Collins & Quillian 1969). The mid-70s, however, saw the introduction of findings on the gradedness of category membership and on the privileged status of some categories that such "spreading activation" theories did not encompass (Rips et al. 1973; Rosch & Mervis 1975; Rosch et al. 1976). These findings subsequently gave rise to a family of "similarity-based" approaches proposing that semantic information is encoded in feature-based representations – category prototypes or representations of individual instances – and that retrieval of semantic information depends in some way upon the similarity between a probe item and these stored representations (Smith & Medin 1981). Like spreading-activation theories, similarity-based approaches advanced specific hypotheses about the nature of the

 0140-525X/08 $40.00

stored representations and of the mechanisms by which semantic information is retrieved (e.g., Hampton 1993; Kruschke 1992; Nosofsky 1984; 1986); but these in turn have been subject to serious and challenging criticism arising from a theoretical framework often called the "theory-theory" (Carey 1985; Gopnik & Meltzoff 1997; Keil 1989; Murphy & Medin 1985).

The theory-theory proposes that semantic knowledge is rooted in a system of implicit beliefs about the causal forces that give rise to the observable properties of objects and events. On this view, implicit and informal causal theories determine which sets of items should be treated as similar for purposes of induction and generalization, which properties are important for determining category membership, which properties will be easy to learn and which difficult, and so on. Conceptual development is viewed as arising (at least in part) from change to the implicit causal theories that structure concepts. This framework has been very useful as a springboard for powerful experimental demonstrations of the subtlety and sophistication of the semantic judgments adults and even children can make, and for highlighting the serious challenges faced by similarity-based and spreading-activation theories. In contrast to those frameworks, however, the theory-theory has not provided an explicit mechanistic account of the representation and use of semantic knowledge. The fundamental tenets of the theory-theory are general principles whose main use has been to guide the design of ingenious experiments rather than the formulation of explicit proposals about the nature and structure of semantic representations or the mechanisms that process semantic information.

In what follows, we provide a précis of our recent book, *Semantic Cognition: A Parallel Distributed Processing Approach* (Rogers & McClelland 2004, henceforth simply *Semantic Cognition* in this précis), which puts forward a theory about the cognitive mechanisms that support semantic abilities based on the domain-general principles of the connectionist or parallel distributed processing framework. Our approach captures many of the appealing aspects of spreading-activation and similarity-based theories while resolving some of the apparent paradoxes they face; and it addresses many of the phenomena that have motivated theory-theory and related approaches within an alternative, more mechanistic, framework. The book illustrates how a simple model instantiating the theory addresses, among other things, classic findings from studies of semantic cognition in infancy and childhood; the influence of frequency, typicality, and expertise on semantic cognition in adulthood; basic-level effects in children and adults; and the progressive disintegration of conceptual knowledge observed in some forms of dementia. In this précis, however, we focus on phenomena that were central to the critical reaction against similarity-based theories and that subsequently motivated the appeal to theory-based approaches. These phenomena are briefly summarized in Table 1, and are explained in further detail in what follows. We emphasize these particular phenomena because they are often thought to challenge the notion that semantic abilities might arise from general-purpose learning mechanisms, and to support the view that such abilities must arise from initial domain-specific knowledge, via domain-specific learning systems.

These issues are central to questions about what makes us uniquely human. Do we possess, at birth, and by virtue of evolution, a set of highly specialized cognitive modules tailored to support knowledge about particular domains? Or do our advanced semantic abilities reflect the operation of a powerful learning mechanism capable of acquiring, through experience, knowledge about all semantic domains alike? A key point of our book is that the learning mechanisms adopted within the connectionist approach to cognition are quite different from classical associationist learning; that the capabilities of connectionist models have been under-appreciated in this respect; and that such models can provide an intuitive explanation of how domain-general learning supports the emergence of semantic and conceptual knowledge over the course of development. The models we describe employ domain-general learning mechanisms, without initial knowledge or domain-specific constraints. Thus, if they adequately capture the phenomena listed in Table 1, this calls into question the necessity of invoking initial domain-specific knowledge to explain semantic cognition.

The particular models we will use throughout our discussion are variants of a model described by Rumelhart (Rumelhart 1990; Rumelhart & Todd 1993), which in turn built on previous proposals by Hinton (1981; 1986). We will therefore begin, in section 1, with a description of Rumelhart's model and how it works, followed by a brief explanation of the more general theory the model is intended to exemplify. In section 2, "Accounting for the phenomena," we will consider how the theory explains the phenomena listed in Table 1, using simulations with variants of the Rumelhart model to illustrate the substantive points. With a more complete understanding of the implications of the theory before us (sect. 3), we then consider, in section 4, "Contrasting the PDP and theory-based approaches," how our theory relates to the theory-theory. In section 5, "Principles of the PDP approach to semantic cognition," we summarize more general aspects of the

TIMOTHY ROGERS is an Assistant Professor of Psychology at the University of Wisconsin–Madison, a position he has held since 2004. He received his Ph.D. in Psychology from Carnegie Mellon University. His research focuses on the cognitive, biological, and computational mechanisms that support human behavior in conceptual tasks such as language comprehension, induction and inference, object-recognition and categorization. He is currently interested in understanding how semantic, control, and perceptual systems interact in a variety of everyday tasks.

JAMES L. (JAY) MCCLELLAND is a Professor of Psychology and the Director of the Center for Mind, Brain, and Computation at Stanford University. McClelland received his Ph.D. in Cognitive Psychology from the University of Pennsylvania in 1975. He then went to the University of California at San Diego where he collaborated with David E. Rumelart in research leading to the two-volume work, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. He later spent twenty-two years at Carnegie Mellon University in Pittsburgh where he was a founding Co-Director of the Center for the Neural Basis of Cognition. His research interests include representation and learning in hippocampus and neocortex, decision processes, and semantic cognition.

Table 1. *Six key phenomena in the study of semantic abilities*

Phenomenon	Example
Progressive differentiation of concepts	Children acquire broader semantic distinctions earlier than more fine-grained distinctions. For example, when perceptual similarity among items is controlled, infants differentiate animals from furniture around 7–9 months of age, but do not make finer-grained distinctions (e.g., between fish and birds or chairs and tables) until somewhat later (Pauen 2002a; Mandler et al. 1991); and a similar pattern of coarse-to-fine conceptual differentiation can be observed between the ages of 4 and 10 in verbal assessments of knowledge about which predicates can appropriately apply to which nouns (Keil 1989).
Category coherence	Some groupings of objects (e.g., "the set of all things that are dogs") seem to provide a useful basis for naming and inductive generalization, whereas other groupings (e.g., "the set of all things that are blue") do not. How does the semantic system "know" which groupings of objects should be used for purposes of naming and inductive generalization, and which should not?
Domain-specific attribute weighting	Some properties seem of central importance to a given concept, whereas others do not. For instance, "being cold inside" seems important to the concept *refrigerator*, whereas "being white" does not. Furthermore, properties that are central to some concepts may be unimportant for others – although having a white color may seem unimportant for *refrigerator*, it seems more critical to the concept *polar bear*. What are the mechanisms that support domain-specific attribute weighting?
Illusory correlations	Children and adults sometimes attest to beliefs that directly contradict their own experience. For example, when shown a photograph of a kiwi bird – a furry-looking animal with eyes but no discernible feet – children may assert that the animal can move "because it has feet," even while explicitly stating that they can see no feet in the photograph. Such illusory correlations appear to indicate some organizing force behind children's inferences that goes beyond "mere" associative learning. What mechanisms promote illusory correlations?
Conceptual reorganization	Children's inductive projection of biological facts to various different plants and animals changes dramatically between the ages of 4 and 10. For some researchers, these changing patterns of induction indicate changes to the implicit theories that children bring to bear on explaining biological facts. What mechanism gives rise to changing induction profiles over development?
The importance of causal knowledge	A variety of evidence now indicates that, in various kinds of semantic induction tasks, children and adults strongly weight causally central properties over other salient but non-causal properties. Why are people sensitive to causal properties?

current work that we believe to be particularly critical to understanding semantic abilities. In "Broader issues," section 6, we discuss implications of the present work for cognitive science more generally.

The material here is largely excerpted from *Semantic Cognition*, with some restructuring, condensation, and minor corrections. In the interest of providing a relatively succinct overview of the theory, we have omitted substantial detail, both in the range of phenomena to which the model has been applied and in the descriptions of the simulations themselves. Where we feel these details may prove especially useful, we refer the reader to the corresponding section of the book. We have avoided adding new material addressing work completed since *Semantic Cognition* appeared; where relevant, such material will arise in our response to open peer commentary.

1. The PDP framework

As previously mentioned, the models we will use to illustrate the theory are variants of an architecture first proposed by Rumelhart (Rumelhart 1990; Rumelhart & Todd 1993) as illustrated in Figure 1. Rumelhart was interested in understanding how the propositional information stored in a hierarchical propositional model such as that shown in Figure 2 could be acquired and processed by a connectionist network employing distributed internal representations. Thus, the individual nodes in the Rumelhart network's input and output layers correspond to the constituents of propositions – the items that occupy the first (subject) slot in each proposition, relation terms that occupy the second slot, and the attribute values that occupy the third slot. Each item is represented by an individual input unit in the layer labeled *Item*, each relation is represented by the individual units in the layer labeled *Relation*, and the various possible completions of three-element propositions are represented by individual units in the layer labeled *Attribute*. When presented with a particular *Item* and *Relation* pair in the input, the network's job is to turn on the attribute units in the output that correspond to valid completions of the proposition. For example, when the units corresponding to *canary* and *can* are activated in the input, the network must learn

to activate the output units *move*, *grow*, *fly*, and *sing*. The particular items, relations, and attributes used by Rumelhart and Todd (1993) were taken directly from the hierarchical propositional model described by Collins and Quillian (1969; see Fig. 2), so that, when the network has learned to correctly complete all of the propositions, it has encoded the same information stored in that propositional hierarchy.

The network consists of a series of nonlinear processing units, organized into layers, and connected in a feed-forward manner, as shown in Figure 1. Patterns are presented by activating one unit in each of the *Item* and *Relation* layers, and allowing activation to spread forward through the network, modulated by the connection weights. To update a unit's activation, its net input is first calculated by summing the activation of each unit from which it receives a connection multiplied by the value of the connection weight; this is then transformed to an activation according to the logistic transfer function.

To find an appropriate set of weights, the network is trained with backpropagation (Rumelhart et al. 1986a). First, an item and relation are presented to the network,

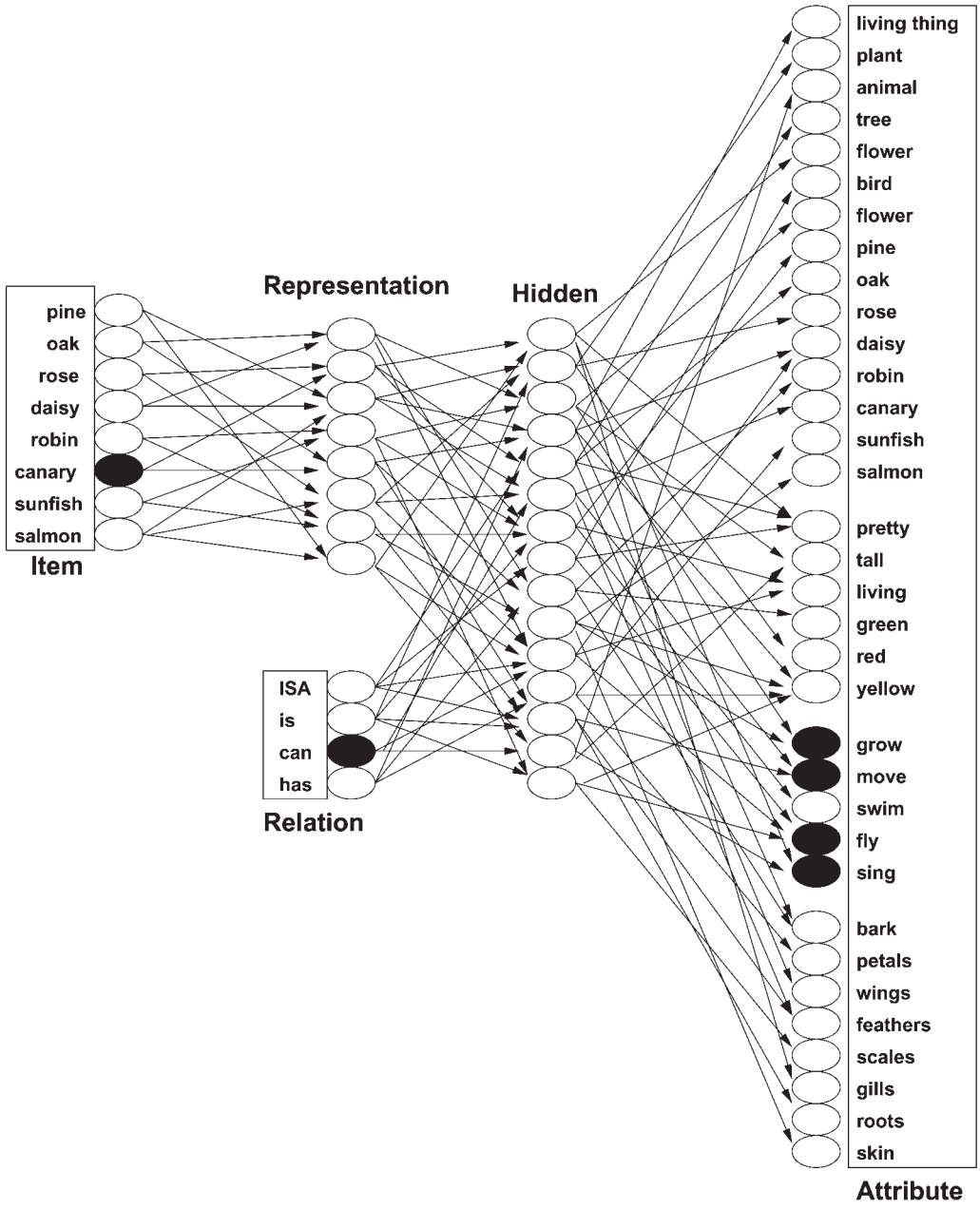

Figure 1. A connectionist model of semantic memory adapted from Rumelhart and Todd (1993), used to learn all the propositions true of the specific concepts (pine, oak, etc.) in the Collins and Quillian model (Fig. 2). Input units are shown on the left, and activation propagates from the left to the right. 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.

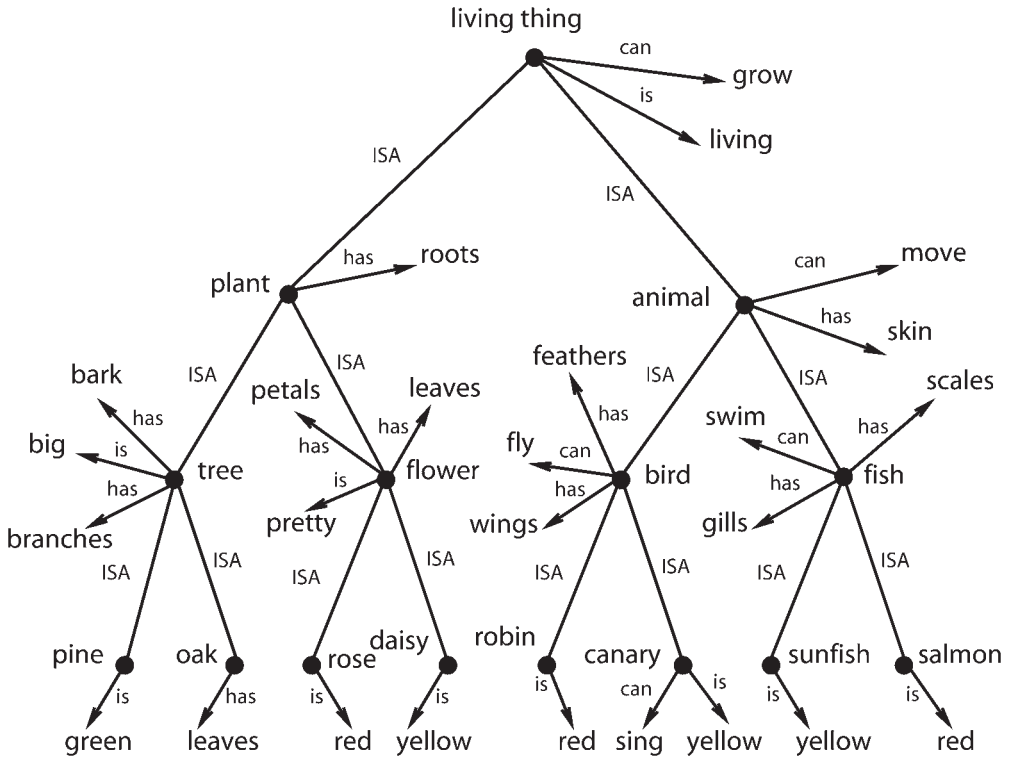

Figure 2. A taxonomic hierarchy of the type used by Collins and Quillian (1969) in their model of the organization of knowledge in memory. The schematic indicates that living things can grow; that a plant is a living thing; that a tree is a plant; and that an oak is a tree. It therefore follows that an oak can grow. The training corpus for the Rumelhart model incorporates all propositions pertaining to the eight subordinate items (pine, oak, rose, etc.) that can be derived from this tree.

and activation is propagated forward to the output units. The observed output states are then compared to the desired or target values, and the difference is converted to a measure of error. The partial derivative of the error with respect to each weight in the network is computed in a backward pass, and the weights are adjusted by a small amount to reduce the discrepancy. Because the model's inputs are localist, all items in its environment are equally distinct from one another in the input – the model's input representation of the robin and canary, for instance, are no more similar to one another than either is to the input representation of the rose. Each individual *Item* unit projects, however, to all of the units in the layer labeled *Representation*. The activation of a single item in the model's input, then, generates a distributed pattern of activity across these units. The weights connecting *Item* and *Representation* units evolve during learning, so the pattern of activity generated across the *Representation* units for a given item is a learned internal representation of the item.

Though the model's inputs and outputs are constrained to locally represent particular items, attributes, and relations, the learning process allows it to derive distributed internal representations that do not have this localist character. In contrast to some other connectionist theories, the units that encode learned internal representations in the model have no explicit content in themselves – they do not correspond to semantic features, propositions, images, or other explicit representations. Thus, it is impossible to determine what the network "knows" solely by inspecting the activation of these internal units. Instead, the network's knowledge must be probed by querying it with an appropriate input, then inspecting the response it generates in the output. Although the learned internal representations have no directly interpretable content, they do subserve a critical function: for reasons elaborated further on, they turn out to capture the semantic similarity relations that exist among the items in the network's training environment, and so provide a basis for semantic generalization. Obviously, the model's behavior in this respect depends on the particular values of the connection weights when tested. Since the values of these connection weights change with experience, the model's generalization behavior strongly depends on the extent and nature of its prior experience with the items in its environment.

Although Rumelhart conceived of this network as encoding and processing propositional content, we view the model as a very simple implementation of a more general theoretical approach to semantic cognition (also exemplified in other related work; see Rumelhart et al. 1986c; McClelland & Rumelhart 1986; McClelland et al. 1989; 1995). Under this approach, the main function of the semantic system is to support performance on tasks that require one to generate, from perceptual or linguistic input, properties of objects and events that are not directly apparent in the environment. The representations that support semantic task performance consist of patterns of activity across a set of units in a connectionist network, with semantically related objects represented by similar patterns of activity. In a given semantic task, these representations may be constrained both by incoming information about the item of interest (in the form of a verbal description, a visual image, or other sensory information)

and by the context in which the item is encountered. Thus, we envision that the two parts of the input in the model – the *Item* and *Context* units – represent a perceived object (perhaps foregrounded for some reason to be in the focus of attention) and a context provided by other information available together with the perceived object. Different item/context input pairs provoke different patterns of activation across internal representation units; and the instantiation of any particular pattern of activation propagates forward to allow the system to generate an output specifying the relevant object properties, which are encoded in the model's outputs.

For instance, the situation may be analogous to one in which a young child is looking at a robin on a branch of a tree, and sees that, as a cat approaches, the robin suddenly flies away. The object and the situation together provide a context in which it would be possible for an experienced observer to anticipate that the robin will fly away; and the observation that it does would provide input allowing a less experienced observer to develop such an anticipation. Conceptually speaking, this is how we see learning occurring in preverbal conceptual development: An object encountered in a particular situation gives rise to implicit predictions which are subsequently met or violated. (Initially the predictions may be very general or even null, and are inherently graded). The discrepancy between expected and observed outcomes then serves as the basis for adjusting the connection weights that support prediction – thus allowing experience to drive change in both the internal representations of objects and events and the predictions about observable outcomes. In the Rumelhart model, the presentation of the "object" corresponds to the activation of one of the *Item* input units; the situation in which the item is encountered corresponds to the activation of one of the *Context* units; the child's expectations about the outcome of the event may be equated with the model's outputs; and the presentation of the actual observed outcome is analogous to the presentation of the target for the output units in the network. On this view, the environment provides both the input that characterizes a situation as well as the information about the outcome that then drives the process of learning. This outcome information will consist sometimes of verbal, sometimes of nonverbal information, and in general is construed as information filtered through perceptual systems, no different in any essential way from the information that drives the *Item* and *Context* units in the network.

We can also see that there is a natural analog in the model for the distinction drawn between the perceptual information available from an item in a given situation, and the conceptual representations that are derived from this information. Specifically, the model's input, context, and targets code the "perceptual" information that is available from the environment in a given episode; and the intermediating units in the *Representation* and *Hidden* layers correspond to the "conceptual" representations that allow the semantic system to accurately perform semantic tasks.

In what follows, we will show how these simple ideas account for a surprisingly broad variety of phenomena in the study of semantic cognition, paying particular attention to the six phenomena listed in Table 1. Accounting for the phenomena will allow us to illustrate certain interesting properties of the model, which in turn will allow us to articulate the general theory more completely.

2. Accounting for the phenomena

2.1. Progressive differentiation of concept representations

Although infants from a very young age are sensitive to perceptual similarities among objects in their world (e.g., Eimas & Quinn 1994; Mareschal 2000), there is now considerable evidence that knowledge about semantic similarity relations is acquired somewhat later and follows a predictable developmental trajectory (e.g., Mandler & McDonough 1993; 1996; Mandler et al. 1991). Specifically, children appear to acquire broader semantic distinctions earlier than more fine-grained distinctions. For example, when perceptual similarity among items is controlled, infants differentiate animals from furniture around 7–9 months of age, but do not make finer-grained distinctions (e.g., between fish and birds or chairs and tables) until somewhat later (Mandler et al. 1991; Pauen 2002a). A similar pattern of coarse-to-fine conceptual differentiation can be observed over the elementary school years in assessments of knowledge about which predicates can appropriately apply to which nouns (Keil 1979).

The contention that children acquire broad semantic distinctions before narrower ones seemingly contradicts an alternative long-standing view that children acquire "basic-level" concepts like dog or car prior to more general (e.g., animal, vehicle) or specific (labrador, limousine) concepts (e.g., Mervis 1987). The main support for this view stems from two sources. First, preferential-looking studies have shown that infants as young as 3 months of age are capable of "categorizing" at the basic-level. For instance, habituation to photographs of cats will generalize to novel pictures of cats, but not to photographs of horses, suggesting that the infants treat the different cats as similar to one another and as different from the horses (Eimas & Quinn 1994). Such results are only observed, however, when perceptual similarity is high within category and low between categories (e.g., Quinn & Johnson 2000). Hence, they may not reflect the infant's pre-existing semantic knowledge about cats and horses, but may instead indicate an ability to rapidly extract information about perceptual similarity over the course of the experiment (as indeed very young infants have been shown to do in random-dot category learning studies; see Bomba & Siqueland 1983). In contrast, recent studies by Pauen (2002a; 2002b) suggest that, when perceptual similarity is closely controlled, preverbal infants in object-manipulation tasks differentiate more general semantic categories prior to basic-level categories.

Second, studies of lexical acquisition have shown that, for fairly familiar items, children learn basic-level labels (e.g., "dog") prior to more general ("animal") and more specific ("labrador") labels (Brown 1958; Mervis 1987). On our reading of the literature, these findings are robust, but they reflect constraints on word learning that arise sometime after children have begun to differentiate concepts at both general and basic levels. That is, the general-before-basic pattern documented in the work of Mandler et al. (1991) and Pauen (2002a) occurs between

7 and 9 months of age, before children have begun to name things; and the basic-before-general pattern observed during word learning arises because, by the time children are learning to name, they are already representing items from different basic-level categories as quite distinct from one another, even if they are from the same general semantic domain.

In Chapter 5 of *Semantic Cognition*, we show that the basic-before-general trend in naming can coexist in the model with general-before-basic differentiation of the underlying conceptual representations. We also provide a detailed treatment of basic-level effects in lexical acquisition and in adulthood and consider how and why such effects change with expertise and in some forms of dementia. In this précis, we focus on understanding the coarse-to-fine differentiation of concepts that occurs in preverbal infants when perceptual similarity is controlled, because a full understanding of the mechanisms that produce the phenomenon in the model will provide the basis for our explanation of all of the remaining phenomena.

We trained the network shown in Figure 1 with the same corpus of propositions used by Rumelhart and Todd (1993). The corpus contains all of the propositions true of each of the eight specific concepts (pine, oak, etc.) shown in the propositional hierarchy displayed at the top of the figure. To see how the network's internal representations change over time, we stopped training at different points during learning and then stepped through the eight items, recording the states of the representation units for each. The top part of Figure 3 shows these activations at three points during learning. Initially, and even after 50 epochs of training as shown, the patterns representing the items are all very similar, with activations hovering around 0.5. At Epoch 100, the patterns corresponding to various animal instances are similar to one another, but are distinct from the plants. At Epoch 150, items from the same intermediate cluster, such as rose and daisy, have similar but distinguishable patterns, and are now easily differentiated from their nearest neighbors (e.g., pine and oak). Thus, each item has a unique representation, but semantic relations are preserved in the similarity structure across representations.

The arrangement and grouping of the representations shown in the bottom of Figure 3 reflects the similarity structure among the internal representations, as determined by a hierarchical clustering analysis. At 50 epochs the tree is very flat, and any similarity structure revealed in the plot is weak and arises from the random initial values of the connection weights. By Epoch 100 the clustering analysis reveals that the network has differentiated plants from animals: all the plants are grouped under one node, while all the animals are grouped under another. At this point, more fine-grained structure is not yet clear. For example, oak is grouped with rose, indicating that these representations are more similar to one another than is oak to pine. By Epoch 150, it is apparent that the hierarchical relations among the concepts is fully captured in the similarities among the learned distributed representations.

To better visualize the process of conceptual differentiation that takes place in this model, we performed a multidimensional scaling of the internal representations for all items at 10 different points during training. The solution is plotted in Figure 4. The lines trace the trajectory of each item's representation throughout learning in the two-dimensional compression of the representation state space. The labeled end points of the lines indicate the final learned internal representations after 1,500 epochs of training. The figure shows that the items, which initially are bunched together in the middle of the space, first divide into two global clusters based on animacy (plant/animal). Next, the global categories split into smaller intermediate clusters, and finally the individual items are pulled apart. In short, the network's representations appear to differentiate in relatively discrete stages, completing differentiation of the most general level before progressing to successively more fine-grained levels. Like children, the model seems to distinguish fairly broad semantic distinctions prior to more specific ones. What accounts for this stagelike progressive differentiation?

To understand this, first consider how the network learns about the following four objects: the oak, the pine, the daisy, and the salmon. Early in learning, when the weights are small and random, all of these inputs produce a similar pattern of activity throughout the network. Since oaks and pines share many output properties, their similar patterns produce similar error signals for the two items, causing the weights leaving the *oak* and *pine* units to move in similar directions. Because the salmon shares few properties with the oak and pine, the same initial pattern of output activations produces a different error signal, and the weights leaving the *salmon* input unit move in a different direction. What about the daisy? It shares more properties with the oak and the pine than it does with the salmon or any of the other animals, and so its weights tend to move in a similar direction as the other plants. Similarly, the *rose* representation tends to be pushed in the same direction as all of the other plants, and the other animal representations tend to be pushed in the same direction as the salmon. As a consequence, on the next pass, the pattern of activity across the representation units will remain similar for all the plants, but will tend to differ between the plants and the animals.

This explanation captures part of what is going on but does not fully explain why there is such a strong tendency to learn the superordinate structure first. Why is it that so little intermediate level information is acquired until after the superordinate level information? Put another way, why don't the points in similarity space for different items move in straight lines toward their final locations? Several factors appear to be at work, but one is key:

> Properties that covary coherently across items tend to move connections coherently in the same direction, while idiosyncratic variation of properties tends to move weights in opposing directions that cancel each other out.

To see this, consider the fact that the animals all share some properties (e.g., they all can move, they all have skin, they are all called animals). Early in training, all the animals have essentially the same representation. Consequently, any weight change forward from the representation units that are made when processing an individual animal (say, the canary) will produce a similar effect on all of the other animals. For properties shared by animals, this generalization speeds learning: When taught that the canary can move, the network will tend

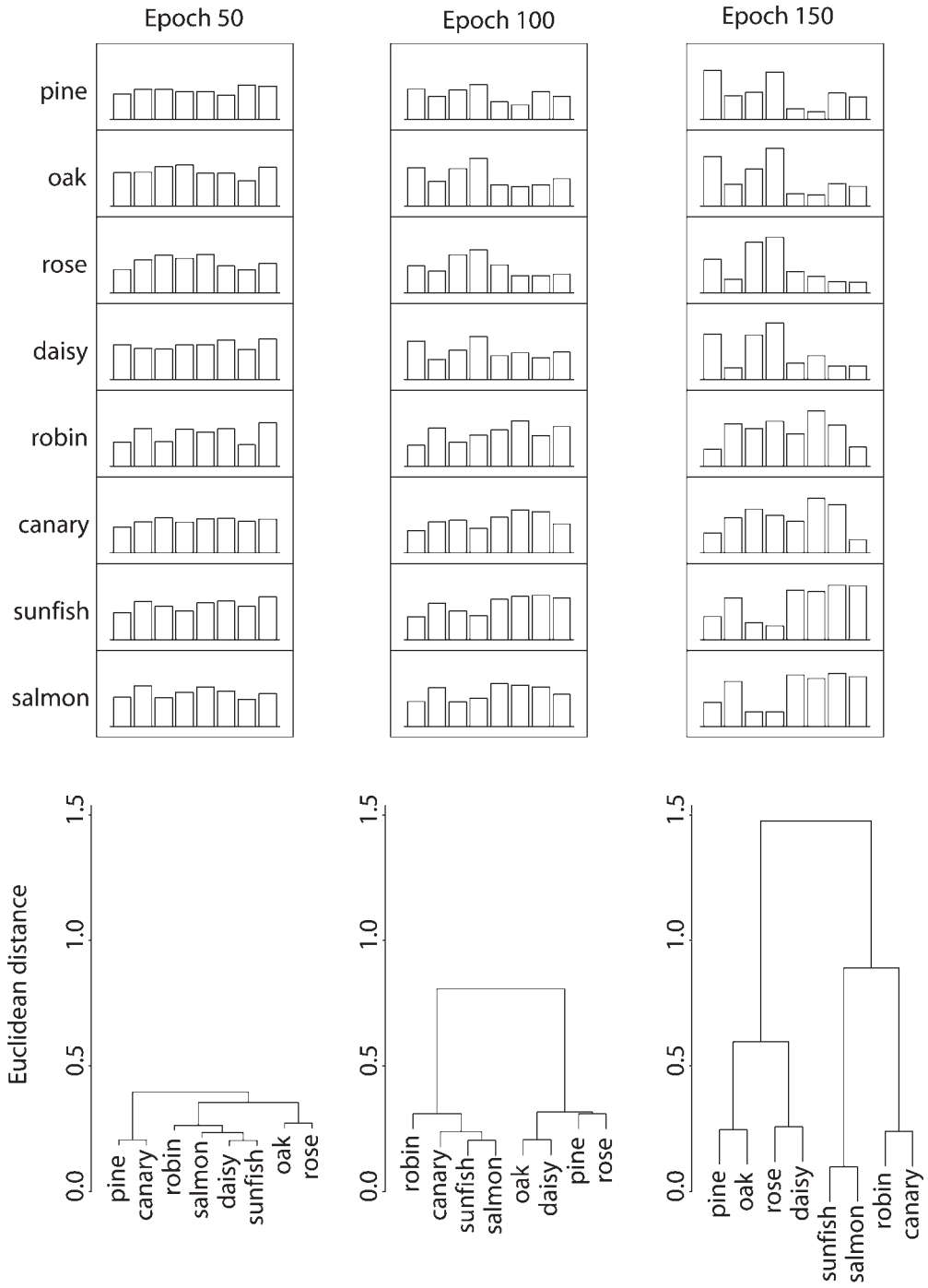

Figure 3. Learned internal representations of eight items at three points during learning, using the network shown in Figure 1. In the top plots, the height of each vertical bar indicates the degree of activation for one of the eight units in the network's *Representation* layer, in response to the activation of a single *Item* unit in the model's input. In the bottom plots, the same data were subjected to a hierarchical cluster analysis that recursively links a pattern or a previously linked group of patterns to another pattern or previously formed group. The process begins with the pair that is most similar (according to a Euclidean distance metric), whose elements are then replaced by the mean of the two items. These steps are repeated until all items have been joined in a single superordinate group. The plots show that, early in learning (50 epochs), the pattern of activation across these units is similar for all eight objects. After 100 epochs of training, the plants are still similar to one another, but are distinct from the animals. By 150 epochs, further differentiation into trees and flowers is evident.

to correctly generalize the property to all animals. Thus, for shared properties, learning accumulates across individual animals, benefiting knowledge for all animals. For properties that differentiate individual animals, on the other hand, this generalization is detrimental to learning: weight changes that help the network learn, for instance, that the canary is yellow or can sing will tend to generalize to other animals. In this case the generalization is usually incorrect, so these weight changes will be reversed by the learning that results when other individual animals are processed. Thus, learning of individuating properties will not tend to accumulate across different examples.

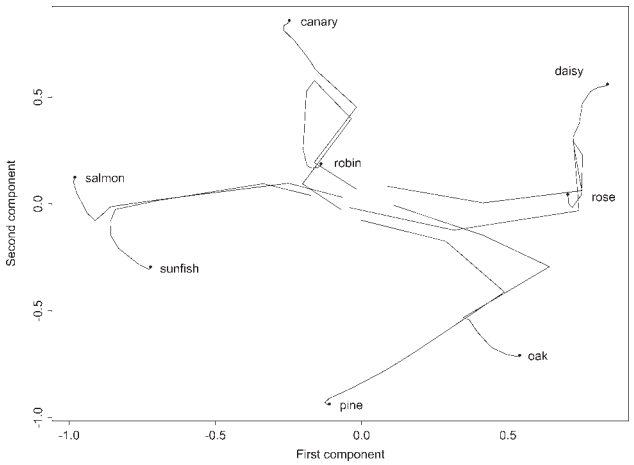

Figure 4. Trajectory of learned internal representations during learning. The Euclidean distance matrix for all item representations was calculated at 10 different points throughout training. A multidimensional scaling was performed on these data to find corresponding points in a two-dimensional space that preserve, as closely as possible, the pairwise distances among representations across training. Thus, the proximity of two points in the figure approximates the actual Euclidean distance between the network's internal representations of the corresponding objects at a particular point in training. The lines indicate the path traversed by a particular item representation over the course of development.

The consequence is that properties shared by items with similar representations will be learned faster than the properties that differentiate such items.

The preceding paragraph considers how the structure of internal representations affects learning in the weights projecting forward from the *Representation* layer. What about the weights projecting from the *Item* input to the *Representation* layer, which after all determine the similarity structure of the internal representations in the first place? We have seen that items with similar outputs will have their representations pushed in the same direction, whereas items with dissimilar outputs will have their representations pushed in different directions. The question remaining is why the dissimilarity between, say, the fish and the birds does not push the representations apart very much from the very beginning. The key to this question lies in understanding that the magnitude of the changes made to the representation weights depends on the extent to which such changes will reduce error at the output. This in turn depends on the configuration of the weights projecting forward from the *Representation* layer. If, given a particular configuration of forward weights, changes to the activation of *Representation* units will not strongly influence the total error at the output level, then the weights projecting into the *Representation* layer will not change. In other words, we can point out a further very important aspect of the way the model learns:

> Error back-propagates much more strongly through weights that are already structured to perform useful forward-mappings.

We can illustrate this by observing the error signal propagated back to the representation units for the canary item, from three different kinds of output units: those that reliably discriminate plants from animals (such as *can move* and *has roots*), those that reliably discriminate birds from fish (such as *can fly* and *has gills*), and those that differentiate the canary from the robin (such as *is red* and *can sing*). In Figure 5, we show the mean error reaching the *Representation* layer throughout training, across each of these types of output unit when the model is given the canary (middle plot) as input. We graph this alongside measures of the distance between the two bird representations, between the birds and the fish, and between the animals and the plants (bottom plot); and also alongside of measures of activation of the output units for *can sing*, *is yellow*, *has wings*, and *can move* (top plot). We can see that there comes a point at which the network is beginning to differentiate the plant and the animal representations, and is beginning to activate *move* correctly for all of the animals. At this time the average error information from output properties like *can move* is producing a much stronger signal than the average error information from properties like *has wings*, *can sing* or *is yellow*. As a consequence, the information that the canary can move is contributing much more strongly to changing the representation weights than is the information that the canary has wings and can sing. Put differently, the knowledge that the canary can move is more "important" for determining how it should be represented than the information that it has wings and can sing, at this stage of learning. Subsequently, the properties that differentiate birds from fish (e.g., has wings) begin to be learned, and to contribute to representational change, so that bird and fish representations are propelled apart; and finally the properties that discriminate subcategories (e.g., canary and robin) are learned and begin to influence representations. Note that these effects are not a simple consequence of the overall frequency of the various properties: *is yellow* (which is true of three items in the corpus) is actually more frequent than *has wings* (which is true only of the two birds); nevertheless the network learns to activate *has wings* first, because this property coheres with other properties that reliably discriminate birds from fish, whereas *is yellow* does not.

The overall situation can be summarized as follows. Initially the network assigns virtually the same representation to all of the items. With just this one representation, the network cannot predict different outputs for different concepts. The only properties that are correctly activated are those that are shared across everything – the *is living*, *can grow*, and *ISA living thing* outputs. All other output properties have their effects on the forward weights almost completely canceled out. However, because the plants have several properties that none of the animals have, and vice versa, weak error signals from each of these properties begin to accumulate, eventually driving the representations of plants and animals apart. At this point, the common animal representation can begin to drive the activation of outputs shared by animals, and vice versa for the plants. This structure in the forward weights in turn allows the properties shared by animals and not plants (and vice versa) to more strongly influence the model's internal representations, relative to properties that differentiate, say, birds from fish. The result is that the individual animal representations stay similar to one another, and are rapidly propelled away

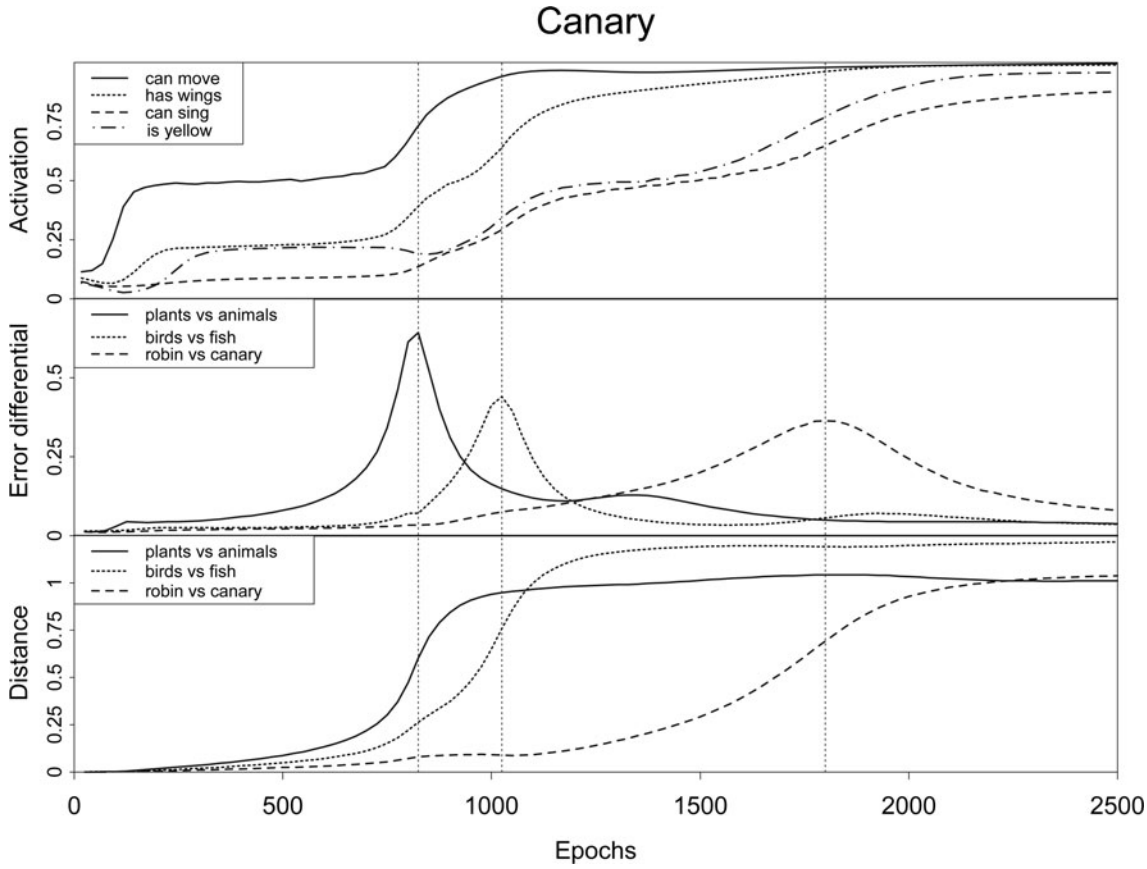

Figure 5. **Bottom**: Mean Euclidean distance between plants and animals, birds and fish, and canary and robin internal representations throughout training. **Middle**: Average magnitude of the error signal propagating back from properties that reliably discriminate plants from animals, birds from fish, or the canary from the robin, at different points throughout training when the model is presented with the canary as input. **Top**: Activation of a property shared by animals (*can move*) or birds (*can fly*), or unique to the canary (*can sing*), when the model is presented with the input canary can at different points throughout training.

from the individual plant representations. Very gradually, however, the weak signals back-propagated from properties that reliably discriminate birds from fish begin to accumulate, and cause the representations of these subgroups to differentiate slightly, thereby providing a basis for exploiting this coherent covariation in the forward weights. This process continues through successive waves of differentiation all the way down to the subordinate level, so that idiosyncratic properties of individual items are eventually mastered by the net.

In short, there is a kind of symbiosis of the weights into and out of the representation units, such that both sets are sensitive to successive waves of higher-order or coherent covariation among output properties. Each wave begins and peaks at a different time, with the peaks occurring at times that depend on the strengths of the corresponding patterns of covariation. The timing of different waves of differentiation, and the particular groupings of internal representations that result, are governed by high-order patterns of property covariation (corresponding to the eigenvectors of the property covariance matrix; see *Semantic Cognition*, pp. 96–104). Stronger patterns will drive differentiation earlier than weaker patterns; and the properties that differentiate very broad categories tend to exhibit stronger patterns of coherent covariation than those that differentiate more specific categories.

2.2. Category coherence

"Coherence" is a term introduced by Murphy and Medin (1985) to capture the observation that, of the many ways of grouping individual items in the environment together, some groupings seem more natural, intuitive, and useful for the purposes of inference than others. For example, objects that share feathers, wings, hollow bones, and the ability to fly seem to "hang together" in a natural grouping – it seems appropriate to refer to items in this set with a single name ("bird"), and to use the grouping as a basis for knowledge generalization. By contrast, other groupings of objects are less intuitive, and less useful for purposes of inductive inference. For example, the set of objects that are blue prior to the year 2010 and green afterward constitutes a perfectly well-defined class, but it doesn't seem to be a particularly useful, natural, or intuitive grouping. The second issue we consider is: How does the semantic system "know" which groupings should support productive generalization, and which should not?

The commonsense answer to this question is that the semantic system construes as similar the groupings of items that have many properties in common. Murphy and Medin (1985) argued, however, that similarity alone is too underconstrained to provide a solution to this problem. They emphasized two general difficulties with the notion that category coherence can be explained

solely on the basis of the learned similarities among groups of items. First, the extent to which any two objects are construed as similar to one another depends upon how their properties are weighted: A zebra and a barber pole may be construed as very similar to one another if the property *has stripes* is given sufficient weight. In order for a similarity-based account of category coherence to carry any authority, it must explain how some attributes of objects come to be construed as important for the object's representation, while others do not. Moreover, as R. Gelman and Williams (1998) have pointed out, the challenge is not simply to derive a set of feature weightings appropriate to all objects, because the importance of a given attribute can vary across different types of items. This observation leads to an apparent circularity under some perspectives: A given object cannot be categorized until an appropriate set of feature weights has been determined, but such a set cannot be recovered until the item has been categorized.

R. Gelman and Williams (1998), Murphy and Medin (1985), Keil (1989), and others (Gopnik & Meltzoff 1997; Gopnik & Wellman 1994; Wellman & Gelman 1997) have suggested that the challenge of selecting and weighting features appropriately might be resolved with reference to naive theories about the causal relationships among object properties. That is, certain constellations of properties "hang together" in psychologically natural ways, and are construed as "important" to an object's representation, when they are related to one another in a causal theory. For example, wings, feathers, and hollow bones may be particularly important for representing birds, because they are causally related to one another in a person's naive theory of flight. On this view, causal domain theories constrain the range of an item's attributes that are relevant to the task.

The second argument against correlation-based learning accounts of coherence stems from the observation that knowledge about object–property correspondences is not acquired with equal facility for all properties. For example, Keil (1991a), initially paraphrasing Boyd (1986), wrote that

> although most properties in the world may be ultimately connectable through an elaborate causal chain to almost all others, these causal links are not distributed in equal density among all properties. On the contrary, they tend to cluster in tight bundles separated by relatively empty spaces. What makes them cluster is a homeostatic mechanism wherein the presence of each of several features tends to support the presence of several others in the same cluster and not so much in other clusters. Thus, the properties tend to mutually support each other in a highly interactive manner. To return to an example used previously, feathers, wings, flight, and light weight don't just co-occur; they all tend to mutually support the presence of each other, and, by doing so, segregate the set of things known as birds into a natural kind.
>
> Boyd's claim is about natural kinds and what they are, not about psychology. At the psychological level, however, we may be especially sensitive to picking up many of these sorts of homeostatic causal clusters such that beliefs about those causal relations provide an especially powerful cognitive "glue," making features cohere and be easier to remember and induce later on. (Keil 1991a, p. 243)

The progressive differentiation process just illustrated suggests some answers to the important issues raised by Keil (1991a), Murphy and Medin (1985) and others. To make these answers explicit, we considered how a variant of the Rumelhart model would learn about items described by *is*, *can*, and *has* properties (as before), with some properties co-occurring together in coherent clusters and others distributed independently. The specific patterns are shown in Figure 6. Each of the items

	is A	can B	has C	is D	can E	has F	is G	can H	has I	is J	can K	has L	is a	can b	has c	is d	can e	has f	is g	can h	has i	is j	can k	has l
1	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0
2	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1	0	0	0
3	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0
4	1	1	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1
5	0	0	0	1	1	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	1	0	0
6	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0
7	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1
8	0	0	0	1	1	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0
9	0	0	0	0	0	0	1	1	1	0	0	0	0	0	1	0	0	0	1	1	0	0	0	0
10	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0
11	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	1	1	1	0
12	0	0	0	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	1
13	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	1	1	0	0	0	0	0	0	1
14	0	0	0	0	0	0	0	0	0	1	1	1	0	0	0	0	0	0	1	1	1	0	0	0
15	0	0	0	0	0	0	0	0	0	1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
16	0	0	0	0	0	0	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	1	1	0

Figure 6. Training patterns for the model (excluding names) in the simulation of category coherence. Individual item patterns are labeled 1–16, and the different properties are labeled with letters. Properties on the left (labeled with uppercase letters) are "coherent," in that they always occur together. Properties on the right (labeled with lowercase letters) are not coherent, because they do not co-occur reliably with one another. Every instance has three coherent and three incoherent properties, and every property appears as a target for four items.

(numbered 1–16 in the figure) was assigned six properties, and each attribute appeared as a target for four items. Hence, all properties were equally frequent in the model's training environment, and all items had an equivalent number of properties. As Figure 6 indicates, however, half of the properties are coherent in that they co-occur together in the same 4 objects, whereas others are incoherent, in that they vary independently of one another across items.

This structure provides an analog in the model to the coherent clusters of properties described by Keil (1991a) in the quotation above. In the real world, such clusters may arise from "homeostatic causal mechanisms," as Keil suggests; for the model, however, such homeostatic causal mechanisms are not directly accessible. What is accessible instead is the coherent covariation of properties across items and contexts produced by such mechanisms. We have assigned arbitrary labels to the items and the properties to avoid any sense that the actual properties are intended to be realistic, and to focus attention on the issue at hand, which is that of coherent covariation versus idiosyncratic distribution.

The top part of Figure 7 shows a hierarchical clustering of the model's internal representations at three points during learning. Since all properties occur in exactly 4 items, any individual property taken in isolation could, in theory, provide some basis for "grouping" a set of four items together in the model's internal representations – for example, considering just the *is-d* property, the model might have reason to "group together" items 3, 6, 10, and 13. From Figure 7, however, it is clear that the model discovers representations that are organized primarily by the coherent properties. The network represents as similar those items that have coherent properties in common (such as items 1–4); and it represents other groups of four that happen to share an incoherent property (such as *is-d*) as different from one another. The reason is exactly that explored in section 2.1: Because the items that share property A also happen to share properties B and C, the error signals generated by all of these properties push the representations of all of these concepts coherently in the same direction. Attributes that vary coherently together will exert a greater degree of constraint on the model's internal representations.

As a consequence, such properties will also be easier for the network to acquire. In the bottom part of Figure 7, we plot the activation of each item's six attributes (when queried with the appropriate relation) throughout training, averaged across five different training runs. Coherent properties are shown as solid lines, and incoherent properties are shown as dashed lines. The model learns very quickly to strongly activate the coherent properties for all 16 items, but it takes much longer to activate each item's incoherent properties. Because all units were active as targets equally often, and all items appeared in the training environment with equal frequency, this difference is not attributable to the simple frequency with which items or properties appear in the environment. The network is sensitive to the coherent structure of the environment apparent in the way that attributes are distributed across items; it shows an advantage for learning and activating an item's "coherent" attributes. That is, the model is especially sensitive to the sorts of "homeostatic causal clusters" to which Keil (1991a) suggests humans may also be especially sensitive.

2.3. Illusory correlations

Children and adults can sometimes be shown to attest to beliefs that directly contradict their own experience. For instance, when shown a photograph of an echidna – a furry-looking animal with eyes but no discernible feet – children may assert that the animal can move "because it has feet," even though, when asked, they agree that there are no feet to be seen in the photograph. Or conversely, when shown a stone statue of a humanoid being, they may attest that it cannot move "because it doesn't have any feet," even when the statue's "feet" are clearly visible (Massey & Gelman 1988).

Such illusory correlations are important because they appear to indicate some organizing force behind children's inferences that goes beyond "mere" associative learning. That is, such phenomena appear to indicate a commitment to beliefs that contradict direct perceptual experience – and so, whatever mechanism supports the belief, it must be built upon something other than learning from direct perceptual experience. Perhaps the child holds an implicit theory of biological motion under which "having feet" is precisely the quality that causes the ability to move under one's own power. Such a theory might then be used to infer that any new animal, because it can move, must have feet, even if you can't see them; and that any new artifact, because it cannot move, must not have feet, notwithstanding appearances to the contrary. Under this view, a child's implicit theoretical commitments leads her to ignore or discount object–property correspondences not suited to the theory, or to enhance or even invent such correspondences, even when they are not present in actual experience. Illusory correlations are thus sometimes taken as evidence for the role of implicit causal theories in conceptual knowledge (Keil 1989; Murphy & Medin 1985).

Our simulations offer a different explanation: Perhaps illusory correlations arise as a by-product of sensitivity to coherent covariation. That is, perhaps children strongly infer that the echidna must have feet, appearances to the contrary, because they observe that it has fur and eyes, and these properties strongly tend to co-occur with feet in other animals. To illustrate how this could be, we trained the model with a variant of the original Rumelhart corpus, which we extended to include four items in each of the previous categories (flowers, trees, birds and fish), as well as a set of 5 four-legged animals (a dog, cat, mouse, goat and pig). The specific patterns (see *Semantic Cognition*, Appendix B) were not intended to accurately capture all of the actual properties of the corresponding items; we employed this extended corpus simply because the original training set was a bit too simple to address all of the phenomena of interest. The extended corpus adheres to the similarity structure from the original corpus: Items from the same intermediate category (e.g., fish, flower) tend to have many properties in common; items from the same broad domain (plant or animal) tend to have more properties in common with one another than with items from the contrasting domain. The slightly larger training set allows us to examine what happens with individual items that diverge slightly from

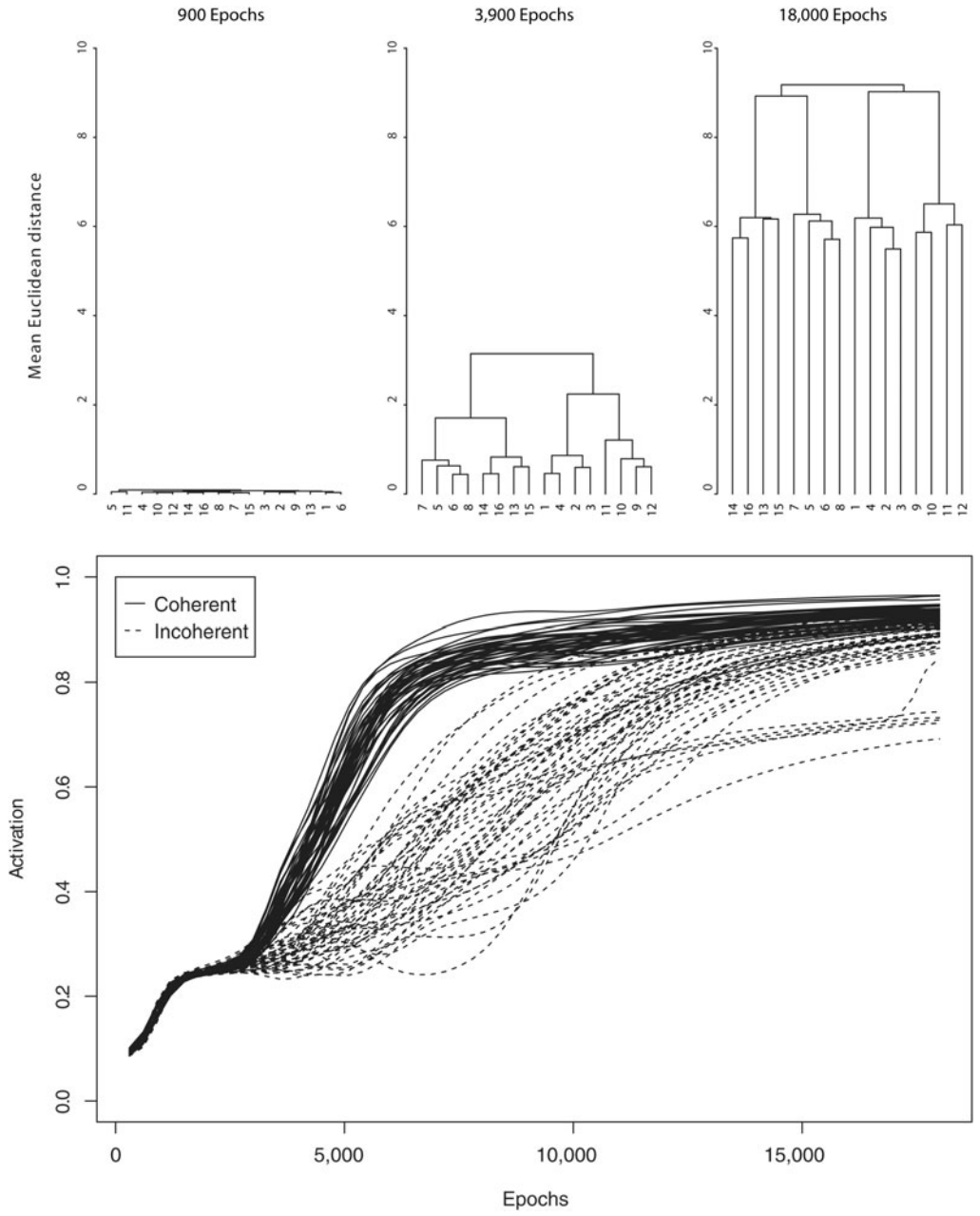

Figure 7. **Top:** Hierarchical cluster analysis of the model's internal representations at three points during learning. Each item is represented with its corresponding number as shown in Figure 6. Although every property in the training corpus is shared by some grouping of four items, the model organizes its internal representations with respect to shared "coherent" properties. **Bottom:** Activation of the correct output units for all 16 items when the network is queried with the corresponding item and context. Coherent properties are shown as solid lines, and incoherent properties are shown as dashed lines. The network quickly learns to activate the all of the coherent properties for all of the items, but takes much longer to learn the incoherent properties. Both plots show data averaged over five separate training runs.

a pattern of coherent covariation among members of a given category.

We investigated the model's responses to two queries throughout learning. First, we considered its activation of the property *has leaves* in response to the item *pine. Has leaves* is a property that covaries coherently with other properties of plants; it is not, however, true of the pine. Second, we investigated its activation of the property *can sing* when queried with the item *canary.* The canary is the only bird (and indeed, the only animal) that can sing in this corpus, so *can sing* represents a relatively idiosyncratic property. Figure 8 shows the activation of the *has leaves* unit and the *can sing* unit when the network is probed with the inputs *pine has* and *canary can*, respectively, at different points throughout training. At Epoch 1,500, the network has been trained repeatedly to turn off the *has leaves* unit when presented with *pine* has as input. Nevertheless, it strongly activates the *has leaves* unit in response to this input. Like the children in R. Gelman & Williams' (1998) study, the network attributes to the object a property that, on the basis of its experience, it clearly doesn't have. Similarly, by Epoch 1,500 the network has

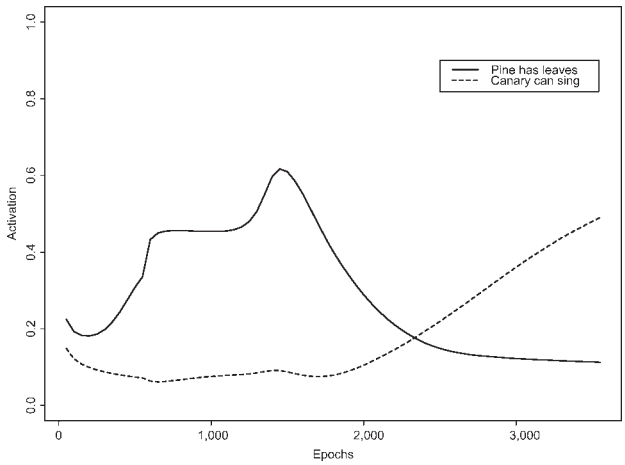

Figure 8. The activation of the *has leaves* and *can sing* output units across the first 5,000 epochs of training, when the network is probed with the inputs *pine has* and *canary can*, respectively. At epoch 1,500, the network has been trained 150 times to turn off the *has leaves* unit in response to the input *pine has;* and to turn on the unit *can sing* in response to the input *canary can.* Despite this, the network still activates the *has leaves* unit for the pine tree, and fails to activate the *can sing* unit for the canary.

repeatedly been "told" that the canary can sing. Despite this, it shows no tendency to activate the output *can sing* when asked what a canary can do. That is, the network appears to create an illusory correlation between the pine and the property *has leaves* that does not exist in its environment, and to ignore the strong correlation that does exist between the canary and the property *can sing.*

The simulation thus demonstrates that "illusory correlations" can arise from a domain-general correlational learning mechanism that is sensitive to coherent covariation among object properties – the higher-order patterns of covariation may overwhelm learning of weaker pairwise object–property correspondences that violate the higher-order regularities.

2.4. Domain-specific attribute weighting

For many theorists (Carey 1985; R. Gelman & Williams 1998; Keil 1991a; Murphy & Medin 1985), a key motivation for the claim that concepts are rooted in naive domain theories stems from the observation that children at fairly young ages can use quite different kinds of information to govern induction for items from different conceptual domains. In one of many experiments demonstrating such effects, Macario (1991) presented children with novel objects varying along two dimensions (color and shape). When the children were led to believe the objects were a kind of food, they most often generalized a new fact about the items on the basis of shared color; but when led to believe they were a kind of toy, they more often generalized on the basis of shared shape. Thus, the children appeared to weight color more heavily than shape for food items, but shape more heavily than color for toys (see also Jones et al. 1991; Smith 2000). Such phenomena appear to indicate a paradox: To "categorize" an object, one must know which of its properties are important; but one cannot know which properties are important until one knows what kind of thing it is.

We have seen that sensitivity to coherent covariation leads the model to weight some properties more strongly than others. Can the same processes explain patterns of domain- or category-specific attribute weighting? To answer this question, we conducted a simulation designed to capture the pattern of data observed in Macario's experiment. To the training patterns employed in the previous simulation, we added four new properties: *is bright*, *is dull*, *is big*, and *is small.* We assigned these properties to the familiar objects in the network's environment (the plants and animals) in such a way that size, but 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, all the trees were big and all the flowers were small, but a given tree or flower could be either bright or dull; whereas all the birds were bright and all the fish were dull, though a given bird or fish could be either big or small. Of course, these attributions are not perfectly valid, but they allow us to illustrate how the network learns in domain-specific ways about attribute "importance." Does the learning process described above come to selectively weight size more than brightness for plants, and brightness more than size for animals?

We trained the model for 3,000 epochs, on all items and relations, at which point it had learned to correctly activate all output properties except for specific names and idiosyncratic properties above a threshold of 0.7. We then used a technique called *backpropagation-to-activation* to investigate how the model would represent various novel objects varying in their size, brightness, and other observable qualities represented by output units. In a recurrent model that included projections back from output properties to *Representation* units, such an item could be represented just by activating its observed properties and allowing this information to feed back to the *Representation* units. Backpropagation-to-activation allows us to accomplish a similar effect in a feed-forward model – for instance, we can investigate how the model would represent a novel item, given just the information that it "is a bird," or given more detailed information, for example that it "is large," "is bright," and "has roots." (Details regarding the technique are given on pp. 63–66 of *Semantic Cognition*.)

We assigned brightness and size attributes to four "novel" test items as shown in Table 2. In the first simulation run, we also assigned to these items an attribute shared by the plants (*has roots*); in the second, we assigned to them an attribute shared by animals (*has skin*). In both

Table 2. *Distribution of attributes across four test objects in the simulation of category-specific attribute weighting*

	bright	**dull**	**big**	**small**
Object 1	1	0	1	0
Object 2	1	0	0	1
Object 3	0	1	1	0
Object 4	0	1	0	1

runs, we used backpropagation-to-activation to derive an internal representation for each item, by backpropagating from the output units corresponding to *bright*, *dull*, *big*, *small*, *roots*, and *skin*. We then examined the similarities among the four test item representations in each case.

Figure 9 shows the results of a hierarchical cluster analysis on the network's internal representations of the four test objects, when they share a property common to plants (left-hand figure) or animals (right-hand figure). When the network is "told" that the objects all have roots like the plants, it groups them on the basis of their size; when "told" that they all have skin like the animals, it groups them on the basis of their brightness. That is, the network seems to "know" that brightness is more important than size for representing animals, but that the reverse is true for plants. Like the children in Macario's (1991) experiment, it represents different similarities among a group of items, and consequently it will generalize from one to another differently, depending upon the superordinate category to which the items belong.

To understand why this happens, consider how the network comes to represent an object that is bright and big, compared to one that is bright and small. When the objects both share a property with the plants, such as *has roots*, the network must assign to them representations that lie somewhere within the space spanned by the predicate *has roots*. Within this region, the only objects that are big are the trees, which exhibit coherent covariation of several other properties; whereas the only objects that are small are the flowers, which have their own set of coherent properties, different from those of the trees. Thus, the *bright-big* test object will receive a representation similar to the trees, whereas the *bright-small* objects will receive a representation similar to the flowers. The property *is bright* does not vary coherently with other properties within the plant domain, and, as a consequence, exerts little influence on representations among the plants.

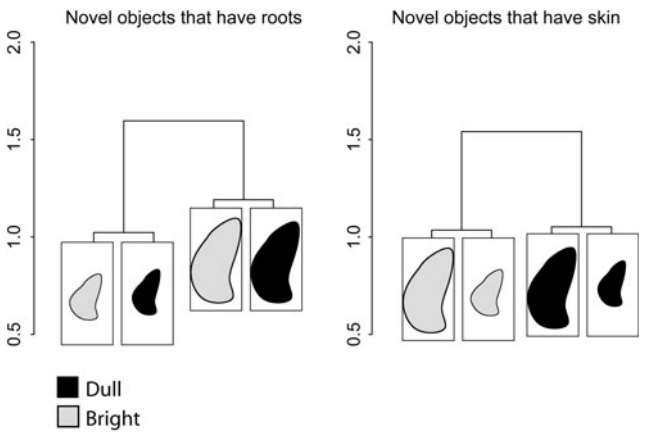

Figure 9. Hierarchical cluster analysis of the model's representations of test objects varying in brightness and size, and sharing a property common either to all animals or to all plants. When the objects share a property common to the plants (*has roots*), the network groups them on the basis of their size, which is important for discriminating flowers from trees. However, when the objects share a property common to animals (*has skin*), the network groups them on the basis of their brightness, which is important for discriminating birds from fish in the network. Thus, the network has learned that brightness is "important" for animals, but not for plants.

The opposite consequence is observed when the same test objects share a property with animals. In this case, they must receive representations that fall within the region of semantic space spanned by the predicate *has skin*. Within this subspace, all the fish are dull, and all the birds are bright. In order to activate the property *is bright*, both objects must be represented as similar to the birds. The property *is big* does not vary coherently with other properties in this domain. Thus, both big and small objects fall within the same small region of semantic space (i.e., proximal to the other birds) and hence are represented as similar to one another. What we see here is that domain-specific constraints on attribute weighting do not require pre-existing knowledge about which properties are important for which conceptual domain. Such constraints can be learned, and there is no chicken-and-egg problem – category-specific attribute weighting can be explained by the sensitivity of a domain-general learning mechanism to patterns of high-order covariation among stimulus properties.

2.5. Induction and conceptual change

An important source of information on the development of conceptual knowledge comes from studies of inductive projection, where children at different ages are asked to answer questions about the properties of novel and familiar objects. In some cases, they may be taught a new fact about an item (e.g., "this dinosaur has warm blood"), and then asked whether the fact is true about other kinds of objects (e.g., "Do you think this other kind of dinosaur also has warm blood?"). In other cases, they may simply be asked about properties of presumably unfamiliar things (e.g., previously unfamiliar animals), or about properties of things that may be somewhat familiar but where it is unlikely they have learned directly about the property in question (e.g., "Do you think a worm has a heart?"). In a series of influential experiments, Carey (1985) showed that children's answers to such questions change in systematic ways over development. Since generalization and induction are key functions of the semantic system, these patterns provide an important source of information about developmental change in the structure of semantic representations.

Carey (1985) used such changing induction profiles in an effort to diagnose developing children's causal theories. She suggested that a concept like *living thing* is rooted in an emergent theory of biology, which is constituted in part of knowledge about the causal mechanisms that give rise to the shared properties of living things. All living things breathe, eat, reproduce, grow, and die; Carey (1985) proposed that 10-year-olds (and adults) realize that all of these properties are consequences of the same underlying causal (biological) mechanisms. By contrast, she suggested, 4-year-olds conceive of these biological facts as arising from the same social and psychological mechanisms that also give rise to other various aspects of human behavior: Something might grow because it "gets the idea" from other things that grow, for example. The later-developing conception of animals and plants as both belonging to the same conceptual domain depends upon the acquisition of a theory of biological causation. Thus conceptual

reorganization – change over time in the way that concepts are organized – reflected, for Carey (1985), change to causal theories. And yet, although conceptual reorganization is so central to Carey's work, she has relatively little to say about the mechanisms that lead to change – indeed, in some subsequent writings, Carey and others seem to find it a mystery how theory change is even possible (Carey & Spelke 1994; Fodor 2000).

Here we consider some of Carey's findings on inductive projection in developing children between the ages of 4 and 10, and present simulations indicating how analogs of these patterns may be seen in the behavior of the Rumelhart model as it gradually learns from experience. We will not attempt to simulate the specific patterns of inductive projection seen by Carey and others; rather our focus will be on showing that the different types of changes that she points to as indicative of underlying theory change can be seen in the changing patterns of inductive projection, and in the underlying representations, within the model. These kinds of changes can be briefly enumerated as follows: (1) Patterns of inductive projection change over development; (2) they can differ for different kinds of properties; (3) such patterns tend to become more specific to the particular type of property over the course of development; and (4) patterns of inductive projection can coalesce as well as differentiate.

To understand how these patterns of reorganization might arise within the model, consider that the particular properties the model must activate in response to a given item depends upon the context in which the item is encountered. In the Rumelhart model, there are four different contexts, which require the model to generate an item's names (*ISA*), behaviors (*can*), parts (*has*), or other properties such as color (*is*). We have stressed up to now how knowledge of a concept evolves across the *Representation* units in the model. In this layer, a given item is always represented with the same pattern, regardless of the context in which the model is queried. The Rumelhart model does, however, provide for context-dependent representations on the *Hidden* layer, where information from the relational context units comes together with the context-independent representation on the *Representation* units. It is to these representations that our attention now turns.

When a new property is associated with a representation in the *Hidden* layer, the likelihood that it will also be activated by a different item will depend on the input from both the *Representation* and the *Relation* layers. Because different relational contexts emphasize different similarity relations, the model will come to generalize different kinds of features in different ways; and these patterns will themselves change over development, as the model gains increasing experience with each of the different contexts. (The range of contexts provided in the model is highly restricted, but should be sufficient to illustrate how context sensitivity can be achieved in the model). To explore how these factors influence the model's inductive projection behavior, we investigated its tendency to project different kinds of newly learned nonsense properties from one item to others, at two different points during training with the same corpus used in the previous section.

Specifically, we added a new output unit to the *Attribute* layer to represent a new nonsense property called *queem.* No occurrences of the novel property *queem* occurred during this overall training, which we take as providing the background developmental experience onto which a test of inductive projection can be introduced. To assess inductive projection in the model, we stopped training after 500 or 2,500 epochs of training with the corpus, and taught the network a new fact about the maple tree: either that the maple *can queem*, that the maple *has queem*, or that the maple *is queem.* We adjusted only the weights received by the new nonsense property from the *Hidden* layer, so that acquisition of the new fact was tied to the network's representation of the maple in the given relational context. (In the book we discuss how the same effect could be achieved by fast hippocampal learning of the type proposed by McClelland et al. 1995.) In each case, when the network had learned the new property, we queried it with the other items in its environment to determine how it would extend the new property *queem.*

The results are shown in Figure 10. Early in learning, the network generalizes the novel property from the maple to all of the plants, regardless of whether it is a *can*, *has*, or *is* property; there are slight differences in its handling of the *is* property compared to the others, in that it tends also to generalize to some degree to the animals as well. By Epoch 2,500, however, the model has learned a much stronger differentiation of the different contexts; the *can* property continues to generalize to all the plants, while the *has* property now generalizes only to the other trees. The *is* property also generalizes predominantly to the other plants, but not so evenly, and it generalizes to some extent to other things (with which the maple happens to share some superficial attributes). Thus, when the network has learned that the "maple is queem," it shows some tendency to generalize the novel property to items outside the superordinate category; it shows no such tendency when it has been taught that "queem" is a behavior (i.e., *can* property) or a part (i.e., *has* property).

The model behaves as if it "knows" that different kinds of properties extend across different sets of objects; and, just as in Carey's studies, this knowledge undergoes a developmental progression, such that the model only gradually sorts out that different kinds of properties should be extended in different ways. The reason is that, just as the network's internal representations of objects in the *Representation* layer adapt to the structure of the environment, so too do its context-sensitive representations over the *Hidden* layer. That is, the weights leading from the *Representation* and *Relation* layers into the *Hidden* layer adjust slowly, to capture the different aspects of similarity that exist between the objects in different contexts. Items that share many *can* properties generate similar patterns of activity across units in the *Hidden* layer when the *can* relation unit is activated. The same items, however, may generate quite different patterns across these units when one of the other *Relation* units is active in the input.

In Figure 11, we show a multidimensional scaling of the patterns of activity generated across the *Hidden* units, for the same 16 items in two different relation contexts, after the model has finished learning. (We excluded the mammal representations from this figure for clarity. They are distributed somewhere in between the birds and fish in all three plots.) The plot in the middle shows the learned similarities between item representations in the *Representation* layer. The top plot shows the

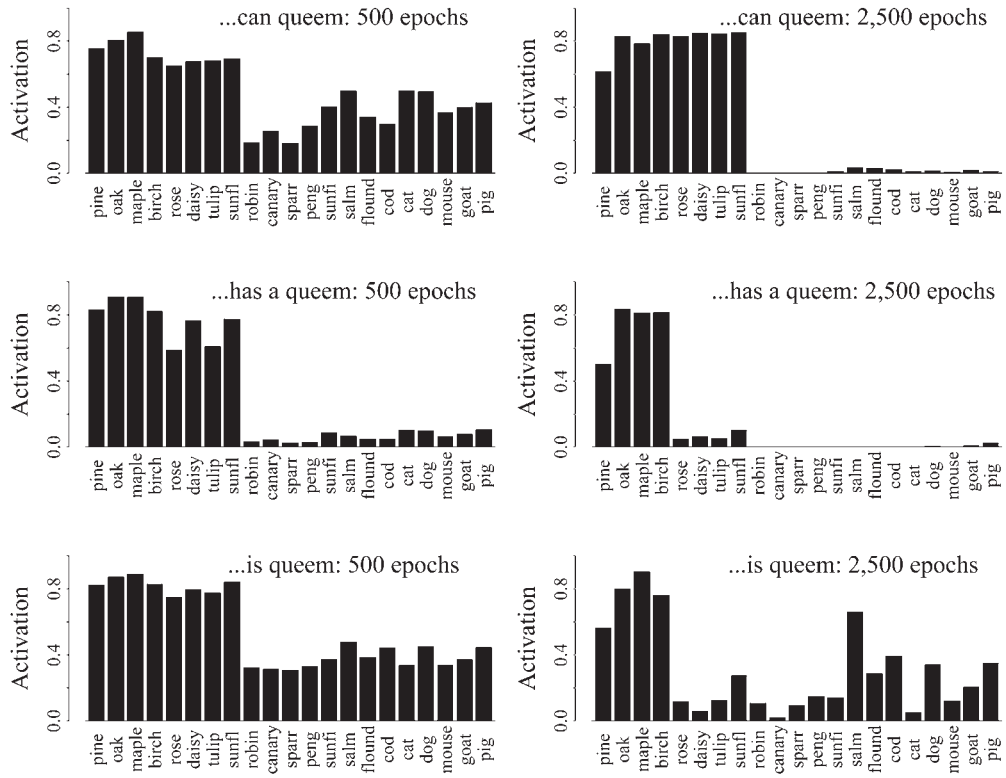

Figure 10. Barplot showing that activation of the nonsense property *queem* when the network is queried with various inputs, after it has learned that the maple *can queem*, *has a queem*, or *is queem.* If the network learns the new property after 500 epochs of training, the property generalizes across the entire superordinate category, regardless of the relation context. However, when the network is taught the novel property after 2,500 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.

similarities across *Hidden* units for the same items in the *is* context, whereas the bottom plot shows these similarities in the *can* context. In the *can* context, all the plants receive very similar representations, because they all have exactly the same set of behaviors in the training environment – the only thing a plant can do, as far as the model knows, is grow. As a consequence, the model generalizes new *can* properties from the maple to all of the plants. By contrast, in the *is* context, there are few properties shared among objects of the same kind. Thus, the network is pressured to differentiate items in this context, and as a result, it shows less of a tendency to generalize newly learned *is* properties. The other relation contexts not shown in the figure (*has* and *ISA*) also remap the similarity relations among the items in the model's environment to some extent. These representations are generally similar to those found in the *Representation* layer, since the similarity structure of the concepts within both the *has* and *ISA* contexts track fairly well the overall similarity structure. The similarity structure differs in subtle ways in each context, however, and these differences exert a subtle influence on the context-sensitive representations.

These changing induction profiles all involve learning to treat items differently in different situations or contexts, which is clearly an important part of the developmental progression charted in Carey's work. But Carey suggests that true conceptual change involves more than simply tailoring one's concepts to particular situations. Instead, the emergence of a concept such as *living thing*, which encompasses plants and animals and allows for induction across these items on the basis of knowledge about shared biological mechanisms, would seem to require a more deep-rooted restructuring of base concepts: Whereas younger children treat animals and plants as effectively unrelated for purposes of induction, by age 10 children seem to appreciate that all living things share certain core properties and are governed by common biological causal forces, so that the concept *living thing* begins to support induction for certain kinds of properties. This achievement thus indicates the coalescence of formerly unrelated concepts within a single conceptual domain.

Although we have seen that concepts may differentiate in the Rumelhart model, the processes we have discussed thus far would seem to preclude the possibility of coalescence with development. Moreover, Carey (1985) also suggested that other forms of conceptual change are commonly observed in development: Rather than reflecting proper subsets or supersets of earlier concepts, later-emerging concepts may entail a complete reorganization of earlier concepts.

These patterns of developmental change are not only consistent with the PDP (parallel distributed processing) framework, but in fact the explanation suggested by the framework shares much in common with Carey's (1985) own ideas about the forces that drive conceptual change in development. The key observation is that, although living things may have many properties in common (e.g., they all have DNA, they all breathe, they all grow and

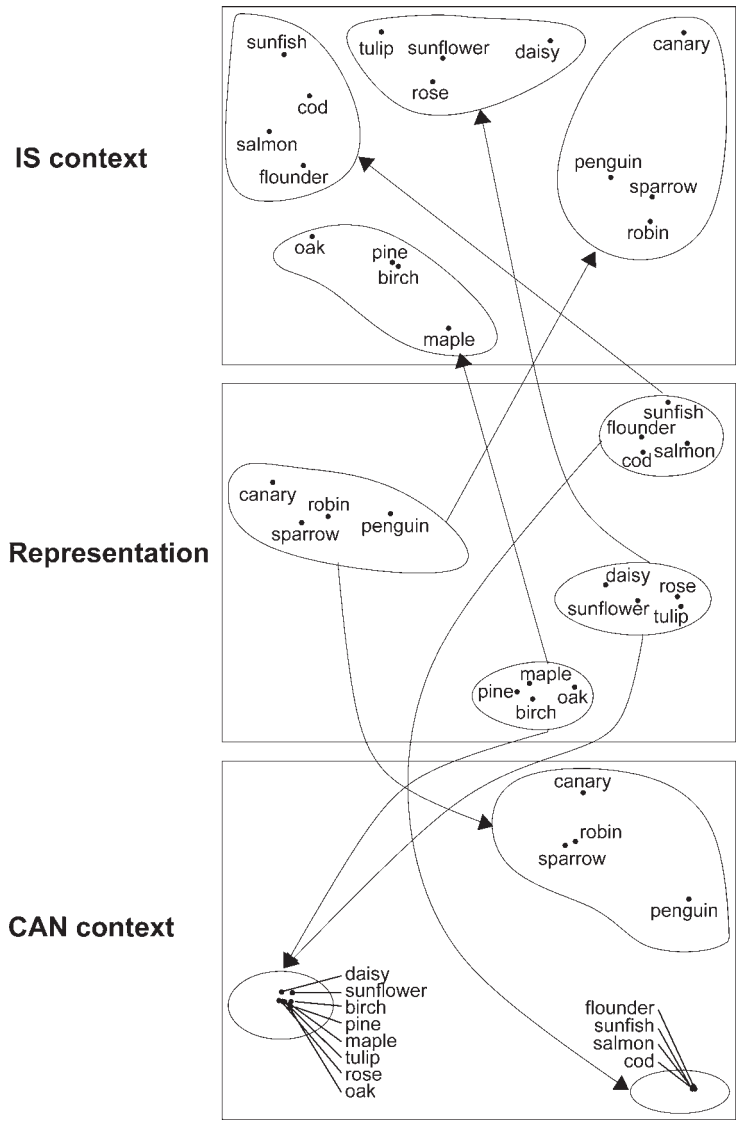

Figure 11. Multidimensional scaling showing the similarities represented by the model for objects in different relation contexts. The middle plot shows the similarities among object representations in the *Representation* layer. The top graph shows the similarities among the same objects in the *Hidden* 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.

die), many of these shared properties are non-obvious (S. A. Gelman & Wellman 1991). For example, animate objects may be considered members of the same class by virtue of sharing various internal organs, but these properties are not apparent in their outward appearance. By contrast, properties that are less diagnostic of an item's ontological status are more readily apparent in the environment. For example, an object's shape, color, texture, parts, and patterns of motion are apparent every time the object is encountered. Information about its insides, its metabolic functions, or other aspects of its behavior may be only sporadically available. Moreover, opportunities for acquiring this information likely change as the child develops; for example, children presumably acquire a great deal of non-obvious biological information when they attend school.

The account of conceptual reorganization consistent with these observations, then, is as follows: Early concepts are shaped by coherent covariation among the most frequently available object properties – outside, observable properties experienced whenever the object is encountered – but such properties may not adequately capture the "deep" structure organizing concepts like *living thing*. Other properties, such as the insides of objects and certain of their behaviors, are encountered less frequently and in fairly selective contexts; however, across contexts, such properties exhibit strong patterns of coherent covariation with one another and with some of the more frequently encountered surface properties. As children gain experience with these coherent-but-rare properties, sensitivity to coherent covariation drives such properties to become "more important" than the very frequent but incoherent

surface properties, leading to a reorganization of internal representations. This view appears to be very similar to Carey's notion that conceptual reorganization arises from the increasing assimilation of knowledge about non-obvious properties, but she provides no mechanism whereby such assimilation can actually lead to the relevant underlying change.

To make this account concrete, consider that, although different contexts evoke somewhat different similarity relations in the model's environment, there is also some important cross-domain structure. For example, the *has*, *can*, and the *ISA* (i.e., name) properties exhibit considerable coherent covariation: If an animal has wings and feathers, chances are good that it can fly and is called a "bird"; if it has scales and fins, it can likely swim and is called a "fish"; and so on. In contrast, the *is* properties (i.e., *is red*, *is yellow*, *is pretty*) are more idiosyncratically distributed – they are shared by items that otherwise have little in common. Let us consider the possibility that many of the coherently covarying properties are non-obvious; that is, they are only observed in specific contexts rather than each time the object is encountered, while the remaining "obvious" properties occur quite frequently in different contexts. The assumption appears plausible on the face of it: For instance, children experience what dogs look like on the outside every time they encounter a dog, but only learn about what the dog has on the inside in specialized and infrequent situations, such as science class.

What happens in a model analog of this situation, in which patterns of coherent covariation apparent across different specific contexts are reflected only weakly, if at all, in the information that is available every time a particular object is encountered? To investigate this question, we considered how the context-invariant representations over the *Representation* units in the network evolved under a training regime in which the *is* properties – which, as noted earlier, are distributed in a relatively arbitrary manner – were available every time an item was encountered, but the other properties were only available less frequently, contingent on a particular context. Specifically, the *is* properties were made a part of the target pattern for learning, regardless of which context unit was active in the input, while all other attributes remained contingent on the context. For example, when presented with *robin has* in the input, the model was given as the target for learning all of the *is* properties true of robins, as well as all of the *has* properties. Similarly, when presented with *robin can* in the input, the model was given all of the robin's *is* properties, as well as its *can* properties. As a result, the information coded in the *is* patterns was more frequent than the information coded in the other contexts; and the *is* information became independent of context, while the information associated with other contexts remained context-dependent. We trained the model with these patterns and examined the resulting internal representations (excluding the 5 mammal items simply to keep the cluster plots uncluttered) at different points during learning. We emphasize that there was no change over time in the training in this simulation; the regime described here remained constant throughout the entire training process. Such changes would, of course, contribute to reorganization of representations (see *Semantic Cognition*, pp. 283–88), but are not essential.

The results of this simulation are shown in Figure 12. After 500 epochs of training, the model has divided the

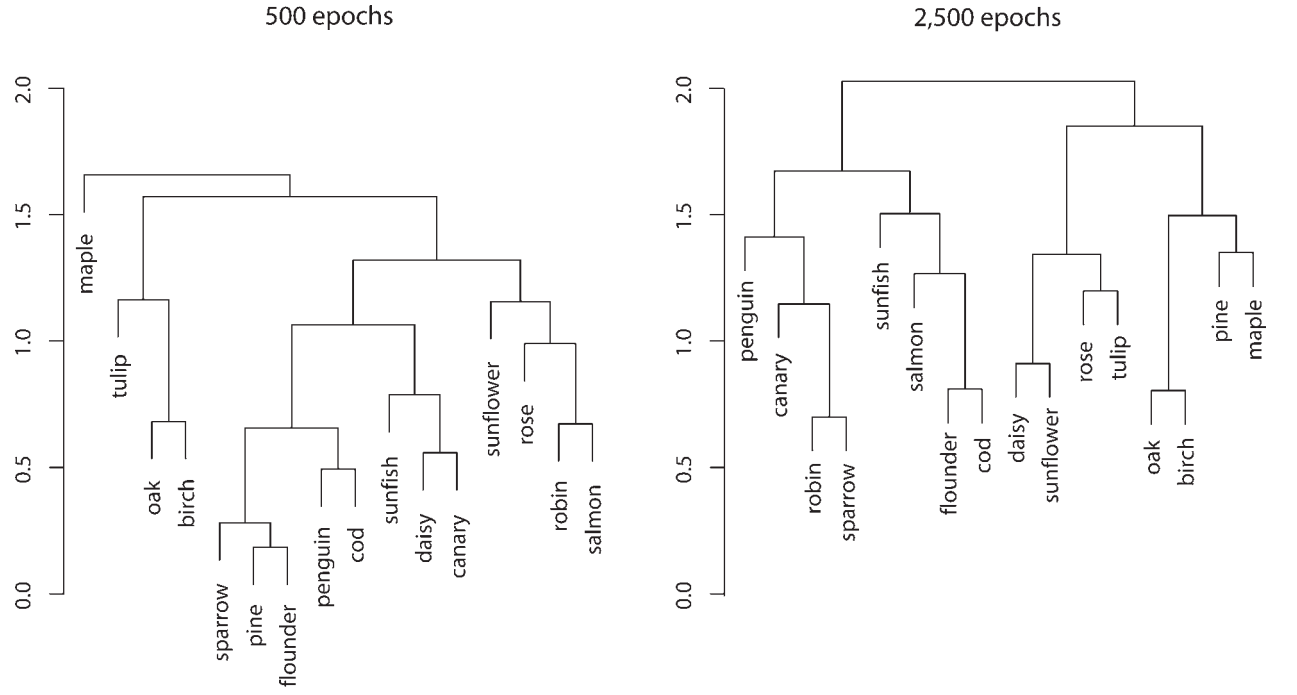

Figure 12. Hierarchical cluster plot of the model's internal representations in a simulation where the model was always required to activate *is* properties in every different context, so that such properties were both (a) more frequent and (b) less context-dependent. Earlier in learning, the model shows an organization of internal representations based largely on the more frequent *is* properties; later the internal representations have reorganized to capture the less frequent but more coherent structure apparent across the different contexts.

items into several small clusters that do not correspond well to global semantic categories. These clusters are organized largely, although not completely, by overlap of the superficial but frequent *is* properties: For example, the right-most cluster includes three red items (rose, robin, salmon) as well as the sunflower (likely because it *is pretty* like the rose and *big* like the salmon); the next cluster consists of yellow items (sunfish, daisy, canary); and so on. (In reality, there is some degree of coherent covariation of color with other object properties. The inconsistency with nature allows us to illustrate key properties of the workings of our model.)

Later in the model's development, the representations have reorganized to capture more fully the shared structure present across the other contexts. And, Figure 12 shows that both differentiation and coalescence can occur in the model: Clusters like the sunfish, daisy, and canary split apart to take their place in the later structure, and new groupings like the general clusters *plants* and *animals* coalesce later in learning. Hence, it is not the case that the later model's representations form a simple superset or subset of the earlier model's representations. Instead, the later model's internal representations find no obvious progenitor in the earlier model's representations.

In summary, the model provides two ways of understanding the changing induction profiles that, for Carey (1985), signaled underlying theory change. First, children may grow increasingly sensitive to the demands of particular situations or contexts, in which different properties and consequently different similarities are highlighted, so that items treated as similar for purposes of induction in some situations may be treated as quite different in others. Second, the "domain-general representations" – those that are acquired as a result of experience across many different contexts – are nevertheless influenced both by the frequency with which different kinds of information are encountered across different situations, and by the coherent covariation of properties across different contexts. Frequently encountered properties will strongly shape the first representations that emerge; but less frequently encountered properties can exert a strong influence on representational change later in learning, if these properties covary coherently with other properties observed in different situations. Thus, both the changing induction profiles observed in children's behavior, and the kind of representational change that Carey emphasizes as indicative of theory-change, can be understood as arising from the same domain-general learning mechanisms described earlier.

3. The importance of causal knowledge in semantic cognition

To this point, we have described simulations illustrating how the PDP theory can explain a range of phenomena motivating the view that conceptual knowledge is rooted in implicit domain theories. We have not yet addressed, however, three lines of evidence that most directly support the idea that causal knowledge contributes importantly to human semantic cognition. Here we illustrate how the PDP theory could be extended to encompass these phenomena; we then consider whether the theory is best considered an alternative to, or an instantiation of, the theory-theory.

3.1. Inductive inferences are constrained by knowledge of event sequences

First, several studies demonstrate that knowledge about the sequence of events through which an object comes to have its observed properties can influence how an adult or child conceives of the object (e.g., Ahn 1998; Ahn et al. 2002; Gopnik & Sobel 2000; Keil 1989). In Keil's "transformation" studies, for example, children were told stories about a raccoon that undergoes a series of interventions and ends up looking like a skunk (Keil 1989). Some children were told that the raccoon was wearing a skunk costume; others were told that it was dyed black and had a stripe painted down its back; still others were told that it received an injection when it was young that caused it to grow up looking and smelling like a skunk. After hearing the story, all children were shown a picture of a skunk and told "now the animal looks like this." When asked to decide if it was a raccoon or a skunk, the youngest children tended to choose skunk, regardless of which transformation story they had been told; but older children tended to choose skunk only in conditions where the mechanism of change could be construed as biological (for instance, when the raccoon was given an injection and "grew up into" a skunk). Thus, for older children, the decision as to whether the animal was "really" a raccoon or a skunk depended upon the causal mechanism by which it exhibited the visual properties of a skunk.

To understand how the PDP approach might be extended to address these issues, we rely upon a generalization of the Rumelhart model, illustrated in Figure 13. In the Rumelhart model, items co-occur together with contexts, and both are represented with static, externally applied patterns of activation across corresponding units. In contrast, the "contextual" information in the generalized model includes (1) other simultaneously present aspects of the situation, and (2) an internal representation of prior events leading up to the current input that can influence the current input's interpretation. We suggest that, just as the Rumelhart network can learn to generate different outputs for the same item depending upon the (static) context in which it is encountered, the generalized model should be able to generate different outputs for a given item depending upon the temporal context – the particular sequence of events that precedes its appearance.

We base this suggestion on previous studies of such recurrent network models. A key appeal of recurrent models is that, after learning, processing can be highly sensitive to temporal context: The response generated by a given input strongly depends upon the sequence of preceding inputs, as captured by a learned internal representation. Such models have been brought to bear on a broad range of phenomena relating to knowledge about sequential structure (e.g., Cleeremans 1993; Cleeremans & McClelland 1991; Elman 1990; 1991; Rohde & Plaut 1999), including models of language comprehension. For example, in St. John's (1992) work on story comprehension, if a named individual has been placed in the role of a waiter greeting and seating guests early in an event

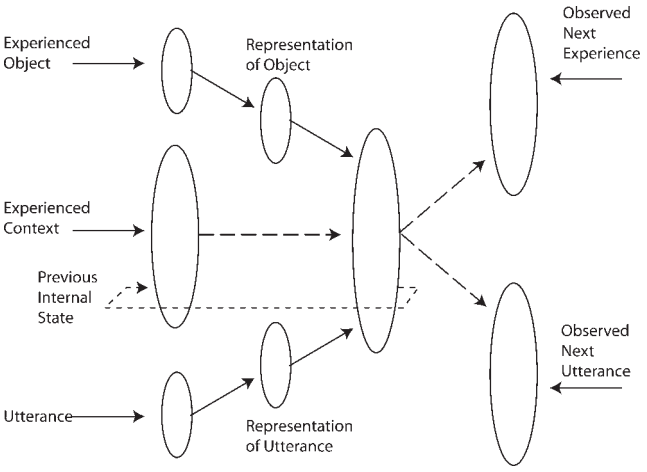

Figure 13. A sketch of a network architecture of the sort we envision will be necessary to capture the acquisition of causal knowledge from event sequences and the convergent use of verbal as well as other modalities of experience to jointly constrain the emergence of semantic knowledge. The diagram is intended to suggest a general correspondence with the Rumelhart network, in which a given item is encountered in a particular relational context, and potential completions of the event are to be predicted. Here we indicate how the contextual representation can be influenced by preceding internal representations (via a time-delayed connection indicated by a dotted line), so that predictions about the current input can vary depending upon the preceding sequence of events. The illustration also shows how verbal inputs and predictions can be interfaced with inputs and predictions from other modalities. The dashed arrows indicate projections that may include recurrent connections.

sequence characterizing a visit to a restaurant, then the model will expect this individual to be the one who brings the food and the check, and not to be the one who eats the food or pays for it.

Such studies suggest that a learning mechanism like the one we have sketched out could provide the basis for understanding phenomena like those documented by Keil and others. For instance, children are likely to have had considerable experience with event sequences involving costumes and disguises. Those of us who have been parents or caregivers to young children may recall how terrifying such costumes and disguises can be for children when they are very young, perhaps because at that point the children do not yet have an acquired appreciation that the costumes only create a temporary change in appearance. But after a child repeatedly witnesses and/or participates in various kinds of costume events, he or she apparently comes to appreciate that the visible surface properties of animate things can be strikingly but also reversibly affected, and that many other properties remain unchanged. A child can dress up as Dracula and his friend as E.T., or vice versa, but other sources of information will indicate that many of the costumed individual's properties are maintained throughout. Furthermore, both he and his friend will revert to their prior appearance when they take their costumes off. Through such experiences, we suggest, the child learns to maintain an internal representation of a costumed individual that retains the properties that the person had before putting on the costume, rather than the properties known to be possessed by the things they appear to be while they are wearing the costume.

In addition to this direct learning from experiences with individuals in costumes, we also suggest that verbal inputs in the form of statements that are made by others during costume-wearing events (e.g., statements like "That's just your friend Sally dressed up like E.T."), as well as movies or stories about costume wearing events, will contribute to the acquisition of knowledge about costumes. We don't suggest that children will need to have had experience specifically with raccoons in skunk costumes, but only that they will need to have had experience with other animate objects in costumes, because we would expect them to generalize across different types of animals, due to their having similar underlying representations. Similarly, children may not need to have direct experience with sequences in which specific animals are given injections in order to draw conclusions from the story in which the raccoon that received an injection "grew up into" a skunk. Perhaps they will think the raccoon is now "really" a skunk because many times animals transform naturally from one apparent "kind" to another as they grow up, the transformation of caterpillars to butterflies and tadpoles to frogs being two clear examples.

Of course, we understand that some readers may remain to be convinced that this kind of story about the influence of causal knowledge on semantic cognition could ever work in practice. Although we cannot allay all such concerns without extensive further work, we can point to an existing simulation that addresses issues related to those arising in the "costume" experiments reviewed earlier. The simulation in question addresses knowledge about the continued existence of objects even when they are out of view. When an object A moves in front of another object B, object B disappears from view – a situation analogous to that in which a costume C is put on by an individual D, so that the individual no longer looks like itself even though it actually remains the same inside. In the case of object permanence, we know that object B is "still there" despite appearances to the contrary; and in the case of a costume, we know that despite appearances it is still individual D standing in front of us, even though the costume replaces D's visible attributes with what might be a very different set of properties.

Munakata et al. (1997) demonstrated how a very simple recurrent network could learn to maintain representations of objects that are no longer visible from simple event sequences involving objects hidden by occluders. The essential element of the simulated event sequences was that objects hidden by the occluder became visible again when the occluder moved away. In order to correctly predict that this would occur, the network learned to maintain a representation of the object during the part of the event sequence when it was hidden by the occluder. Although, of course, costumes provide far more complex situations than this, this simulation illustrates the fundamental property required for a system to employ knowledge about an item's prior status in order to maintain a consistent internal representation of the item when it is subjected to certain transformations, rather than treating it as having been fundamentally transformed by the

alteration. We believe that similar processes may also underlie acquisition and use of knowledge about the consequences of more complicated transformations documented by Keil (1989) and others.

3.2. Children strongly weight inferred causal properties when generalizing newly learned names

In a very different series of studies, Gopnik and Sobel (2000) have shown that (1) children make inferences about the causal properties of novel items, and (2) they use these inferences to govern their decisions about how names should generalize. (We consider Gopnik and colleagues' more recent work on inferring causal properties later.) In the canonical paradigm, children are shown a machine called a "blicket detector." The blicket detector flashes and makes music when certain blocks (blickets) are placed on it; but nothing happens when other blocks (non-blickets) are placed on it. In early studies with this device, the authors showed that children would use the apparent causal potency of a given object, rather than its appearance, to decide whether it is a blicket or not. That is, shown a small yellow block that is called a blicket and activates the detector, and a tall red block that does not, children will then call another block a blicket if it activates the detector, regardless of its color or size, generalizing the name to other objects based on their causal powers, not on their color or shape. The experiment thus shows that children appear to lend special weight to "causal" properties in their inductive inferences.

We consider such phenomena to reflect the operation of mechanisms similar to those described in previous sections. The children in Gopnik and Sobel's experiment may not have had much experience with blocks that produce music when they are placed on certain boxes; but no doubt they have had experience with other kinds of objects with specific causal properties. Keys, switches, fasteners, bank cards, batteries, openers, remote controls, and many other objects in everyday use have specific causal powers of this type. And, such objects can vary considerably in shape or other aspects of their appearance, while remaining consistent in their causal potency. Batteries, for instance, come in many shapes, sizes, and colors, but have similar causal consequences. We suggest that people learn to represent such objects (and, indeed, all other types of objects) through exposure to event sequences in which they interact with other objects, with partially predictable consequences. Furthermore, we suggest that the words we use to refer to such objects covary more consistently with their causal properties than with surface attributes such as shape and size, with the consequence that these causal properties become more important in determining how such object's names will generalize.

3.3. The role of explanations in causal and other semantic tasks

The third source of evidence that children's concepts depend upon causal theories is simply that they can provide explicit explanations for their semantic judgments. In one study, Massey and Gelman (1988) showed children photographs of novel objects and, for each, asked them to decide whether it could move itself up and down a hill. After making their judgment, children were asked to explain them. Their responses seemed to the authors to reveal an underlying process of causal inference. For example, when a child says "it can move up and down the hill because it has feet," this indicates to Massey and Gelman that, in making their judgment, the child is consulting an underlying theory in which "having feet" is precisely the property that causes the ability to move autonomously. The models we have described may explain how the child is able generate the judgment itself, but how can they account for this introspective ability to explain the reasons for the judgment?

The difficulty with this argument is that the explanations people give for their own behavior are often at complete variance from the factors that demonstrably govern their responding (Nisbett & Wilson 1977). Indeed, people are remarkably poor at judging whether they are capable of explaining even very familiar causal scenarios (e.g., how a toilet works; see Wilson & Keil 2000). Such findings suggest that the explicit explanations people proffer for their own judgments do not necessarily shed light on the mechanisms that support the judgments; and this may be true even when there is a degree of concordance between the behavior and the explanation. That is, we do not believe that the overt explanations people produce provide much insight into the processes that support their semantic judgments.

We do accept that overt explanations constitute one of the various kinds of responses that people can learn to generate from a given situation; and we suggest that a shared intuitive sense of what "counts" as an explanation is one of the things that could be learned within a model like that shown in Figure 13 (see *Semantic Cognition*, Ch. 8.). On this view, explanations can shape, and can be shaped by, our internal semantic representations of witnessed events, just like other varieties of experience and behavior; however, the propositions that appear in overt explanations do not necessarily play a causal role in generating semantic judgments.

4. Contrasting the PDP and theory-based approaches

The variety of phenomena and the arguments emphasized by theory-based approaches demonstrate clearly that adults and even young children can be quite sophisticated in their semantic abilities, and we often find ourselves in agreement with some of the claims of theory-based approaches. For example, we agree with theory-theorists that "semantic knowledge" encompasses more than just list-like knowledge about the properties of objects – it includes knowledge about how objects interact with one another, how certain properties and situations give rise to other properties and situations, and so on. In Table 8.1 in *Semantic Cognition*, we enumerated several points of agreement between our position and theory-based approaches. In this section of the précis, however, we will attempt to bring out the key differences. We should note that we are contrasting our view with a theory-based approach that is more of a prototype than a specific theory held by any individual investigator. Several important contributors expressly do not endorse all of the properties we attribute to some version of the

theory approach. For instance, Gopnik (see Gopnik & Meltzoff 1997; Gopnik & Wellman 1994), a major proponent of theory-theory, considers the possibility that theory-like knowledge may be acquired using a domain-general mechanism, albeit one that may be especially attuned to the detection of causal relations (Gopnik et al. 2004). Also, Murphy (2002) eschews the theory approach in favor of what he calls the "knowledge approach," even though he was one of the early protagonists of theory-based approaches (Murphy & Medin 1985), and he expresses doubt about domain specificity, innateness of domain knowledge, and even that causal knowledge plays a special role in semantic cognition.

The first point of contrast lies in the question of whether the knowledge that underlies semantic task performance necessarily depends on initial (i.e., innate) principles that provide the seed or skeleton on which the development of semantic cognition depends. Many researchers in the theory-theory and related traditions appear to favor the view that some initial principles are necessary to serve as a base for further elaboration of conceptual knowledge. The argument for innateness, however, sometimes rests on little more than the suggestion that known learning procedures seem inadequate to explain the acquisition of the abilities children possess (Keil 1994). Even the very simple networks that we have employed can acquire domain-specific behaviors similar to those that putatively arise from naive domain theories. Thus, the observation of domain-specific behaviors in children provides little reason to infer innate domain-specific theories (or innate domain-specific constraints leading to such theories).

To be clear, we do not contend that there are no initial constraints of any kind on learning or development. We accept, for example, that some animals may be endowed with an initial bias to link taste with sickness but not with electric shock (Garcia & Koelling 1966), and that perceptual mechanisms have evolved to facilitate, among other things, the representation of external three-dimensional space and the segregation of the perceptual world into objects. Where we appear to differ from many theorists is in our feeling that, for many aspects of semantic knowledge, there is no clear reason at present to rely so heavily upon the invocation of initial domain-specific principles. Mechanisms exist that can learn to behave in domain-specific ways based on experience, without the need for extensive initial domain-specific commitments.

A second point of contrast with theory-based approaches lies in the question of whether semantic abilities are fundamentally rooted in causal knowledge. We certainly agree that children learn about and rely upon knowledge of causal properties and relations, and that this knowledge constitutes a part of their semantic knowledge. We do not accept, however, the need to attribute special status to causal knowledge; and we don't believe that causal knowledge necessarily carries with it any real appreciation of mechanism. For us, causal knowledge, together with all other forms of semantic knowledge, inheres in the configuration of weights that allows the semantic network to generate expectations about the likely outcomes of particular event sequences. Properties that enter into causal relationships with other properties are, by definition, associated with more predictable outcomes across different events. Hence such properties will covary coherently with other properties, and consequently, they will be quickly learned and strongly weighted by the learning mechanisms we have described. Also, we fully accept that words like "cause" are part of language and that such words can influence how we think about event sequences – possibly leading us on some occasions to assign greater centrality to events that are described as causes rather than effects. We simply hold that such phenomena do not require that causal knowledge be construed as fundamentally different from other kinds of semantic knowledge.

Third, the theory-theory has what we believe is an important and related set of weaknesses, at least as it has been developed up to now. Specifically, theory-theory is for the most part noncommittal about the nature of the representations and processes that underlie semantic task performance and the development of semantic abilities. The most systematic statements of the approach (Gopnik & Meltzoff 1997; Gopnik & Wellman 1994) contain no specification of mechanisms for the representation, use and acquisition of the knowledge underlying semantic task performance. Instead, the authors of these works simply suggest that it is useful to think of the child's knowledge as being, in some respects, analogous to a scientific theory. The subsequent effort by Gopnik et al. (2004) to characterize children's inferences as conforming to normative rules of causal inference does not really alter this lack of commitment to an underlying mechanism – indeed, Gopnik et al. (2004) explicitly eschew any such commitment.

Lack of commitment to mechanism can, of course, be a virtue when any such commitment would be premature. In such cases the theory simply remains underspecified. Without a more mechanistic specification, however, the analogy to explicit scientific theories brings with it a tendency to attribute properties of such theories to naive domain knowledge, whether such attribution is intended or not. In our view, this tendency can be counterproductive, because there are important properties of scientific theories that naturalistic human semantic knowledge does not actually have. Real scientific theories are explicit constructions, developed as vehicles for sharing among a community of scientists a set of tools for deriving results (such as predictions and explanations) using explicit, overtly specified procedures that leave a trace of their application through a series of intermediate steps from premises to conclusions. As far as we can tell, few theory-theorists would actually wish to claim that these properties of real scientific theories are also characteristic of the intuitive domain knowledge that underlies the performance of children or adults in naturalistic semantic tasks.

We suspect, however, that these aspects of real scientific theories occasionally filter into the thinking of researchers. For example, Spelke et al. (1992) speak of children reasoning from principles stated in propositional form. This idea may provide a useful basis for deriving predictions for experiments, whether or not anyone actually believes that the principle is held in explicit propositional form and enters into a reasoning process that follows specified rules of inference. But it may also carry additional implications that lead to unjustified conclusions. For example, the notion that a theory contains explicit principles and/or rules carries with it the tendency to suppose that there must be a mechanism that constructs such principles

and/or rules. Yet it is easy to show that the full set of possible principles or rules vastly outstrips those that children are said to actually use; and that the subset that children are said to use is underdetermined by actual evidence. Thus, the tacit invocation of explicit principles or rules ends up motivating the suggestion that there must be initial domain constraints guiding at least the range of possible principles that might be entertained (cf. Chomsky 1980; Keil 1989). If, however, behavior is not governed by explicit principles or rules, it is only misleading to consider the difficulties that would arise in attempting to induce them. By proposing that learning occurs through the gradual adaptation of connection weights driven by a simple experience-dependent learning process, the PDP approach avoids these pitfalls and allows us to revisit with fresh eyes the possibility that structure can be induced from experience.

With these observations in mind, we are now in a position to consider the relationship between the PDP approach to semantic cognition and theory-based approaches. One possible stance would be to suggest that the PDP framework constitutes an implementation of a theory-based approach – one that simply fills in the missing implementational details. Though in some ways this suggestion is appealing, we have come to feel that such a conclusion would be misleading, since the representations and processes captured by PDP networks are quite different from the devices provided by explicit scientific theories. While the knowledge in PDP networks may be theory-like in some ways, it is expressly not explicit in the way it would need to be in order to constitute a theory by our definition. Thus, we would argue that the PDP framework provides a useful alternative framework for understanding the acquisition, representation, and use of semantic knowledge.

5. Principles of The PDP approach to semantic cognition

We consider here the core principles underlying our approach to semantic cognition – those aspects of the simple model implementation to which we are strongly committed. The model itself is obviously greatly simplified. We have discussed some of the ways the model might be extended; and we envision that a more complete model may involve additional elaborations that we have not foreseen. The following principles capture, however, aspects of the simple model that we believe will prove critical to any such future account; they are considered at length in *Semantic Cognition* (Ch. 9).

1. *Predictive error-driven learning.* Our current work grows in part out of a long-standing effort to apply the PDP framework to aspects of cognitive development (McClelland 1989; 1994; Munakata & McClelland 2003). This work has stressed how predictive error-driven learning may provide the engine for knowledge acquisition in a wide range of domains, including language, object permanence, and causal reasoning; we believe that the same engine drives semantic knowledge acquisition.

2. *Sensitivity to coherent covariation.* The models we have considered are strongly sensitive to patterns of coherent covariation among the properties that characterize different items and contexts; we propose that such sensitivity is critical to understanding many aspects of semantic cognition.

3. *The convergence principle.* Sensitivity to coherent covariation is not a property of all networks that might be trained with predictive error-driven learning. Rather, such sensitivity requires that error signals for all sources of information about an item converge, at some point in the network, on the same set of connection weights. In the Rumelhart network, such convergence occurs at the first layer of weights projecting from *Item* to *Representation* layers – error signals from all output units, across all contexts, influence how these weights change, and permit the network to detect patterns of coherent covariation among them. Other network architectures considered in *Semantic Cognition* (Ch. 9) do not have this property and so will not be sensitive to coherent covariation, and nor will they exhibit the interesting behaviors critical to our account of semantic abilities.

4. *Distributed representation.* Something that sets the PDP approach to human cognition apart from some other connectionist approaches is the stipulation that representations are distributed: the same units participate in representing many different items, with each individual representation consisting of a particular pattern of activity across the set. Importantly for the current work, distributed representations promote generalization: what is known about one item tends to transfer to other items with similar representations. Although our models do employ localist input and output units, these never communicate with each other directly – their influences on one another are always mediated by distributed internal representations.

5. *Weak initial differentiation.* A specific property of the Rumelhart model, very important to the way that it functions, is that the network is initialized with very small random connection weights, so that all items initially receive nearly identical distributed representations. The important consequence of this choice is that at first, whatever the network learns about any item tends to transfer to all other items. This allows for rapid acquisition and complete generalization of information that is applicable to all kinds of things; but it also induces in the network a profound initial insensitivity to the properties that individuate particular items. Different items are treated as effectively the same until considerable evidence is accumulated indicating how they should be distinguished, based on patterns of coherent covariation. After each wave of differentiation, there remains a tendency to treat those items not yet distinguished as very similar. In general, this property of the network imposes a very strong tendency to generalize, instead of capturing idiosyncratic differences between items.

6. *Gradual, structure-sensitive learning.* Our simulations depend on slowly and gradually adjusting the weights during learning, so that weight changes are not dominated by any single experience or a limited set of experiences, but tend to benefit processing for all items and all contexts. We believe that learning in a real environment requires the assimilation of statistical properties, some of which may be strong and of fairly low-order, but others of which are much subtler and infrequently encountered. The environment so characterized favors slow learning for reasons discussed in McClelland et al. (1995) and *Semantic Cognition* (pp. 65–66).

7. *Activation-based representation of novel objects.* If learning in the semantic system is a gradual and incremental process, then it cannot mediate the ability to immediately use new information obtained from one or a few experiences. To explain such abilities, we propose that the semantic system can dynamically construct useful internal representations of new items and experiences – instantiated as patterns of activity across the same units that process all other items and events – from the knowledge that has accumulated in its weights from past experience. In our book we have implemented this principle using backpropagation-to-representation – a process that allows the feed-forward Rumelhart network, given some information about a novel object's observed properties, to assign it an internal representation (See *Semantic Cognition*, pp. 63–65 and 69–76, for details and discussion). The important point is that the representations so assigned are not stored in connection weights within the semantic system. Instead, the representations are used directly as the basis for judging semantic similarity and making inferences about the object's unobserved properties and behaviors in other situations. To allow such representations to be brought back to mind in another situation, they can be stored via the complementary fast-learning system in the hippocampus; and with repetition, these representations can be gradually integrated in the connection weights in the neocortical learning system.

It must be noted that a system adhering to the principles described above has several limitations; specifically, it tends to be quite insensitive to idiosyncratic properties of individual objects and learns very slowly. In light of this and other considerations, McClelland et al. (1995) extended earlier ideas of David Marr (1971) in arguing that it is crucial to provide a second, complementary learning system that relies on sparse, non-overlapping representations rather than densely overlapping, distributed ones, and in which large weight changes can be made based on one or a few presentations of novel information. This allows knowledge of idiosyncratic properties of individuals to be learned rapidly and generalized very narrowly, complementing the positive features of the slow-learning system. McClelland et al. (1995) identify the fast-learning system with the medial temporal lobes, and the slow-learning system primarily with the neocortex. Such a system would support a wide range of important functions that are quite domain general; as such, both the slow-learning cortical system and the fast-learning hippocampal system are, in our view, parts of a general-purpose, cross-domain learning system.

6. Broader Issues

In the final chapter of *Semantic Cognition*, we touch on some broader issues in cognitive science that relate to the specific issues in conceptual development that have been our focus here. In the next sections we summarize briefly the points we made in that discussion that have not already been covered in this précis.

6.1. Thinking and reasoning

As is often the case with PDP models, we suspect that our models will arouse in some readers a feeling that there is some crucial element of cognition that is missing. Even those who feel generally favorable toward our approach may have a sense that there is something to human conceptual abilities that goes beyond implicit prediction and pattern completion. Do we really think this is all there is to semantic cognition? What about "thinking"?

A suggestion explored both in Hinton's (1981) early work and by Rumelhart et al. (1986c) is that temporally extended acts of cognition – what one would ordinarily call "thinking" – involves the repeated querying of the processing system: taking the output of one prediction or pattern completion cycle and using that as the input for the next. Rumelhart illustrated the basic idea with a mental simulation of a game of tic-tac-toe, in which a network trained to generate the next move from a given board position simply applied its successive predictions to its own inputs, starting with an empty board. Hinton used a similar idea to suggest how one might discover the identity of someone's grandfather from stored propositions about fathers: One could simply complete the proposition "John's father is" and from the result construct a new probe for the father of John's father. A slightly more general idea is that thinking is a kind of mental simulation, not only encompassing internally formulated propositions or sequences of discrete game-board configurations, but also including a more continuous playing out of imagined experience. This perspective is related to Barsalou's proposals (e.g., Barsalou et al. 2003), and seems to us to be quite a natural way of thinking about thinking in a PDP framework.

6.2. Relationship between PDP models and Bayesian approaches

Over the last several years there has been considerable interest in the idea that various aspects of human cognition, including many aspects of semantic cognition, can be characterized as a process of Bayesian inference (see, e.g., Anderson 1990; Oaksford & Chater 1998). What is the relationship between these ideas and the approach we have taken here?

One perspective might be that they are distinct alternative frameworks for thinking about human cognition. In our view, however, Bayesian approaches are not replacements for connectionist models, and nor for symbolic frameworks. Rather, they provide a useful descriptive framework that can be complementary to these other more mechanistic approaches. Indeed, Bayesian approaches are often cast largely at Marr's (1982) computational level – specifying, for example, a normative theory for inference from evidence under uncertainty. It is a further matter to provide a model at what Marr called the algorithmic level, which specifies the processes and representations that support the Bayesian computation. Connectionist models are cast at this algorithmic level and are thus not inconsistent with normative Bayesian approaches.

It is worth noting that many connectionist models were either designed to be, or were later shown to be, implementations of Bayesian inference processes (McClelland 1998). For example, the Boltzmann machine (Hinton & Sejnowski 1986) and Harmony theory (Smolensky 1986) are general-purpose frameworks for deriving optimal (Bayesian) inferences from input information, guided by

knowledge built into connection weights; and the stochastic version of the interactive activation model (McClelland 1991; Movellan & McClelland 2001) has this property, also. The backpropagation algorithm implements a Bayes optimal process in the sense that it learns connection weights that maximize the probability of the output given the input (subject to certain assumptions about the characteristics of the variability that perturbs the observed input–output patterns), as several authors pointed out in the early 1990s (MacKay 1992; Rumelhart et al. 1995).

Connectionist models might therefore be viewed as specifying the actual algorithms that people use to carry out Bayesian computations in specific task situations. There is, however, one important point of difference between our approach and most such models that we are aware of. Unlike the highly distributed connectionist models that are the focus of our own work, the Bayesian models generally operate with a set of explicitly enumerated alternative hypotheses. For example, in Bayesian theories of categorization, an item is assigned a posterior probability of having come from each of several possible categories, and each category specifies a probability distribution for the features or attributes of all of its members. In our PDP approach there are no such categories, but rather each item is represented in a continuous space in which items are clustered and/or differentiated to varying degrees. We hold that the use of distributed representations has desirable computational consequences, and it will be interesting to explore further how they might be encompassed within a Bayesian framework.

6.3. Semantic cognition in the brain

The neural basis of semantic cognition has been the focus of a great deal of recent research using a variety of methodologies. Investigations of semantic impairment following brain damage and functional imaging studies of healthy adults both support the general conclusion that semantic processing is widely distributed across many brain regions. One widely held view for which substantial evidence now exists is that the act of bringing to mind any particular type of information about an object evokes a pattern of neural activity in the same part or parts of the brain that represent that type of information directly during perception and action (Martin & Chao 2001).

Our simple and abstract model can be brought into line with this work by placing the input or output units representing different types of information in different brain regions (Rogers et al. 2004), so that units coding different kinds of movement are located in or near brain regions that represent perceived movement, those coding color are in or near regions mediating color perception, and so forth. In addition to these units, however, our theory calls for a convergent representation: a set of representation units that tie together all of an object's properties across different information types. Such units might lie in the temporal pole, which is the focus of pathology in the purest and most profound semantic disorder, semantic dementia (Mummery et al. 2000). Others (Damasio 1989; Barsalou et al. 2003) have emphasized the potential role of this region as a repository of addresses or tags for conceptual representations. We emphasize that the patterns of activation in these areas are themselves "semantic" in two respects. First, their similarity relations capture the semantic similarities among concepts, thereby fostering semantic induction. Second, damage or degeneration in these areas produces a pattern of degradation that reflects this semantic similarity structure. Distinctions between items that are very similar semantically tend to be lost as a result of damage to this area, whereas distinctions between highly dissimilar area concepts are maintained (Rogers et al. 2004).

Note that we do not contend that these representations contain a "copy" of semantic features, propositions, images, or other explicit content. In agreement with many others, we believe that this content is instantiated in sensory, motor, and linguistic representations closely tied to those that mediate perception and action – roughly corresponding to the input and output units in the Rumelhart model. Instead, the intermediating "semantic" representations that, we suggest, are encoded in anterior temporal lobe regions are like the learned internal representations acquired in the Rumelhart model. They capture similarity structure that is critical for semantic generalization and induction, and that determines which explicit properties are "important" for a given concept; but they do not encode directly interpretable semantic information.

7. Conclusion

It is clear to us that our efforts are only one step toward the goal of providing an integrative account of human semantic cognition. The principles stated here are very general, and we expect they will remain the subject of ongoing debate and investigation. The form that a complete theory will ultimately take cannot be fully envisioned at this time. We do believe, however, that the small step represented by this work, together with those taken by Hinton (1981) and Rumelhart (1990), are steps in the right direction; and that, whatever the eventual form of the complete theory, the principles exemplified in this précis will be instantiated in it. At the same time, we expect that future work will lead to the discovery of additional principles, not yet conceived, which will help the theory we have laid out here to gradually evolve. Our main hope for this work is that it will contribute to the future efforts of others, thereby serving as a part of the process that will lead us to a fuller understanding of all aspects of semantic cognition.

Open Peer Commentary

Semantic cognition or data mining?

doi:10.1017/S0140525X08005906

Denny Borsboom and Ingmar Visser

Department of Psychology, University of Amsterdam, 1018WB Amsterdam, The Netherlands.

d.borsboom@uva.nl **http://users.fmg.uva.nl/dborsboom/**
i.visser@uva.nl **http://users.fmg.uva.nl/ivisser/**

Abstract: We argue that neural networks for semantic cognition, as proposed by Rogers & McClelland (R&M), do not acquire semantics

and therefore cannot be the basis for a theory of semantic cognition. The reason is that the neural networks simply perform statistical categorization procedures, and these do not require any semantics for their successful operation. We conclude that this has severe consequences for the semantic cognition views of R&M.

If Rogers & McClelland (R&M) have done what they say they have done in *Semantic Cognition* (Rogers & McClelland 2004), then they have solved one of the most vexing problems in philosophy, cognitive science, and psychology: namely, the problem of giving a *mechanistic* explanation of why humans are able to have thoughts and beliefs that are *about* objects, states of affairs, propositions, and the like. That is, should R&M's claims hold true, so that the neural network they propose indeed acquires semantics, then they should be considered to have solved the long-standing problem of *intentionality* (Brentano 1874/1995). It is thus important to consider the question whether the neural network proposed indeed does this. That is, we need to evaluate whether the network really acquires "internal representations of objects" (*Semantic Cognition*, p. 69), so that it would be accurate to state, say, that the network "represents the daisy, the sunflower, the canary, and the sunfish as similar to one another" (p. 287), or that it has "Like the children, ...inferred that the object [an echidna] can move and that it has legs." (p. 252).

Let us consider an example to see whether such claims hold up to scrutiny. We read, on page 215, that the network displayed "a tendency to extend the name 'dog' . . . to robins." So, according to R&M, their network has learned to apply the name "dog" to some objects, namely, dogs, and now extends that name to other objects, namely, robins. Obviously, R&M cannot be taken to seriously mean this. Their network has never seen dogs or robins. Hence, it has never applied the name "dog" to any dogs whatsoever and could not possibly "extend" that name to robins, simply because there aren't any robins around to extend it to. Now, if the network cannot be said to learn to apply the concept of "dog" to dogs, or to extend that concept to robins – as we think is painfully obvious – then what *can* it be said to do?

Clearly, the network can be said to respond to certain input patterns with certain output patterns. What do these patterns consist of? R&M assert that the network is trained with the corpus of Collins and Quillian (1969), that is, a set of concepts ("dog," "robin") with associated properties ("is a living thing," "can fly," etc.). Literally speaking, however, this is incorrect. What the network is trained with, and probed to reproduce, is a pattern of zeroes and ones that was *extracted* from Quillian's corpus. "Extracted by whom?," one may now plausibly ask – after all, we are talking about semantic cognition here, and if it were the *neural network* that extracted these patterns, say, from its observations of dogs and robins – real or simulated – then that would at least count for something. The disappointing answer, however, is that these patterns were not extracted by the network, but were carefully put in place by R&M themselves.

A second problematic aspect of the network architecture is that the network learns to associate a label, say, "dog," with the corresponding properties, for example, "animal," "tail," "bark," and so forth. Children, when learning to make sense of the world, do exactly the opposite: They see a bunch of features, and their task is to learn that there are similarities between objects based on these features, and that, if similarity is large enough, objects should be clustered into a single concept. The network presented by R&M seems to be learning language by reading a dictionary, which is most certainly not the way children learn a language. Moreover, the network by R&M fails to capture an important aspect of categorization learning, which is the sudden of stepwise nature of that learning process (Schmittmann et al. 2006).

But wouldn't *some* semantic cognition be necessary for the network to operate as well as it does? Certainly not. Despite the superficial resemblance of a connectionist model to the neural structures in our heads, all that the network of R&M does is to carry out a statistical categorization procedure. No semantic cognition is required for this purpose, as can be easily verified by looking at other statistical programs of this type, for instance, those incorporated in statistical packages like SPSS. It is ironic, in this respect, that R&M use clustering and multidimensional scaling programs to study the "representations" that the network "has" (e.g., see *Semantic Cognition*, Ch. 7), for their network is exactly such a program. Moreover, given that such wonderful statistical techniques as multidimensional scaling and hierarchical clustering are available to describe these learning processes, why not just stick to those techniques rather than use an overparameterized version of them implemented in a neural network (cf. Ripley 1996)? The usually quoted advantage of neural networks over statistical techniques is that they behave better in the face of noisy data; that advantage, however, is almost certainly completely lost in the R&M network, because it has localist rather than distributed representations.

We submit that the neural network of R&M does not acquire any semantics whatsoever, and as a result their theory cannot be a theory of semantic cognition. It seems to us that R&M have two options in responding to this charge, of which one is hopeless and the other absurd. The hopeless route to defuse the criticism is to attempt to show that a neural network in fact does something qualitatively different from, say, principal components analysis. This is hopeless because, despite all the fancy talk of "neurons" that "fire" to create "distributed representations" and the like, the sobering fact is that neural networks, in general, can be shown to do nothing but principal components analysis, maximum likelihood regression, and so on (cf. Hadley [2000] for relationships between classical computational models and neural networks; and Sarle [1994] for an overview of equivalencies between neural networks and different regression techniques and principal components analysis). Alternatively, R&M may take the other horn of the dilemma and argue that *humans* do nothing but carry out statistical categorization procedures in acquiring semantics. This is not logically inconsistent, to be sure, but it has some very unhappy consequences: If there is nothing to semantic cognition except for the sort of data mining that a neural network does, then we are logically committed to the thesis that SPSS does semantic cognition when we tell it to factoranalyze a data set. For anybody remotely familiar with either SPSS or factor analysis, this would seem a position sufficiently absurd to dismiss out of hand.

Inductive reasoning and semantic cognition: More than just different names for the same thing?

doi:10.1017/S0140525X08005918

Aidan Feeney, Aimee K. Crisp, and Catherine J. Wilburn

Department of Psychology, Durham University, Stockton-on-Tees TS17 6BH, United Kingdom.

aidan.feeney@durham.ac.uk **www.dur.ac.uk/aidan.feeney**
a.k.crisp@durham.ac.uk
c.j.wilburn@durham.ac.uk

Abstract: We describe evidence that certain inductive phenomena are associated with IQ, that different inductive phenomena emerge at different ages, and that the effects of causal knowledge on induction are decreased under conditions of memory load. On the basis of this evidence we argue that there is more to inductive reasoning than semantic cognition.

Rogers & McClelland's (R&M's) deeply impressive book, *Semantic Cognition* (2004), outlines a plausible alternative to theory-theory, and it shows how parallel distributed processing (PDP) models can capture many phenomena in the literatures on categorisation, naming, and concepts. We are interested in implications of R&M's approach for what we know about inductive reasoning and its relation to knowledge. Our understanding of

their position is that in many cases inductive reasoning and semantic cognition are just different names for the same thing. However, in our view, although knowledge is very important to an understanding of thinking, there are limits to what can be explained by recourse to knowledge and the processes by which it is attained. First, we will describe some effects, our own and other people's, which appear to challenge accounts that equate thinking with semantic cognition. Then we will speculate as to what kinds of account might best capture those effects.

The literature on deductive reasoning contains the clearest evidence that there is more to thinking than semantic cognition. For example, Handley et al. (2004) used a belief bias task where 10-year-old participants were asked to reason about arguments the validity and believability of whose conclusions had been orthogonally manipulated. Participants also completed measures of inhibitory control and working memory. Successful performance on this task calls for the inhibition of outputs from semantic cognition, and Handley et al. observed that inhibitory control and working memory were independent predictors of the ability to respond in accord with logical validity.

Of course, R&M make no claims about deduction. However, some of our own work asks whether inductive reasoning can be wholly captured by fast and parallel knowledge-based processes or whether slow, resource-demanding processes also play a role. For example, Feeney (2007) studied inductive projection using arguments with multiple premises. Such arguments can be used to study whether people are sensitive to diversity and amount of evidence when evaluating inductive arguments, and sensitivity to these phenomena has been modelled in wholly similarity-based ways (Osherson et al. 1990; Sloman 1993). Feeney showed that a measure of IQ is associated with people's sensitivity to these principles. The results are complex, but particularly in the case of diversity, those participants who scored highest on the IQ test tended to be most sensitive to the diversity of the premises. One interpretation of correlations between IQ and performance on particular thinking tasks is that they indicate the involvement of slow, symbol-manipulating processes in thinking (see Stanovich 1999). That is, inductive reasoning is more than semantic cognition, and is based on more than processes that allow for the calculation of similarity between representations.

A related finding concerns when sensitivity to properties of the premises of an inductive argument develops. Wilburn and Feeney (2007) have shown that sensitivity to diversity begins to emerge at age 7, whereas sensitivity to amount of evidence does not begin to emerge until age 13. We interpret this finding as suggesting that in a category-based inductive argument, sensitivity to amount of evidence requires the reasoner to know that larger samples make for sounder inferences, whereas sensitivity to diversity can be demonstrated on the basis of similarity calculations alone. This finding also suggests that there is more to thinking than mere similarity.

Like R&M (*Semantic Cognition*, Ch. 8), we have also been concerned with the effects of knowledge about causal relations on inductive generalisation. A particularly interesting case comes from Medin et al. (2003), who demonstrated the category-based conjunction fallacy. They compared strength ratings for the following argument:

Lead has Property X, therefore pipes and plumbers have Property X

to the mean strength ratings for the causally *near* generalisation from lead to pipes and to the causally *distant* generalisation from lead to plumbers. (We term lead and pipes *causally distant* because the reasoner has to infer the involvement of pipes to explain the transmission of Property X from lead to plumbers.) Medin et al. (2003) demonstrated that, on average, people commit the conjunction fallacy. That is, the strength rating for the argument with the conjunctive conclusion is higher than the average strength rating for other two arguments.

Feeney et al., (2007) followed up on this finding and showed that the near generalisation from lead to pipes is rated strongest, whereas the distant generalisation from pipes to plumbers is rated weakest. In addition, we found that participants highest in IQ were more likely to rate the near generalisation stronger than the conjunctive argument. In further follow-up experiments (Crisp et al., under review) we asked participants to concurrently perform a working memory task whilst rating generalisation strength. The secondary task increased rates of the conjunction fallacy observed when ratings for the conjunctive argument were compared to the distant case, but not when compared to the near case. Our interpretation of these findings is that in the distant case, people resisted the conjunction fallacy because they explicitly reasoned about causal relations and reconstructed the causal chain linking, for example, lead to plumbers. Having reconstructed the causal chain, they assigned equally high-strength ratings to distant and conjunctive arguments. A concurrent task impeded their ability to engage in this causal reasoning in the distant case, whereas it had no effect in the near case because the stronger causal relation was immediately available. The same basic pattern was obtained when participants were encouraged to answer quickly. Thus, the individual differences, secondary task, and speeded task data suggest that some effects of knowledge on thinking are moderated by processes that are associated with IQ and working memory, and which take time.

Our preferred explanation for these findings is that there are at least two types of thinking (see Evans 2006; Sloman 1996; Stanovich 1999), a fast and associative form of thinking, and a slower and sequential type of thinking. The first type of thinking performs, among other operations, similarity calculations, whereas the second type applies rules and makes some (but not all) inferences about causal relations. It has been studied by researchers interested in models (Johnson-Laird 2006), rules (Rips 1994), or simulations (Evans & Over 2004) for reasoning. R&M's models of semantic cognition appear more relevant to the first type of thinking than they do to the second.

Impressed as we are by R&M's book, we cannot see how their current models can capture our data. Of course, R&M have anticipated our concerns and questions (see *Semantic Cognition*, pp. 371–73), but it may take another book to convince us of their answer.

ACKNOWLEDGMENT

Aimee Crisp and Catherine Wilburn are supported by ESRC postgraduate training awards.

Context, categories and modality: Challenges for the Rumelhart model

doi:10.1017/S0140525X0800592X

James A. Hampton

Psychology Department, City University, London EC1V OHB, United Kingdom.

hampton@city.ac.uk

http://www.staff.city.ac.uk/hampton

Abstract: Three issues are raised in this commentary. First, the mapping of semantic information into the different layers could be done in a more realistic way by using the *Context* layer to represent situational contexts. Second, a way to differentiate category membership information from other property information needs to be considered. Finally, the issue of modal knowledge is raised.

The parallel distributed processing (PDP) approach to modeling cognition has provided a healthy redress of the balance between empiricist and rationalist accounts of human thought. Following Chomsky's demolition of behaviorist theories of thought and language, it was assumed for many years that the mind was a symbol-processing machine, following algorithmic, syntactic rules to solve problems, achieve goals, and so forth. The

discovery that PDP networks can behave in systematic rule-following ways has been matched by growing evidence that in many important respects our psychological processes are also only approximately rule-governed, so that a felicitous marrying of model and data has been achieved. In *Semantic Cognition*, Rogers and McClelland (2004) show how the Rumelhart model can learn to accurately associate properties with their respective noun concepts, while at the same time showing the general influence of the similarity structure of the knowledge being represented. Just as Rosch (1978) proposed, the mind is sensitive to the correlational structure of the world and the concepts we learn correspond to the complex covariation of different properties across semantic domains. The Rumelhart model provides the missing mechanism for how this arises, while at the same time modeling a wide range of now familiar prototype effects such as basic levels, typicality and category-based induction.

As presented, the model does not aim to represent the actual contents of anyone's semantic memory, and so there is still much detail to explore and develop. The following comments are suggestions about directions in which the model could usefully be taken, both to demonstrate its explanatory power and test its limits.

Use of the Context Relation layer. The *Context* or *Relation* layer is currently used to determine the type of relation between the noun concept (e.g., *pine*) and a property (e.g., *ISA tree, CAN grow, IS tall*). This use of the *Context* layer appears arbitrary and could lead to difficulties in a more realistic conceptual domain. The *Context* layer is clearly a vital part of the architecture of the model and cannot be omitted. But perhaps the *Context* layer might more usefully encode just that – context. Barsalou (2003) has reviewed evidence that the properties generated to a noun concept relate to an imagined situational context – so that, for example, very different properties would be generated for a car seen in a parking lot versus a car from the point of view of a driver. Typicality structure can also be highly context dependent (Barsalou 1987; Roth & Shoben 1983). Output property units could then encode whole properties (can grow, is tall) undifferentiated by their syntax. Syntactic form is a poor guide to the relatedness of properties. In the model most of the "is" relations were visually based. But in real life an "is" relation can encode any number of non-perceptual and abstract properties such as *is valuable*, *is annoying*, or *is bad for your health*. Grouping properties by syntax may not correspond to any real-world structure. An alternative suggestion to try here would be to use the *Context* layer to input the type of property (part, appearance, function, behavior, origin, etc.) using semantic rather than syntactic criteria to determine types.

Category information is not just another property. Categorical *ISA* relations have traditionally been treated very differently in studies of semantic memory from other properties. Knowing the category membership of an item will normally provide a much broader range of useful inferences about it than will knowledge of a salient property. The *ISA* relation captures the kind of thing that the item is, whereas properties just capture a particular property. Category information is also verified more rapidly (Hampton 1984). The model does not reflect this difference structurally, although it is notable that all of the input items reappear as *ISA* output units. How could the model be asked whether it had learned the properties of superordinate categories – for example, that trees have roots or that fish have gills?

Quantification and modality. A difficulty for any similarity-based model is the handling of extensional reasoning and quantified statements. When it has mastered its knowledge domain, the model will correctly verify that a robin is a robin, a robin is a bird, and a robin is red. It will not be able to explain, however, that *a robin is a robin* is tautologically true, *a robin is a bird* is necessarily true (assuming that any non-bird could never resemble a robin sufficiently to belong in that class), whereas *a robin is red* is generically true – only being true of most adult robins (or in Europe only of adult males). Knowledge of what actually exists is not primarily the job of semantic memory, but the model clearly lacks a way to handle truth under different quantifiers.

Failure to consider the truth of statements extensionally is quite possibly an advantage of the model given that people are also bad at it, and succumb to similarity-based "non-logical" effects when reasoning about category membership (e.g., Hampton 1982; Jönsson & Hampton 2006). But it would be worth exploring whether the model can learn the difference between properties that are necessarily true and those that are typically true.

The reverse side of the coin is whether the model can determine which properties can be expected to co-occur and which may not. Suppose that backpropagation to representation is used to find a representation of an item that is large and yellow, and has petals, as opposed to an item that has roots, gills, and feathers. Can it be demonstrated that some representations are found rapidly and with low residual error (even although the properties have not co-occurred in the training set), whereas others are impossible to represent without a high degree of error. Modal intuitions of necessity and possibility (Rips 2001) are an important aspect of semantic cognition, and it would be a bonus for the research program to show how the network can also match such intuitions.

Structured models of semantic cognition

doi:10.1017/S0140525X08005931

Charles Kemp[a] and Joshua B. Tenenbaum[b]

[a]*Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213;* [b]*Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA 02139.*

ckemp@cmu.edu **http://www.charleskemp.com**

jbt@mit.edu **http://web.mit.edu/cocosci/josh.html**

Abstract: Rogers & McClelland (R&M) criticize models that rely on structured representations such as categories, taxonomic hierarchies, and schemata, but we suggest that structured models can account for many of the phenomena that they describe. Structured approaches and parallel distributed processing (PDP) approaches operate at different levels of analysis, and may ultimately be compatible, but structured models seem more likely to offer immediate insight into many of the issues that R&M discuss.

It is widely accepted that cognition can be understood at multiple levels of analysis, but there are different claims about the nature of these levels (Broadbent 1985; Marcus 2001; Rumelhart & McClelland 1985; Smolensky 1988). In *Semantic Cognition* (2004), Rogers & McClelland (R&M) appear to suggest that parallel distributed processing (PDP) approaches and structured approaches lead to proposals at the same level of analysis, and are therefore competitors. Like some previous researchers (Smolensky 1988), we believe that these two paradigms are compatible, and that they aim for explanations at different levels of analysis.

Since R&M treat structured approaches as the competition, they naturally emphasize the problems they see with structured models of cognition. Among other criticisms, they suggest that structured approaches cannot capture typicality, exceptions, and the graded inferences that are characteristic of human learning (*Semantic Cognition*, p. 44); that there are few attempts to explain how taxonomic hierarchies might be acquired (pp. 13, 31); and that structured approaches do not explain why people make very different inferences when reasoning about different kinds of properties (e.g., "has cold blood" vs. "weighs ten tons," p. 34).

If PDP approaches and structured approaches operate at different levels of analysis, then many phenomena (e.g., graded inferences and learning) will turn out to be compatible with

both approaches. Some structured models suffer from the limitations described by R&M, but others do not. Models that combine probabilistic inference with structured representations allow for noise and exceptions, and are able to account for graded generalizations and typicality effects (Anderson 1991; Kemp & Tenenbaum 2003; Tenenbaum 2000). Combining probabilistic inference with structured representations also leads to a principled account of how these representations might be acquired, and there are probabilistic models that learn categories (Anderson 1991), taxonomic hierarchies (Kemp et al. 2004), abstract schemata (Kemp et al. 2007a), and logical theories (Kemp et al. 2008)

Structured approaches also help to explain how different kinds of knowledge support inferences about different kinds of properties. Shafto et al. (2005) have shown that a probabilistic model that incorporates a taxonomic hierarchy accounts for inferences about taxonomic properties (e.g., "has sesamoid bones"), and that a model that relies on a structured representation of a food web predicts inferences about disease properties (e.g., "carries *E. Spirus* bacteria"). As R&M point out, allowing for multiple structured representations can raise some challenging problems, but none of these problems seems insurmountable. For example, there are models that learn multiple representations of the relationships between a set of categories (Shafto et al. 2006) and models of inductive reasoning that combine the knowledge embedded in multiple representations (Kemp et al. 2007b).

Although psychologists should ultimately aim to understand semantic cognition at multiple levels of analysis, it is useful to consider whether the current generation of structured models shows more or less promise than the current generation of PDP models. The community as a whole may pursue multiple approaches, but individuals will need to decide which of these approaches is most deserving of their time and attention. R&M provide a comprehensive description of the benefits that PDP approaches can provide, but there are two primary reasons why structured approaches appear more promising to us.

First, formalizing commonsense knowledge is a major challenge for models of semantic cognition. R&M suggest that structured models may be "too restrictive or constraining to capture the nuances of human knowledge of all sorts of things" (*Semantic Cognition*, p. 44). We are drawn toward the opposite conclusion, and feel that there are fundamental aspects of human knowledge that are naturally captured using structured representations but are difficult for current-generation PDP models to incorporate. Several examples can be found in the literature, but here we focus on two: standard PDP models do not provide a compositional system for building complex concepts out of simpler pieces (Fodor & Pylyshyn 1988) and find it difficult to discover abstract relational laws (Marcus 2001). The work of R&M appears to suffer from both of these limitations. First, it is not clear how the model will handle relations of several different arities (e.g., *CAN[canary, fly]*, and *CAN[cat, eat, canary]*), or cases involving nested relations (e.g., *CAUSE[EATS(canary, food), GROWS(canary)]*). Second, it seems likely that the model will struggle to acquire abstract knowledge about relations: for instance, it is not clear whether the model can recognize that a relation (e.g., *ISA[·,·]*) tends to be transitive, or that two relations (e.g., *HAS[·,·]* and *POSSESSES[·,·]*) are near identical in meaning. Abstract knowledge of this sort should support inferences about novel categories: for example, given that a dax is a wug and a wug is a zav, a dax is likely to be a zav. Extensions of the R&M approach may be able to overcome many of the limitations we identified, and PDP models of relational learning and reasoning (Doumas et al. 2008; Shultz & Vogel 2004) may help to show the way. At present, however, PDP models seem less successful at capturing complex systems of relational knowledge than models that rely on predicate logic as a representation language.

A second advantage of structured models is that they make more direct contact with previous psychological research on semantic cognition. As R&M suggest in their opening chapter (*Semantic Cognition*, Ch. 1), many psychologists who have written about categorization, the theory-theory, and related topics have seen the need for structured representations. To mention only one example, the work of Keil (1979) is motivated in part by the idea that categories are organized into taxonomic hierarchies, and Keil's approach can be directly converted into a structured model that helps to explain how taxonomic hierarchies are acquired and used for induction (Schmidt et al. 2006). The work of R&M may encourage psychologists to develop new experiments and theories that explore the notion of distributed representations. At present, however, the most profitable interactions between computational approaches and psychological studies seem likely to be organized around structured models.

Predictions about the future of any discipline are notoriously unreliable, but we will venture two of them. First, researchers will eventually understand how structured semantic representations are instantiated in the brain. Second, deep insights at the neural level will only be possible once we have a deep understanding of the computations supported by structured semantic representations.

Semantic cognition: Distributed, but then attractive

doi:10.1017/S0140525X08005943

Emilio Kropff and Alessandro Treves

Cognitive Neuroscience sector, SISSA – International School for Advanced Studies, 34014 Trieste, Italy; Kavli Institute for Systems Neuroscience and Centre for the Biology of Memory, MTFS, NO-7489 Trondheim, Norway.

emilio.kropff@ntnu.no **http://folk.ntnu.no/kropff/**

ale@sissa.it **http://people.sissa.it/~ale/**

Abstract: The parallel distributed processing (PDP) perspective brings forward the important point that all semantic phenomena are based on analog underlying mechanisms, involving the weighted summation of multiple inputs by individual neurons. It falls short of indicating, however, how the essentially discrete nature of semantic processing may emerge at the cognitive level. Bridging this gap probably requires attractor networks.

The simple feed-forward Rumelhart model presented in *Semantic Cognition* (Rogers & McClelland 2004) is clearly a valuable tool to study the formation and structure of semantic knowledge in terms of general connectionist systems. It allows an investigation of those properties that arise naturally in any network of neural-like elements, which individually sum a large number of inputs through experience-modifiable weights. Though rudimentary in certain respects, it is a huge step beyond conceptually more sophisticated constructs based on mere logic, such as the theory-theory approach, which explicitly eschew the distributed nature of the neural mechanisms underlying cognitive processes, and in so doing voluntarily confine themselves to the philosophical domain. It is not clear, however, whether the Rumelhart model, even when elaborated and made more complex, can exhaust all potential insight that neural network approaches may yield on the architecture of semantic cognition. More critically, it is not clear whether it can satisfactorily address the central challenge laid out by the logic-based approaches – that of accounting for the apparently discrete nature of much of cognition.

The surprising resilience of non-distributed conceptual accounts (inasmuch as these are the product of the minds of their proponents and followers) after all indicates the same cognitive process they purport to elucidate: that is, the tendency of semantic cognition to articulate itself in the form of discrete logical steps. To anchor cognition in the facts of neuroscience – and, in particular, in the essentially graded, analog nature

of the underlying neural computations – is indeed sacrosanct, provided the apparent discreteness of the mind eventually emerges. Networks trained with backpropagation seem to go a certain distance in the right direction, but it remains doubtful whether they can fully bridge the gap between analog neural computation and discrete mental operations. Developed twenty years ago, the backpropagation paradigm does not reflect, in its claim for neural plausibility, several phenomena that have been observed more recently, including the tendency of patterns of neural activity to sometimes fall into discrete attractor states (Akrami et al. 2006; Wills et al. 2005). We propose, with many others of course, to go beyond networks trained with backpropagation, and consider cortical networks that store memories based on known neurobiological mechanisms, as captured by associative plasticity "Hebbian" rules. The Hopfield model (1982), of equivalent abstract simplicity to the Rumelhart model, shows how discrete *attractor* states can govern the neural dynamics of networks with recurrent connections, such as had earlier been invoked by David Marr (1971) to account for associative memory processes in the hippocampus.

To address semantic memory issues – for example, the ones derived from the distributed representation of concepts in diverse cortical areas – attractor networks comprised of multiple equivalent modules, as in the scheme envisaged by Valentino Braitenberg (see Braitenberg & Schüz 1991), may offer a convenient perspective (Treves 2005). "Hidden" units in such multimodular associative models promise to play as important a role as in feed-forward networks. We have shown how an optimized Hebbian learning rule that includes local information about statistical biases in the activation of each unit (reflecting the "popularity" of semantic features) can store and retrieve semantic patterns of activity, overriding the limitations of classical autoassociative memory models (which were originally designed to store, effectively, orthogonalized representations). The optimized learning rule leads to interesting similarities with semantic memory phenomena, such as category specific deficits (Kropff & Treves 2007; Warrington & Shallice 1984). Observed distributions of feature popularity obtained with experimental approaches (such as feature norms; see McRae et al. 2005) appear, however, to hover around values that are too high to be compatible with retrieval (Kropff, forthcoming). Hidden units, functioning for example as conjunction detectors, could solve this problem by lowering the typical feature popularity. If so, it would be of great interest to understand their contribution to categorization, as this type of discretization of conceptual spaces emerges during the learning process; as well as to other phenomena studied in the book, such as illusory correlations or concept reorganization.

ACKNOWLEDGMENTS

The preparation of this commentary was partially supported by HFSP grant No. RGP0047/2004-C. Alessandro Treves is grateful for the hospitality of the Institute for Advanced Studies at the Hebrew University of Jerusalem.

A sneaking suspicion: The semantics of emotional beliefs and delusions

doi:10.1017/S0140525X08005955

Angus W. MacDonald, III

Department of Psychology, University of Minnesota, Minneapolis, MN 55455.
angus@umn.edu
http://www.psych.umn.edu/research/tricam/

Abstract: This commentary challenges Rogers & McClelland (R&M) to use their model to account for delusional belief formation and maintenance. The gradual development of delusions and the nature of disconnectivity in Capgras delusions are used to illustrate the role of emotional salience in delusions. It is not clear how this kind of emotional saliency can be represented within the current architecture.

The elegance of *Semantic Cognition: A Parallel Distributed Approach* (Rogers & McClelland 2004; henceforth *Semantic Cognition*) is in the explanatory power of several very simple principles. Rogers & McClelland (R&M) show the principles of distributed units with sigmoidal activation functions; varying connection strengths; and predictive, error-driven learning account for a whole slew of nontrivial phenomena from cognitive and developmental psychology and linguistics. This is not entirely a news flash – as the authors point out, they have built upon a strong foundation of past connectionist models in this area. R&M's novel contribution is the sweeping scale of the cognitive phenomena they simulate with a uniquely robust architecture. Indeed, their simulations suggest this is how we know what we know, and the basis for how we reason about the world.

The more I read, the more I wanted R&M to expand beyond their trees, flowers, birds, fish, and mammals, and predict when my children will stop believing in Santa Claus, or why Anselm of Canterbury believed existence was a virtue, thereby concluding God existed. These seemed well within the explanatory power of the model (the first deriving from competing error-driven learning, and the second from coherent covariation). In this commentary, I will limit my comments to another domain that did not immediately appear to be within the explanatory power of the model. This is the formation and persistence of delusions.

Delusions are beliefs that arise in the absence of evidence (they are self-evident to the believer and generally very personally poignant), and are resistant (though not immune) to invalidating evidence. Upon occasion, delusions can be quite fanciful and bizarre. A number of dementias lead to delusional beliefs, including some kinds of brain trauma, late-stage Alzheimer's disease and AIDS, Parkinson's disease, affective disorders, and my particular domain of expertise, schizophrenia. Although experimental evidence is sparse as to the nature of cognitive changes during the development of a delusional psychosis, clinicians' reports are generally consistent about the typical progression. A delusion begins as a vague *sense* of something, for example, that someone has malicious intentions toward me. Over the span of weeks or months, a delusion begins to crystallize; for example, the government is spying on me, then the FBI is following me, then the FBI is tracking my thoughts using an implanted device. Perhaps the FBI's reasons for this elaborate artifice also become clearer to me – "I'm a threat to the government" slowly morphs into "I've been chosen by God to bring down this heretical regime." Although delusions take on many forms, they are not *random.* They are characterized by personal significance and are emotionally charged. No one develops the delusion that popcorn comes from barley or that the pavement just happens to be made of worn-out carpet. Delusions experienced by depressed people are nearly always depressing; delusions experienced by manic people are nearly always expansive. Delusions appear to emerge not out of random associations, or even a random breakdown of associations. There is a landscape that appears to constrain the kinds of delusions that people will report (and presumably experience).

The landscape of one type of delusion is particularly constrained. Capgras delusions occur across several forms of dementia, including schizophrenia. In a typical Capgras case, the patient believes a loved one has been replaced by an identical imposter. One prominent hypothesis used to account for this phenomenon observes that visual form is processed by paths ascending to occipital cortex, whereas its emotional salience can be processed by subcortical paths that project directly to the amygdala (Morris et al. 1999). A Capgras delusion, therefore, may occur when the occipital pathway is intact (thereby allowing facial recognition), but the subcortical pathway is lesioned (denying the perceiver the usual emotional salience of the face) (Ellis & Young 1990; Frith 2004). There is some support for this conjecture:

Five psychiatric patients with Capgras delusions, in contrast to controls, showed no modulation of their autonomic responses to familiar and unfamiliar faces (Ellis et al. 1997). Indeed, across the different classes of delusions, there is reason to believe that an abnormal *feeling* drives an *a posteriori* cognitive explanation that is manifest as a delusion (Frith 2004).

It is not clear how the landscape of delusional susceptibility is manifest by the model described by R&M. In accounting for the possibility of "preparedness" to learn, they readily acknowledge the possibility that the strength of some pathways are not random and may have evolved stronger links (*Semantic Cognition*, p. 368). This is one way to incorporate emotional experiences. However, this brief consideration of the development of delusions, and evidence for the relationship between delusions and changes in emotional salience, is not clearly reconcilable within the current model. Given the boost to psychopathology research derived from "breaking" other connectionist models (e.g., Cohen & Servan-Schreiber 1992), many readers will welcome any and all efforts R&M might make to account for these distressing phenomena. How can this model be "broken" to make it delusional?

ACKNOWLEDGMENTS

Preparation of this commentary was supported by a McKnight-Land Grant Fellowship from the University of Minnesota and NIMH grant MH18269.

A crosslinguistic perspective on semantic cognition

doi:10.1017/S0140525X08005967

Asifa Majid and Falk Huettig

Max Planck Institute for Psycholinguistics, Nijmegen 6500AH, The Netherlands.

Asifa.Majid@mpi.nl

http://www.mpi.nl/Members/AsifaMajid

Falk.Huettig@mpi.nl

http://www.mpi.nl/Members/FalkHuettig

Abstract: Coherent covariation appears to be a powerful explanatory factor accounting for a range of phenomena in semantic cognition. But its role in accounting for the crosslinguistic facts is less clear. Variation in naming, within the same semantic domain, raises vexing questions about the necessary parameters needed to account for the basic facts underlying categorization.

Rogers & McClelland (R&M) set the ambitious goal of accounting for a wide array of experimental findings in semantic cognition. The ability of their simple distributed connectionist model to account for such a range of phenomena is impressive. However, the authors do not consider crosslinguistic naming data at all in their book, *Semantic Cognition* (Rogers & McClelland 2004). We believe that these data pose serious problems for the model in its current form.

One of the critical properties of the R&M model is its sensitivity to "coherent covariation" of features, for example, "has wings," "has feathers," "can fly" are features that coherently covary in birds. Sensitivity to higher-order inter- and intra-category features enables their model to exhibit some of the core characteristics of semantic cognition. For their model to be psychologically plausible, it is important to be able to demonstrate that there are indeed higher-order correlations in the world and that semantic categories straightforwardly map these correlations. In the examples that R&M consider, it seems likely that this could be so – plant and animal categories, for instance, appear to be highly constrained in the ways that the authors outline (see also Malt 1995). But for many other categories it is not clear that coherent covariation alone will account for the facts.

Consider event categories as an example. Recent cross-cultural research suggests that there may be coherent covariation among features of events, as there are for object categories. But despite this, the specific instantiation of categories found in different cultures vary substantially. Majid and colleagues (Majid et al. 2007a; forthcoming), for example, analyzed naming data collected from 28 typologically, genetically, and geographically diverse languages. Using correspondence analysis, Majid et al. (2007a; forthcoming) found that there were a small number of dimensions that accounted for the semantic categories of cutting and breaking across languages. The first, and most important of these dimensions, was a continuous one that distinguished events where the location of separation in the object was predictable from those where the location was unpredictable (roughly corresponding to "cut" and "break" events). But despite the fact that all languages loaded very highly on this, and the other, dimensions, individual categories from specific languages were very different.

To illustrate this, compare the sheer number of categories used for this domain from two of the languages of the sample – Tzeltal, a language spoken in the Highlands of Mexico, and Yélî Dnye, a language spoken on Rossel Island, an isolated island in Papua New Guinea. Tzeltal speakers used more than 50 different verbs to describe the 60-odd videoclips (Brown 2007), while Yélî Dnye speakers used only three (Levinson 2007). Or consider a more detailed contrast – Dutch, Swedish, and Mandarin have a specific category used for events of "cutting-with-scissors," whereas most other languages lump these events with other events of predictable separation.

How can it be that languages correlate very highly on the dimensional structure but vary so much in the specific categories that they have? One way to reconcile these findings is that the dimensions uncovered by the correspondence analysis are exactly those exhibiting coherent covariation among features. For cutting and breaking events in the real world, there is a tight correlation between the kind of instrument used, the object it is used on, the manner in which it is used, and the end state of the object. Sharp instruments, such as knives, are typically used with rigid but pliable objects, in a deliberate manner to achieve a clean separation. Across the board, languages are sensitive to these regularities, just as R&M propose. But R&M under-emphasize – and perhaps underappreciate – that different languages come to different conventionalized solutions about how to make reference to these constellations.

For example, English speakers have a hierarchical system of verbs that they can call upon when deciding how to name an event. The same event could be labeled *cut* or *slice*, *dice*, *chop*, or *break* and *snap*, *smash*, and so on. Swedish speakers, on the other hand, do not have this option open to them. In Swedish, cutting and breaking verbs appear to be organized in a flat structure. There are no general superordinate verbs like *cut* and *break*. As a consequence, Swedish speakers are more consistent namers for these events since there are fewer options for how to label a particular event in their language. English speakers are less consistent because there are equally good alternatives available (Majid et al. 2007b).

This kind of variation in naming highlights the complex interplay of world, concept, and word. Convergent evidence from domains as diverse as event representation (such as the data mentioned above), color (e.g., Kay & Regier 2003), and artifacts (e.g., Malt et al. 1999) suggest that names may indeed be constrained by perceptual attributes, but there is still much variation between languages. It is not clear how R&M's model will be able to account for this sort of variation.

R&M hold an ambivalent position about the role that language plays in categorization. On the one hand, they acknowledge that "experience with spoken language may play some role in concept acquisition prior to the infant's ability to produce speech" (*Semantic Cognition*, p. 145). On the other hand, they stress

that in their model "structure arises from the pattern of covariation of properties of objects, and does not depend on the explicit labeling of the objects" (p. 69). If categories are being created independently of linguistic experience, then how do R&M account for how children come to acquire the particular linguistic system of their community? If language is being used from the earliest phases of acquisition, then R&M need to be more explicit about how labels interact with "coherent covariation."

As we said at the outset, the range of phenomena that R&M tackle are impressive. In their own words, they wish to account for semantic tasks "that require a person to produce or verify semantic information about an object, a depiction of an object, or a set of objects indicated verbally (e.g., by a word)" (*Semantic Cognition*, p. 2). We have our doubts that a single model can really do it all, given that there is evidence that the representations underlying naming and object recognition are distinct. For example, Malt et al. (1999) have demonstrated that object naming and nonlinguistic sorting are only weakly correlated, suggesting that they are at least partially independent.

To summarize, R&M need to illustrate how their model can be appended to accommodate these findings or acknowledge whether another mechanism is needed.

Some suggested additions to the semantic cognition model

doi:10.1017/S0140525X08005979

Jean M. Mandler

Department of Cognitive Science, University of California San Diego, La Jolla, CA 92037-0515.

jmandler@ucsd.edu

http://www.cogsci.ucsd.edu/~jean/

Abstract: Rogers & McClelland (R&M) present a powerful account of semantic (conceptual) learning. Their model admirably handles many characteristics of early concept formation, but it also needs to address attentional biases, and distinguish direct input from error-driven learning, and fast versus slow learning. Not distinguishing implicit and explicit knowledge means that the authors also cannot explain why some coherently varying information becomes accessible and other information does not.

Semantic Cognition (Rogers & McClelland 2004) is an impressive book, with important lessons for researchers interested in how the development of a knowledge system takes place. Rogers & McClelland (R&M) show convincingly that even a simple feed-forward connectionist system can account for a number of aspects of nonverbal knowledge acquisition, such as global (superordinate) learning preceding more specific (basic-level) learning, how naming accelerates basic-level learning, and how generalization, even overgeneralization, takes place. These are fundamental characteristics of concept formation in infancy (Mandler 2004), and an algorithmic account of how it might be accomplished is a fine contribution. The emphasis on coherent covariation is particularly important because it provides a detailed account of the pattern learning that underlies much human knowledge. It also poses a serious challenge to strongly nativist views that require built-in domain-specific constraints. R&M make an excellent case that the structure of the environment itself, in conjunction with a domain-general learning mechanism, is sufficient to produce what superficially appears as domain-specific learning.

It must be noted, however, that the book veers more toward showing the usefulness of their learning model than to provide a realistic model of semantic (conceptual) development. R&M themselves note that distributed representations and fully recurrent networks would be more realistic, and that their simulations work on greatly oversimplified content and restricted contexts. However, they propose some basic developmental principles, and it is to these that I direct my comments. I offer a few suggestions to make their approach a more satisfying account.

First, there are known attentional biases that influence the course of learning. For example, infants are biased to attend to motion from birth. Indeed, it is a large part of what they do attend to in the first months of life. They often do not pay attention to objects' details (Bahrick et al. 2002), in spite of the fact that recognition data tell us those details are already becoming part of the perceptual knowledge system (Eimas & Quinn 1994). These are quite plausibly learned in the way R&M describe. Attentional biases likely affect concept formation more than they do perceptual learning, but in any case they are present in real life, which means that infants do not respond randomly even at the beginning of learning, as all the simulations in this book do. Their presence also means that some things are learned faster than others, not just because of learning how they covary with other attributes, but because of the attention that is paid to them from the start.

R&M show nicely how with coherent variation an attribute such as movement can be used to categorize unmoving objects. This is interesting vis-à-vis our experiments showing that infants correctly categorize static models of animals and non-animals (Mandler & McDonough 1993; 1998). But these real infants had many months of observing that animals move by themselves and non-animals do not. Motion is input, not error-driven learning. In the R&M simulations, infants do not see motion, but only learn to infer it. Unfortunately, this approach distorts what actually occurs in a learning infant. An observed property that consistently differentiates animals and non-animals should be learned very quickly and presumably faster than the way it covaries with other properties, even though such pattern learning eventually becomes important.

A related issue concerns how to relate the simulations of various tasks to developmental time. Simulations using distributed inputs found that the animal versus non-animal discrimination took about 2,500 epochs. This is a discrimination that takes place in the first few months of life (Quinn & Johnson 2000). How should we relate it to the differentiation into basic-level categories, which takes another year or longer? Albeit with a different task but using extremely simple inputs, the relevant simulation still took about the same number of epochs to differentiate birds from fish. It doesn't matter so much how epochs are interpreted as that they be proportional to the amount of time it takes to actually master coarse and fine distinctions. Some pattern learning, such as the covarying properties of animals and plants, does not appear in real life until around age 10 (if at all). Why? Does that imply hundreds of thousands of epochs are needed, or will a small number of well-distributed school instructions make the difference? The book wasn't designed to compare verbal and observational learning, but this is another aspect of fast and slow learning that needs consideration.

Another crucial issue, if we are to understand conceptual development, is why some coherently varying information becomes accessible and other information does not. Only some knowledge is available for conceptual thought. Although R&M distinguish perceptual and conceptual knowledge, their model does not do so; it treats all "semantic cognition" as alike. But self-motion and certain bodily parts become explicitly known, whereas much equally covarying perceptual information does not. For example, why do we know explicitly only a few things about what faces look like, but not the equally coherently varying spatial relations that differentiate men's and women's faces? We have a great deal of perceptual knowledge we use for recognition that remains unconceptualized.

R&M note that they need a second learning mechanism that takes in new information without interfering with established knowledge. This issue, previously discussed by McClelland et al. (1995) in terms of fast and slow learning systems, is not

only relevant to differences between verbal and observational learning, but also potentially relevant to explicit versus implicit knowledge. I regret that R&M did not address how fast and slow learning interact in development. In the relevant simulations new information was rapidly added to an existing knowledge base by a backpropagation-to-representation technique, but how this new information interacts with ongoing learning was not discussed.

Perhaps when a fast-learning system is integrated with the current model, it will help clarify what differs between forming accessible concepts and learning the coherent patterns that underlie many of them. It might also clarify how learning perceptual patterns comes to affect the attentional processes that influence concept formation. R&M's book is an excellent contribution, and hopefully these further difficult issues will be solvable within a connectionist framework.

Concepts, correlations, and some challenges for connectionist cognition

doi:10.1017/S0140525X08005980

Gary F. Marcus[a] and Frank C. Keil[b]

[a]*Department of Psychology, New York University, New York, NY 10012;*
[b]*Department of Psychology, Yale University, New Haven, CT 06510.*
gary.marcus@nyu.edu **http://www.psych.nyu.edu/gary/**
frank.keil@yale.edu
http://www.yale.edu/psychology/FacInfo/Keil.html

Abstract: Rogers & McClelland's (R&M's) précis represents an important effort to address key issues in concepts and categorization, but few of the simulations deliver what is promised. We argue that the models are seriously underconstrained, importantly incomplete, and psychologically implausible; more broadly, R&M dwell too heavily on the apparent successes without comparable concern for limitations already noted in the literature.

Rogers & McClelland's (R&M's) précis target article represents an important effort to bring explicit computational models to bear on questions of concepts and categorization; the sheer breadth of their demonstrations cannot fail to impress.

What, however, do R&M's demonstrations, presented here and in the book (*Semantic Cognition*, Rogers & McClelland 2004), really show? R&M's ambition is threefold: to reinterpret theory-based approaches to concepts, to suggest that domain-specific knowledge emerges solely from general learning principles, and (implicitly) to undermine symbol-manipulation in favor of parallel distributed processing (PDP)-style connectionism.

The success of the enterprise lies in the simulations, and here we have serious reservations. Although the models seem to address some core phenomenon in concepts or conceptual development, few fully deliver what is promised.

Take, for example, "category coherence" (sect. 2.2. of the target article) R&M's ostensible goal is to explain why "some groupings seem more natural, intuitive, and useful for the purposes of inference than others" (sect. 2.2, para. 1), but no matter how well the models track such correlations, they never really get to the crux of the matter: causation. Some properties are correlated in the world by accident, and others because of causal relations, a fact that is apparent even to young children (e.g., Greif et al. 2006); and people's generalization over different sets of "coherent" properties is powerfully mediated by their understanding of causal relations (Rehder 2003; Rehder & Kim 2006); R&M's model, in contrast, literally cannot represent the difference between a correlation of particular strength that is causal and one that is not.

More broadly, the demonstrations they present are (1) seriously underconstrained, (2) importantly incomplete, and (3) psychologically implausible.

1. *Constraints*: Each time R&M introduce a new phenomenon, they also introduce a new *model:* In the five core models of the target article, there are five different training regimes, five different architectures, five different dependent measures – and no effort at reconciling the differences.

2. *Completeness*: Although the current models excel at learning complex correlations between features, they fail to represent abstract operations over variables, structured representations, and contrasts between individuals and kinds; and it is not clear how well they can do any of these things in principle (Marcus 1998a; 1998b; 2001). In consequence, common two-place predicates such as X is a *sister* of Y, or P is a *parent* of Q cannot be properly represented in a fully generalizable way (Marcus 2001); complex notions such as *lectures about movies* versus *movies about lectures* cannot be represented without a combinatorial explosion of input nodes; and the models cannot represent basic distinctions such as the difference between the fact that dogs in general have four legs and that some dog in particular has three legs (Marcus 2001).

3. *Psychological plausibility*: Virtually all PDP models face certain fundamental problems, such as the slowness of their learning (here, acquiring a single fact can take thousands of trials), but the present models face a special set of problems, in terms of the way the learning task itself is defined. The experience of a real child can be thought of as a series of learning episodes that pertain to particular entities and whatever happens to be observable in a given moment: A child might see a particular dog, note some properties of that dog, and then update his or her internal representations. In R&M's framework, the learning experience is entirely different: In part because the model lacks a type-token distinction, the learning regime is designed such that the model *never* experiences any *particular* dog; all its experiences consist, instead, of pre-digested lessons about the properties of *all dogs*, whether or not those properties might be plausibly observed in any given moment. The model is not told "on a particular occasion, I see a dog, and on that occasion the dog is barking," it is told *dogs can bark;* moreover, it is given all such facts *simultaneously.* In Figure 1, the model is told – in one go – that all canaries can grow, move, fly, and sing. Real children rarely have it so easy; at any given moment a child might see a canary that is in the act of flying, without seeing it sing, or see it sing without seeing it fly; the child has to infer that what he or she sees at one moment doesn't preclude other possibilities later. The challenge of putting together observed and unobserved properties – which some might call the heart of categorization – is not adequately addressed. (Likewise, since the model trades entirely in abstractions rather than specific experiences, there is no way to capture, e.g., the human intuition that natural kinds have essences that can persist through drastic transformations; see Keil 1989.)

Still more broadly, perhaps the greatest problem with the current work is not so much the details as the enthusiastic reports of apparent success without comparable concern for limitations. Problems raised before (e.g., Gentner & Markman 1993; Marcus 1998a; 1998b; 2001; Pinker & Prince 1988), receive scant mention, and competing models such as Anderson and Betz (2001), Rehder (2003), Love et al. (2004), Nosofsky (1986), and Kruschke (1992) are never seriously discussed. There is no test, for example, of whether exemplar or ACT-R models would exhibit the same patterns, and hence no way of parceling out what predictions are unique to the present architecture and which would follow from *any* system that was sensitive to intercorrelations.

What we are left with? The simulations do acquire complex, learnable correlations between properties, but do not show that such intercorrelations alone suffice for human reasoning. In a true account of human reasoning, correlations may well serve as the *input* to cognition, but the end product is often far richer; we humans don't just notice contingencies, we seek to *understand* them. When a child sees that animals in cold climates

tend to have heavy fur, she doesn't just note the data; she asks *why*, and that is one thing PDP models just cannot do.

Finely tuned statistical engines of the sort discussed here may well play some role in our conceptual understanding of the world (Keil 1991b; Marcus 2000), but if the current work serves as a guide, such machinery seems unlikely to suffice on its own.

Analogy and conceptual change in childhood

doi:10.1017/S0140525X08005992

John E. Opfer[a] and Leonidas A. A. Doumas[b]

[a]*Department of Psychology, Ohio State University, Columbus, OH 43206;* [b]*Department of Psychological and Brain Sciences, Indiana University, Bloomington, IN 47405.*

opfer.7@osu.edu
adoumas@indiana.edu

Abstract: Analogical inferences are an important consequence of the way semantic knowledge is represented, that is, with relations as explicit structures that can take arguments. We review evidence that this feature of semantic cognition successfully predicts how quickly and broadly children's concepts change with experience and show that Rogers & McClelland's (R&M's) parallel distributed processing (PDP) model fails to simulate these cognitive changes due to its handling of relational information.

Rogers & McClelland (R&M) have presented a powerful response to the theory-theory of concepts (Carey 1985; Keil 1989; Murphy & Medin 1985), the view that knowledge of causal and other abstract relations among entities influences learning, memory, and reasoning. Against this view, in *Semantic Cognition* (Rogers & McClelland 2004) R&M have shown that an artificial learner (their parallel distributed processing [PDP] model) need not represent concepts within a theory to show many classic phenomena of cognitive development, including category coherence, context-sensitive generalization, conceptual reorganization, and the causal status effect.

Despite its ability to simulate these aspects of child cognition, R&M's approach has a fundamental limitation: Their PDP network does not process relational structure the way that children do. That is, within R&M's PDP model, relations (like *ISA*, *can*, and *has*) allow the model to learn the difference between being a predator, being capable of chasing, and actually having prey, much as children do when recognizing that a kitten, even when it hasn't chased any mice, is a miniature predator-to-be. So far, so good; but children, unlike a PDP network, can represent these relations and their fillers in a manner that preserves *relation-filler independence* (i.e., relations and their fillers are represented independently), while simultaneously representing the bindings between roles and fillers in an explicit and dynamic fashion. Thus, children can appreciate how "Fido chases Felix" is like "Felix chases Fido" (same elements involved in the same relation) and how they differ (role-bindings are reversed; e.g., Richland et al. 2006). This capacity requires (1) that relations and objects be coded with the same sets of units regardless of their specific configuration (i.e., the same unit[s] should code for the *chase* relation and for the object Fido regardless of whether Fido is chasing Felix or Felix is chasing Fido), and (2) that the system can create and destroy bindings dynamically. That is, it must be able to bind the units representing the chaser role of the chase relation to the units representing Fido (and explicitly encode that binding) when Fido is doing the chasing, and then bind the same units that represented the chaser role to Felix when Felix is doing the chasing.

Consequently, although the R&M model can simulate some important aspects of cognitive development, it fails to account for several developmental phenomena that entail relational reasoning, such as transitive inference and analogy. These capacities are important because they account for rapid and broad changes in semantic cognition, such as developing the *living thing* concept. For example, Opfer and Siegler (2004) have shown that children can quickly learn abstract categories like goal-directed agent by comparing goal-directed actions (e.g., cats turning toward mice, caterpillars turning toward leaves, and plants turning toward sunlight). Moreover, just as adults interpret ambiguous blobs turning toward goals to be living things (Opfer 2002; Schultz et al. 2004), kindergartners who learned that plants – like animals – are goal-directed also spontaneously induced (without feedback) that plants – like animals – are living things, too (Opfer & Siegler 2004). This zero-trial learning is inconsistent with the hundreds of epochs of direct training required by the R&M model. Further, errors that children actually make during learning – such as assuming that only animals are living things – are consistent with their idea that life requires some kind of goal-directed movement (normally visible only in animals), but this error is never made by the R&M model; moreover, errors made by the model – such as honoring a categorical distinction between sunflower/rose/robin/salmon versus sparrow/pine/flounder – have never been reported in the many studies investigating development of the living things category (for review, see Opfer & Siegler 2004). Thus, while R&M's PDP model can simulate feature-based learning of the living thing category, it does not actually simulate children's relation-based learning of the living thing category.

Children make analogical inferences such as those found in Opfer and Siegler (2004) because they can process relational structure. Relational structures allow us to make alignments between otherwise dissimilar systems (e.g., Gentner 1989; Holyoak & Thagard 1995) and to make inferences based on relational – rather than only featural – commonalities (Opfer & Bulloch 2007). Thus, having learned a predicate like *goal-directed agent*, children can align otherwise dissimilar objects (cats, potted plants) and generalize the properties of cats and other goal-directed agents (e.g., living-thing) to plants as well. These kinds of problems pose a difficulty for R&M's model precisely because it represents neither relations (e.g., goal-directed) nor relation-filler bindings explicitly. Consequently, R&M's PDP model cannot use relational information to drive inference (see also Hummel & Holyoak 2003).

A recent model by Doumas et al. (2008), called DORA, provides a solution to these problems. DORA is a connectionist model that, by virtue of its solution to the dynamic binding problem, can represent relations as explicit symbols that can take arguments. Starting with unstructured representations of objects as simple feature vectors, DORA learns explicit representations of object properties (and later relational roles) via comparison-based intersection discovery. These representations are effectively single-place predicates (represented as collections of nodes) that can be bound to arguments. DORA then links sets of these single-place predicates to form complete multi-place relations (where each of the linked predicates serves as a role of the relation). Importantly, these relational roles can be dynamically bound to arguments. Like its predecessor LISA (Hummel & Holyoak 2003), DORA uses time to carry binding information. Roles are bound to their fillers by systematic asynchrony of firing, where bound roles and fillers fire in direct sequence. For example, to bind Fido to the role chaser and Felix to the role chased, DORA will fire the units representing chaser followed by the units representing Fido, followed by the units representing chased, followed by the units representing Felix.

Unlike R&M's PDP model, successful models of semantic cognition must be able to learn explicit representations of properties and relations from examples and must bind these representations to novel arguments. By exploiting the strengths of structured relational thinking, successful models can make analogies based on common relations and thereby generalize over shared relations, just as children do when learning that, by virtue of being goal-directed agents, plants – like animals – are living things.

Time for a re-think: Problems with the parallel distributed approach to semantic cognition

doi:10.1017/S0140525X08006006

Philip Quinlan

Department of Psychology, University of York, Heslington, York, North Yorkshire YO10 5DD, United Kingdom.

ptq1@york.ac.uk

http://www.york.ac.uk/depts/psych/www/people/biogs/ptq1.html

Abstract: Rogers & McClelland (R&M) have provided an impressive outline of the capabilities of a class of multi-layered perceptrons that mimic many aspects of human knowledge acquisition. Despite this success, in the literature several basic issues are raised and concerns are expressed. Indeed, the problems are so acute that a different way of thinking is called for. In this commentary it is suggested that rational models approach provides a promising alternative.

In 1984 Hinton wrote, "Any plausible scheme for representing knowledge must be capable of learning novel concepts that could not be anticipated at the time the network was initially wired up" (Hinton 1984, p. 26). Despite the 400-plus pages on the wonders of the various kinds of multi-layered connectionist networks discussed by Rogers & McClelland (R&M) in *Semantic Cognition* (Rogers & McClelland 2004), we are faced with the rather difficult issue of plausibility mentioned by Hinton.

In arguing for the benefits of distributed representation, Hinton raised many issues regarding the limitations of traditional semantic networks. How does the store keep track of new items of information? How are new things added to the network? How is it that new things are wired up correctly? Such worries apply with some force to the connectionist models discussed by R&M. In every case, the network structure is fixed and because of this, it is the capabilities of very particular kinds of tightly constrained network architectures that are revealed. However, if it is accepted that learning a new thing corresponds to adding a new node, then Hinton's concerns about incorporating new nodes re-emerge. There is some discussion of learning the new fact that, "a sparrow is a bird" (Rogers & McClelland 2004, pp. 63–69), but where does the sparrow unit come from, and who is in charge of the wiring? Aligned to these concerns are others – sparrows do/may/will molt? (Rather unfortunately, there is also the recurrent tendency in this line of work to graft on new units when the situation calls for it: Need to simulate hierarchical naming? Okay, graft on "*ISA*-general," "*ISA*-basic," "*ISA*-specific" units.)

Perhaps we can side-step some of these issues. Novel distributed representations on the representation units can be introduced at any time during the network's development. So (maybe) we can do away with the item units and focus attention on the sorts of distributed representations that are captured by the networks. Do these embody *the* atoms of meaning? It seems that everyone agrees that, underlying all aspects of knowledge representation and knowledge acquisition, are the atoms of meaning. Variously, semantic primitives, semantic features, cogits, and semantic micro-features have been discussed. Many have struggled in attempting to specify exactly what such things might be. Some have attempted have to individuate these by providing verbal labels (e.g., *gotfrom*-plants), whereas others have conceded that this is probably too hard and proffer that such things may not map directly onto verbal labels. As a consequence, we are much better off with the notion of semantic micro-features – the undefinable essence of knowing. A difficulty with this latter view is that norming studies in which people are simply asked to generate properties for things are reasonably successful in accounting for what it is that we know.

Aside from the basic assumption, about the atoms of meaning, there is little consensus. Clearly, issues remain over what such atoms are and how they operate. If we are to adopt a mental chemistry approach to knowledge acquisition, then we can accept that there is a set of semantic atoms, combinations of which give rise to all kinds of complex concepts. The critical point is that only some but not all such atoms are needed to represent any given entity. By this view it is simply not true that the identical set of atoms is used to build the concepts of "pine," "robin," and "salmon," and so forth. Such a mental chemistry set provides a relatively simple way of thinking about the productivity of thought: "fins," "pink," "gills," "swims" – salmon; "feathers," "red chest," "sings" – robin. When I think about "robin," "not having fins" simply does not enter my head. The alternative is enshrined in the sort of theory underpinning the R&M model. Within their network models, all atoms are used to represent every concept. For a more considered argument as to why this kind of theory does not work, see Fodor and McLaughlin (1990).

We may also ask other basic questions about whether the R&M framework for thinking is useful in other ways. The evidence seems incontrovertible: humans are exquisitely sensitive to environmental statistical regularities. R&M use such evidence to propagate a connectionist manifesto for human knowledge acquisition. Nevertheless the evidence can be used in a different way. There is a different trend emerging and this allows us to countenance the idea that the sorts of learning processes embodied in the models favoured by R&M are radically different from the sorts of learning processes that are going on in humans. We may contrast connectionist counting machines with Bayesian statistical machines as discussed by rational modelers (Steyvers et al. 2006; Xu & Tenenbaum 2007a).

Indeed, it seems that by the rational models view both nativist and empiricist accounts of knowledge acquisition can be accommodated. To take an example from vision, the principle of grouping by proximity may reflect an innate tendency. Things close together on the retina reflect things close together in the real world, so the tendency is to group adjacent things as belonging to a single object. This sort of "prior," as embodied in the visual system, could reasonably reflect the adaptive history of the species. Reasoning about the visual world reflects both this sort of prior and estimates of probability built up over the individual's own developmental history.

There is little point in repeating more cogent arguments. The main intention here is to stress the emerging support for rational models. Such models provide a radically different view of human knowledge acquisition to that discussed by R&M (see Xu & Tenenbaum [2007b], and their provocative account of word learning). The supporting evidence fits rather uncomfortably with the idea of connectionist counting machines.

R&M have provided a very clear description of what various kinds of connectionist models of human knowledge acquisition can do. Given such an impressive list, it seems sensible to ask what it is that such models *cannot* do. Only by addressing these issues will progress be made. My concerns are such that my advice is to look elsewhere for the answers to the problems that remain. It seems to me that a more fruitful framework for thinking is provided by the emerging field based around rational models.

On the semantics of infant categorization and why infants perceive horses as humans

doi:10.1017/S0140525X08006018

Paul C. Quinn

Department of Psychology, University of Delaware, Newark, DE 19716.

pquinn@udel.edu

http://w3.psych.udel.edu/people/faculty/quinn.asp

Abstract: This commentary considers the issues of what should be taken as evidence for semantic categorization in infants and why infants display a surprising asymmetry in the categorization of humans versus nonhuman animals. It is argued that perceptual knowledge should be viewed as a potent source of information for semantic categorization, and that the

asymmetrical categorization behavior arises as a consequence of the frequency and similarity structure of experience.

Two comments on *Semantic Cognition* (Rogers & McClelland 2004) are offered. The first speaks to Chapter 4, "Emergence of Category Structure in Infants," and the second addresses Chapter 5, "Naming Things: Privileged Categories, Familiarity, Typicality, and Expertise."

What counts as semantic categorization in infants?. In Chapter 4, Rogers & McClelland (R&M) examine whether the concepts of infants are semantic, where semantic information is defined as information "not available more or less directly from the perceptual input" (*Semantic Cognition*, p. 2). To address this issue, R&M consider reports of global categorization by infants (Mandler & McDonough 1993; Pauen 2002a). Some readers like myself may have questions about the evidence and the definition.

The responding of the 11-month-olds in Pauen (2002a) and Mandler and McDonough (1993) should be revisited. R&M observe that these infants generalized habituation from the familiarized category instances to the novel instance of the familiar category, and dishabituated to the novel instance of the novel category. However, this was not the pattern of results reported in the original papers. The 11-month-olds dishabituated to the novel instance from the familiar category and the novel instance from the novel category, although more so to the latter. Mandler and McDonough interpreted these results as evidence that the infants recognized the perceptual difference between the novel and familiar instances from the familiarized category, and recognized that the novel instance from the novel category was from a novel conceptual category. But this interpretation may be questioned because the infants did not provide a critical behavioral signature of categorization, namely, equivalent responding to instances of the familiarized category. One might reply that there was positive evidence for categorization observed among Mandler and McDonough's 11-month-olds presented with planes versus birds, but that could have been carried by perceptual cues: planes with silver wheels and vertical tail fins versus birds with textured wings depicting ruffled feathers. Findings that 12-month-olds categorize animals versus vehicles based on texture differences are consistent with this suggestion (Smith & Heise 1992).

However, for argument sake, let us assume that the 11-month-olds had produced unambiguous evidence of categorization. Would that allow one to conclude that infant categorization was semantic? Given the controls in Pauen (2002a), performance may well have been influenced by knowledge that had accrued prior to the experiment. But this should not be surprising given that the infants would have had 11 months to utilize perceptual input systems and a general learning mechanism to acquire a database of information about objects in the world. *And that knowledge may be perceptual.* In the case of animals, stored data may include information about faces, coloring, skeletal appendages, a body shape bounded by curved contours, movement patterns, and species-specific sounds of communication.

This observation raises the question of whether conceptual information of the sort emphasized by Mandler (2000) and Carey (2000) (and subsequently in development by Gelman [2003] and Keil [1989]) should be deemed as necessary to conclude that semantic categorization has occurred. From this commentator's perspective, knowledge about perceptible parts and properties of objects can be semantic knowledge. One could not have much of a concept of cats, for example, without knowing what they looked like and what parts they had. Concepts must include perceptual information, or else they would not be very helpful. Even school children learning about biology must be able to recognize cats in various poses and contexts. It is hard to imagine how a child could even map more abstract attributes acquired through language (i.e., cats have cat DNA) onto their correct object referents without having category representations available from perceptual experience to serve as support structures. By this view, information about perceptual properties seems just as semantic as information about genetics (Hampton et al. 2007), and the way forward is to explain how perceptual and conceptual knowledge are integrated (not dissociated) to form mature concepts (Murphy 2002; Quinn 2004b).

The infant who mistook a cat for a person. In Chapter 5, R&M describe how increased experience with a particular category leads word learners to extend the label for that category to less familiar, similar categories, although not to less familiar, dissimilar categories. In a corresponding simulation, R&M trained a model with dog patterns appearing more frequently than other mammals. Early in learning, the model extended the label dog to goats and even robins, but not to trees. R&M also relate how increased experience with a category that has been linked with expertise acquisition leads to increased ability to differentiate within that category. A matching simulation demonstrated how birds (or fish) became more distinct for a model that was trained with birds (or fish) more frequently than other animals. Importantly, the different tendencies associated with increased experience, towards generalization and differentiation, are explained in a common framework in which the semantic space devoted to the more experienced class becomes larger than that allocated to less experienced classes.

The behavior of word learners and experts that is captured in R&M's simulations may also be observed in infants categorizing humans and nonhuman animals. In particular, 3- to 4-month-olds familiarized with humans generalize familiarization not only to novel humans, but also to cats, horses, and even fish, although they differentiate cars. By contrast, same-aged infants familiarized with horses generalize familiarization to novel horses, but differentiate humans, fish, and cars (Quinn & Eimas 1998). In addition, infants represent humans as subordinate-level exemplars, but represent nonhuman animals (i.e., cats) as summary-level information.

Consistent with R&M's approach, the differences in how infants represent humans versus nonhuman animals have been attributed to infants' greater experience with humans inclusive of repeated encounters with parents, siblings, and caregivers (Mermillod et al. 2004). This experience leads to a larger representational space for humans, thereby allowing individual humans to be represented at the subordinate level and enabling generalization to nonhuman animals. The trends toward increased specificity and generality resonate with findings that experts recognize instances from their domain at more specific levels than novices (Tanaka 2001), and are also better able to recognize commonalities across their domain (Murphy & Wright 1984). These observations in turn suggest a sense in which infants' knowledge of humans may constitute an initial domain of perceptual expertise (Quinn 2005). The specificity and generality effects are further in accord with the dual trends of differentiation and coalescence observed in the trajectory of learning produced in R&M's network simulations.

Subsequent research has suggested that infants' generalization from humans to nonhuman animals is rooted in holistic-configural structure, that is, a head attached to an elongated body with skeletal appendages (Quinn 2004a; Quinn et al. 2007), which is consistent with the finding that experts perceive objects within their domain holistically (Gauthier & Tarr 2002). Humans, cats, horses, and fish (but not cars), share this abstract resemblance, and R&M's networks demonstrated the ability to compute such global structure, extracting commonalities emerging in the pattern of coherent covariation. Notably, such structure is preserved even as the networks (and presumably infants) continue learning and begin differentiating subclasses. On this basis, generalization from humans to nonhuman animals by infants may form a foundation for the construction of a broad, domain-level concept of animal, and may also be a precursor of how children first rely on a human prototype to reason about biological knowledge (Carey 1985). Thus, in accord with R&M, differences in the frequency of exposure to different classes (humans vs. nonhuman animals),

coupled with the similarity structure of those classes (attributes shared by humans and nonhuman animals, but not cars), are important determinants of the growth of category organization during early cognitive development.

ACKNOWLEDGMENTS

Preparation of this commentary was supported by NIH Grants HD-42451 and HD-46526. The author thanks Gregory L. Murphy for helpful comments on an earlier draft.

The development of modeling or the modeling of development?

doi:10.1017/S0140525X0800602X

David H. Rakison[a] and Gary Lupyan[b]

[a]*Department of Psychology, Carnegie Mellon University, Pittsburgh, PA 15213;* [b]*Department of Psychology, Cornell University, Ithaca, NY 14850.*

rakison@andrew.cmu.edu **http://www.psy.cmu.edu/~rakison/**

il24@cornell.edu **http://www.cnbc.cmu.edu/~glupyan/**

Abstract: We agree with many theoretical points presented by Rogers & McClelland (R&M), especially the role of domain-general learning of coherent covariation. Nonetheless, we argue that in failing to be informed by key aspects of development, including the role of labels on categorization and the emergence of constraints on learning, their model fails to capture important features of the ontogeny of knowledge.

The book *Semantic Cognition* by Rogers and McClelland (2004) is an elegant demonstration that a simple parallel distributed processing (PDP) model can exhibit behavior that matches the behavior found in a range of empirical studies on infants' conceptual development. As such, Rogers & McClelland (R&M) make a compelling case that domain-general, rather that domain-specific, mechanisms that are sensitive to lower- and higher-order covariation underpin early concept formation. Although we concur with many of the authors' claims and their general theoretical perspective, in this commentary we propose that R&M have overlooked a number of key points about development which are crucial to consider in modeling early concept formation.

An important aspect of early concept learning overlooked by R&M is the role of verbal labels. Labels affect categorization and concept development in infants as young as 9 months (e.g., Balaban & Waxman 1997; Xu 2002), and their effect continues to grow in the subsequent months (e.g., Fulkerson & Haaf 2003; Nazzi & Gopnik 2001; Waxman & Markow 1995). Thus, labeling may be an important additional mechanism by which infants construe semantically related items as similar to one another in the absence of observable similarities.

Unfortunately, the role of labels cannot be investigated in R&M's network because they are implemented as simple stimulus features (the *ISA* relation). In our opinion, it is erroneous to implement basic-level category labels as features akin to having wings or barking. The principle of coherent covariation gains traction because the feature *can move*, for example, is informative in that not all entities can move and that being able to move predicts other properties. Labels are different: Many things from different semantic categories can move, but only canaries are canaries. From this perspective, the category label is the piece of information that varies most coherently and is most predictive of the item's category.

To explore the consequences of labels on concept formation, a model needs to map multiple exemplars (e.g., many different canaries), to a single label. In the process of learning to associate a single label with multiple category exemplars, the label becomes strongly associated with features most predictive of the category (Lupyan 2005) providing the "glue" that may be necessary for cohering together items from categories with high intra-category variability (Lupyan, in press). Thus, rather than adding a simple feature, labels can be thought to schematize a given stimulus by placing it into a relationship with the other members of the category.

An additional concern relates to the fact that humans, and especially human infants, demonstrate clear limits on learning, whereas connectionist networks are capable of learning essentially any pattern of inputs (Massaro 1988). This point is overlooked in two ways by R&M's model. First, in a number of simulations the model is able to show patterns of behavior that match those of infants only after it receives a level of experience that is unavailable in the real-world or the laboratory setting. The model, for example, has only begun to differentiate conceptually the input stimuli after 50 epochs, but by this time the network has been exposed to over 50,000 trials (Siegler 2005).

Second, and more important, R&M expose the model to all of the covarying input at the same time, yet infants are limited in the amount and kind of correlated information they can process. Before approximately 7 months of age, for example, infants are unable to encode relations among static features (Younger & Cohen 1986), and it is not until around 14 months of age that they can encode object features or whole objects with dynamic motion-related cues such as *can fly* or *can walk* (Rakison 2005). That infants are unable to process certain kinds of information constrains concept learning, but, at the same time, it also facilitates concept learning; that is, it allows infants to learn about more fundamental aspects of things in the world while at the same time ignoring other aspects. R&M's model, in contrast, is exposed simultaneously to a wide range of information which in an infant would probably lead to what William James (1890) called a "blooming, buzzing confusion." R&M argue that they used input features that they consider to be important or salient to infants, but in our view this approach disregards a large database of empirical data that shows to which features infants actually attend in developing concepts (see Madole & Oakes 1999).

Finally, the architecture of R&M's model is sufficiently flexible and powerful to demonstrate learning for any input pattern. Fitting a PDP model to existing data is not the strongest test of the theory advocated by the model (Roberts & Pashler 2000); more powerful support for the theory behind the model is to generate novel predictions that are borne out by empirical studies. Moreover, from our perspective any model that tries to emulate a set of empirical findings with infants or children must take developmental issues into account. We have recently developed such a PDP model for early concept formation that is theoretically compatible with that of that of R&M, but that incorporates development in a number of plausible ways (e.g., increasing over time the number of hidden units and reducing over time the weight-decay parameter of fast but not slow learning links) (Rakison & Lupyan, in press). This developmentally oriented model exhibits behavior that is unintuitive but nonetheless matches that found in infants. For example, 14-month-olds learn relations in simple causal events that are consistent and inconsistent with the real world (e.g., agents possessing moving or static parts), but 16-month-olds demonstrate constraints on learning by failing to learn the inconsistent events (Rakison 2005). From our perspective it is necessary for models to be informed and compatible with key developmental findings and issues if traction is to be made in determining the origins, nature, and development of concepts.

Semantic redintegration: Ecological invariance

doi:10.1017/S0140525X08006031

Stephen E. Robbins

Center for Advanced Product Engineering, Metavante Corporation, Milwaukee, WI 53224.

Stephen.Robbins@metavante.com

stephenerobbins.com

Abstract: In proposing that their model can operate in the concrete, perceptual world, Rogers & McClelland (R&M) have not done justice to the complexities of the ecological sphere and its invariance laws. The structure of concrete events forces a different framework, both for retrieval of events and concepts defined across events, than that upon which the proposed model, rooted in essence in the verbal learning tradition, implicitly rests.

There is no cognition without memory, that is, without the redintegration of events. In *Semantic Cognition* (Rogers & McClelland 2004), the depth and breadth of thought by the authors on semantics is impressive, but I fear the claim that their model dwells in the perceptual/ecological sphere ignores the realities of this sphere, resulting in an inadequate theory of redintegration.

Consider an event: stirring coffee in a cup, using a spoon. The event has a time-extended invariance structure, here defined as the transformations and invariants specifying an event and rendering it a virtual action. The swirling coffee surface is a radial flow field. The constant size of the cup, should it move forward or backward, is specified, over time, by a constant ratio of height to the occluded texture units of the table surface gradient. The tau ratio defined over this flow field supports modulating the hand for grasping the cup (Savelsbergh et al. 1991). Were the cup cubical, its edges and vertices are sharp discontinuities in the velocity flows of its sides as the eyes saccade, where these flows specify the form of the cup (cf. Robbins 2004; 2007). The periodic motion of the spoon is a haptic flow field that carries an adiabatic invariance – a constant ratio of energy of oscillation to frequency of oscillation (Kugler & Turvey 1987). The action of wielding the spoon is defined by an inertial tensor describing the moments of force (Turvey & Carello 1995). It is this entire informational structure and far more that must be supported, globally, over time, by a neural network – by the resonant feedback among visual, motor, auditory, even prefrontal areas. To define the "summary sketch" or compressed format, stored in the hippocampus, of this dynamic, ever changing, global pattern (McClelland et al. 1995) would be interesting.

In Rogers & McClelland's (R&M's) formulation, I present an occurrence, SPOON, in the context, CAN, and the network is trained, via weight adjustments, to respond with STIR. In the causal version, the network is trained to predict the sequelae – the "circular motion of the coffee liquid," the clinking sound, and so forth. What sense is this? This is a remnant of the supremely non-ecological verbal learning tradition, its roots in the semantics-eradication program of Ebbinghaus, which ultimately bifurcated events arbitrarily into components (now feature vectors), for example, SPOON and COFFEE, then asked: How do we learn these components as a paired-associate pair? This is the paired-associate (PA) learning framework of the older cousin-model (McClelland et al. 1995). Paivio's (1971) introduction of imagery, later the elaboration techniques, where SPOON and COFFEE are imagined in a dynamic event, were the first near-ecological cracks in this brute force, syntactic learning framework. Coherent covariation is essentially a low-order, syntactic invariance, insufficient to carry the form of invariance structure described above.

In reality, we are perceiving the spoon as an integral part of a stirring event, with all the event's ongoing invariance structure. Where is the "error?" We need no weight adjustments to "link" the event "components." A spoon scooping and lifting oatmeal is yet another event with an integral and complex set of forces and auditory/visual patterning. A spoon stirring pancake batter is yet a different complex invariance structure. A spoon digging into and cutting grapefruit is yet another. A spoon balancing on the edge of the coffee cup another. Now we have the set: SPOON CAN: [stir, balance, scoop, cut]. If I re-present SPOON, which event will it redintegrate? Presenting SPOON is roughly equivalent to a static event, a resting spoon. There is little structure to this cue-event; it is underspecified, yet common to all. It is like sending an imprecise reconstructive wave with little coherence through a hologram – we reconstruct a composite image of multiple recorded wave fronts, in this case, spoon-related events. It is the classic interference of McGeoch (1942). The brain's neural network does not require error-training to specify this set.

But we can retrieve specific events. The cue must bear the same invariance structure or a sufficient subset. For the coffee stirring, we re-present an abstract rendering of the coffee's radial flow field, or simulate the inertial tensor of the wielding. We are creating, globally throughout the brain, a more constrained, more coherent "reconstructive wave." To reconstruct the batter-stirring event as opposed to the coffee-stirring, we must constrain our wielding cue differently to capture the larger amplitude of the stirring motion and/or the greater resistance of the batter. In essence, we are discussing a "paired associate" paradigm, call it A-B/C/D, that is impossible in the verbal learning realm, for every cue is in effect SPOON, yet with the appropriate ecological, dynamic constraints on a cue event involving a spoon, we can reconstruct each separate event (Robbins 2006a; 2006b). But we also understand, "A SPOON can CATAPULT (a pea)," for the spoon, we know, can be inserted in, and supports the forces/invariance structure of, catapulting. This is precisely the realm of French's (1990; cf. Robbins 2002) devastating critique regarding the Turing Test – we would need to train for all possible object pairs – and the network proposed is no better equipped for these emergent analogies; it is simply not capturing the source of the semantics. It might be held that some analogy program such as Structure Mapping Engine (SME) (Gentner 1983) must now take over, but these must now define, ad hoc, the relevant "features" of catapult and spoon which are then merely algorithmically related (French 1999; Robbins 2002), again precisely because these systems ignore the ecological invariance defining events.

Ecological invariance forces an entirely different principle of retrieval for events and higher-order invariance (classes) across events. Discovering invariance laws, it is heavily argued (Kugler & Turvey 1987; Wigner 1970; Woit 2006; Woodward 2000; 2001), *is* scientific explanation. This subsumes the "causal" explanations of theory-theory. In this, science models itself after the brain in perception, and in this sense R&M, in holding, contra theory-theory, that there is no special mechanism, are perceptively right. But the neural dynamics required for the transformations and invariance noted here is beyond the coherent covariation detection of the model; the lesson of invariance in the ecological sphere has yet to be engaged by either side.

Reading *Semantic Cognition* as a theory of concepts

doi:10.1017/S0140525X08006043

Jesse Snedeker

Department of Psychology, Harvard University, Cambridge, MA 02138.

snedeker@wjh.harvard.edu

http://www.wjh.harvard.edu/~lds/index.html?snedeker.html

Abstract: Any theory of semantic cognition is also a theory of concepts. There are two ways to construe the models presented by Rogers & McClelland (R&M) in *Semantic Cognition*. If we construe the input and output representations as concepts, then the models capture knowledge acquisition within a stable set of concepts. If we construe the hidden-layer representations as concepts, the models provide a simulation of conceptual change.

The primary goal of *Semantic Cognition* (Rogers & McClelland 2004) is to illustrate how connectionist models can provide an explicit theory of the nature and development of semantic processing. But any account of semantic processing is also, necessarily, a

theory of concepts (see Fodor [1998] for discussion). What theory of concepts is implicit in the Rogers & McClelland (R&M) models? Answering this question might allow us to align these models with the rich theoretical literature on the nature of concepts (see Laurence & Margolis [1999] for introduction). There are two ways of thinking about these models as conceptual theories: We can think of the nodes in the input and output layers as concepts, or we can think of the patterns of activation in the hidden layers as concepts.

If we think of the input and output layers as concepts, then these models propose a theory with three sets of concepts that exist prior to the learning process that is being modeled: item concepts which are nodes in the input layer for basic-level kinds (e.g., *salmon*), property concepts in the output layer (e.g., *pretty*), and relation concepts in the input layer that link kinds and properties (e.g., *can*). The model is trained on propositions composed of these concepts and learns to make inferences about other propositions based on knowledge it acquires. On this construal, these are models of knowledge acquisition, not conceptual change, since the set of concepts is defined before training begins and is not altered by the training.

The nature of these concepts varies across the different instantiations of the model. The localist model (*Semantic Cognition*, Chs. 2–4) instantiates a theory of concepts that is strikingly similar to Fodor's (1998) atomistic theory. The concepts in the input and output layer have no internal structure. Their content does not come from the propositions in which they appear – that knowledge appears later and is formulated over the concepts. Instead, the identity of each concept is based on its relation to the world (Fodor's nomothetic relation). The node *pine* means pine solely because it is activated when the model represents pine trees. Its identity is fixed by the training stimuli, which are meant to represent the perceptual experiences of the learner. For example, a training stimulus such as "pine has bark" corresponds to a situation in which the child sees a pine, represents it as a pine, notices the bark, and represents it as bark. Thus, the input to the model presupposes that the learner already possesses concepts that allow her to categorize objects at the basic level and properties at a fairly abstract level (e.g., *can move*).

The distributed version of the model (Ch. 4) instantiates a different theory of concepts. Properties and relations are still modeled as conceptual atoms, while items are modeled as sets of perceptual features. For example, pine is represented by setting the units corresponding to big, green and branches to 1, and all other units to 0. This is simply a connectionist instantiation of the classical theory of concepts; each kind of item is defined by a set of necessary and sufficient features which are (arguably) perceptual in nature.

The alternative is to construe the patterns of activation in hidden layers as concepts. This appears to be what R&M have in mind (see *Semantic Cognition*, pp.140–141). On this construal, the input and output layers are preconceptual primitives that link the conceptual representations in the hidden layer to the world, thus lending them content. The pattern of activation in the *Representation* layer when an item node is activated is the concept for that item. Because these patterns of activation change throughout training, under this construal the models are simulations of conceptual change.

On this interpretation, the model bears a family resemblance to two types of conceptual theories. First, like theory-theories, the model grounds the meanings of some concepts in their relations to other concepts. In particular, the natural kind concepts gain their content from their relations to the predicates created by combining properties and relations. Furthermore, the relative significance that is assigned to each predicate depends largely on the intercorrelations between them. However, in another respect these hidden-layer concepts are similar to prototype theories. In prototype theories, the set of possible concepts is defined by the set of features and the set of possible weights for each feature. In these models the set of possible hidden-layer concepts is a function of the primitives in the *Item, Relation*, and *Property* layers. The model differs from the prototype model in that the function for combining these features is considerably more complex.

On this construal, what knowledge is necessary for acquiring the concepts in the hidden layer? In both the distributed and localist versions of the model, every instance of a basic-level kind receives the same input representation. As we noted earlier, this amounts to the claim that we categorize items and properties in precisely the right way, prior to acquiring the knowledge in the training set. In other words we must have the concepts to learn them (see Fodor 1998). This constraint is relaxed in the localist model in Appendix C, in which two items (cats and dogs) are each represented by five different nodes (e.g., five individual cats). This manipulation does not affect the organization of the hidden layer, leading the authors to conclude that pre-categorization is not necessary for acquiring these concepts. This conclusion seems a bit premature: most items in this simulation were pre-categorized (19 of 21), and all of the relations and properties were. Thus, it is not clear that the relevant structure could be unearthed if all the input nodes represented individual entities or particular instantiations of a property.

One alternative is to argue that the pre-categorization in the input and output nodes is based on perceptual processing rather than conceptual processing. The coherence of this position depends upon arriving at a clear definition of the distinction between conceptual and perceptual processes, and demonstrating that the primitives in the input and output layers can be defined in purely perceptual terms.

Agency, argument structure, and causal inference

doi:10.1017/S0140525X08006055

Alice G. B. ter Meulen

Center for Language and Cognition, University of Gröningen, 9700 AS Gröningen, The Netherlands.

a.g.b.ter.meulen@rug.nl

Abstract: Logically, weighting is transitive, but similarity is not, so clustering cannot be either. Entailments must help a child to review attribute lists more efficiently. Children's understanding of exceptions to generic claims precedes their ability to articulate explanations. So agency, as enabling constraint, may show coherent covariation with attributes, as mere extensional, observable effect of intensional entailments.

Three theoretical concerns regarding parallel distributed processing (PDP) modeling keep puzzling a logically minded cognitive scientist, if it is, as Rogers & McClelland (R&M) claim, designed to provide simulations of the acquisition and deterioration of human semantic cognition.

1. The weighting or strengthening of network relations in the base architecture (see *Semantic Cognition*, Rogers & McClelland 2004, pp. 115, 356) is presumably a linear relation. It is hence transitive, that is, if x is heavier or stronger than y, and y heavier or stronger than z, then x must be heavier or stronger than z. But an arbitrary similarity relation does not need to be transitive, for a canary may be similar to an orange in color, and an orange similar to the moon in being round, but a canary is not therefore similar to the moon. If such similarity judgments constitute the basis for inductive generalizations and clustering (*Semantic Cognition*, p.183), do PDP networks require similarity to be strengthened to a transitive relation? Perhaps an item may be simultaneously included in two distinct clusters? To give a particular illustration of this problem: If the representation space in Figure 3.9 (*Semantic Cognition*, p. 112) were also to

represent the color pink, how could it incorporate both *salmon* and *rose* in one shaded region?

2. According to *Semantic Cognition* (p. 182), *basic* names indicate the "labels that identify an object at an intermediate level of specificity" (e.g., *bird*), and *general* names "identify an item at a more inclusive level" (e.g., *animal*), and *specific* names "identify objects at a more specific level" (e.g., *canary*). So obvious logical entailment patterns should be validated; for example, if every bird is an animal and every canary is a bird, then obviously every canary must be an animal. Clearly, the list of attributes is highly structured by such logical entailments, which presumably facilitates children's review in allocating weights to attributes. Is there any evidence to assume that all the listed attributes are independently reviewed in assigning weights to connections, as suggested by linearly listing all attributes in PDP networks and requiring exhaustive review of these in attributing weights (e.g., the longer the list, the longer it takes to allocate weights)? Don't children quickly detect such valid logical patterns and use them in facilitating their review of attributes, skipping large sets of labels that don't apply simply because they depend on a label that is already known not to apply? In other words, doesn't *logical* knowledge help children in speeding up attribute weighting by making their review more efficient?

3. From an epistemological point of view, generic knowledge is special, "immunized" information. It is preserved in changing worlds and throughout a variety of contexts in virtue of a set of associated, generally accepted exceptions, that are prevented from serving as counterexamples falsifying the generically quantified statement (Carlson & Pelletier 1995). In learning to complete the similarly generic claim *canary can*..., any child will quickly exclude dead, frozen, or wind-up canaries in assigning a heavy weight to the attribute *move.* Having learned to treat a wind-up canary as an exception to the generic claim that canaries can move, the child applies the abstract notion of agency to anything capable of self-determined action (cf. *Semantic Cognition*, p. 131–33). If agency is considered to be one of Gelman's enabling constraints (*Semantic Cognition*, p. 133), then it seems undeniably to play a core role in concept formation at a pre-lingual and hence pre-explanatory phase of development. In the similarity-based learning algorithm of bootstrapping syntax (van Zaanen 2001; 2002), the child learns that if x moves y, then x causes y to move, and accordingly, y has been moved by x, and that therefore y has moved, but also that the reverse entailments do not hold. The causal source of the movement is consistently assigned as thematic role to the subject argument in either the intransitive or active transitive use of the verb *move*, as it is part of its lexical semantic content at least in ordinary English. As the child also knows that agency is required for self-movement, he or she would probably, at a linguistically adequate phase of development, be able to explain that a dead or wind-up canary, which he or she has observed to have moved from one location to another, must have been moved there by someone else. Such causal inference appears in no way logically prior to linguistic acquisition of the corresponding lexical items, nor does it seem to play a role similar to the logical relations alluded to above in structuring the set of attributes. But to conclude from such lexicalization of attributes that "Infants' sensitivity to a property, as indexed by attention, ease of learning, and other measures, is affected by the extent of its coherent covariation with other properties" (*Semantic Cognition*, p. 135), seems theoretically a bridge too far, although coherent covariation must be an extensional, observable counterpart to any obviously intensional logical entailment.

It is an interesting, and to my knowledge, novel claim that the PDP networks should also simulate semantic dementia effects. For instance, in overextending basic level names (*dog*), the model is claimed to be more likely to apply a highly familiar name incorrectly than a less familiar name to similar items (*Semantic Cognition*, pp. 215–17). It is well known that the familiarity of names is directly correlated with their early acquisition and hence retention in early memory, rather than with their frequency in recent usage (ter Meulen 2004). It would constitute an interesting new dimension in the empirical predictive power of PDP simulations, if time were an explicitly represented parameter. Consequently, the weighting should be an overtly dynamic relation, instead of a static time-slice model. In dynamic semantics a plethora of detailed representation systems have been developed that could well serve as inspiration for designing such overtly dynamic PDP systems (Kamp & Reyle 1993; Lascarides & Asher 1993; ter Meulen 1995; 2000; 2006; van Benthem & ter Meulen 1997).

Authors' Response

A simple model from a powerful framework that spans levels of analysis

doi:10.1017/S0140525X08006067

Timothy T. Rogers[a] and James L. McClelland[b]

[a]*University of Wisconsin-Madison, Department of Psychology, Madison, WI 53706;* [b]*Center for Mind, Brain and Computation, and Department of Psychology, Stanford University, Stanford, CA 94305.*

ttrogers@wisc.edu http://concepts.psych.wisc.edu
mcclelland@stanford.edu http://psychology.stanford.edu/~jlm

Abstract: The commentaries reflect three core themes that pertain not just to our theory, but to the enterprise of connectionist modeling more generally. The first concerns the relationship between a cognitive theory and an implemented computer model. Specifically, how does one determine, when a model departs from the theory it exemplifies, whether the departure is a useful simplification or a critical flaw? We argue that the answer to this question depends partially upon the model's intended function, and we suggest that connectionist models have important functions beyond the commonly accepted goals of fitting data and making predictions. The second theme concerns perceived in-principle limitations of the connectionist approach to cognition, and the specific concerns these perceived limitations raise for our theory. We argue that the approach is not in fact limited in the ways our critics suggest. One common misconception, that connectionist models cannot address abstract or relational structure, is corrected through new simulations showing directly that such structure can be captured. The third theme concerns the relationship between parallel distributed processing (PDP) models and structured probabilistic approaches. In this case we argue that there the difference between the approaches is not merely one of levels. Our PDP approach differs from structured statistical approaches at all of Marr's levels, including the characterization of the goals of cognitive computations, and of the representations and algorithms used.

Our book, *Semantic Cognition: A Parallel Distributed Processing Approach* (Rogers & McClelland 2004), has provoked a wide range of reactions, ranging from supportive suggestions for new domains to address, to friendly amendments of aspects of our proposals, to points of criticism, either of specific elements of our argument or of the overall approach that we have taken. We thank the commentators for engaging seriously and thoughtfully with our work. We are gratified that the range of replies is broad enough to raise most of the reactions we have encountered as we have presented our work to different audiences.

We wish to address both the specific challenges, misunderstandings, queries, and criticisms raised by the various

commentators, as well as the core issues underlying their alternative perspectives on our theoretical framework. To these ends, our response is organized around three themes, each addressing specific issues raised in the commentaries, and each touching on more general issues regarding the connectionist enterprise. These include (1) a discussion of how computational models should be used to support cognitive theories, (2) a treatment of putative limitations on the scope of our theory, and (3) a consideration of the relationship between our theory and a range of approaches cast in terms related either to Anderson's (1991) "rational analysis" or Marr's (1982)"computational" level of analysis.

R1. What is the relationship between a model and a theory?

We begin by considering a set of questions raised by commentators that might seem at first glance to be fundamental, but which, we will argue, are best viewed as concerns about particular details of specific models. For instance, **Snedeker** wonders whether our use of localist representations in many of the models amounts to endorsing Fodor's claim that concepts are innate; **Quinlan** takes us to task for adding new units to the model in order to represent additional items in the environment; **Marcus & Keil** suggest that we use different models or measures of model performance to substantiate each theoretical claim; **Rakison & Lupyan**, **Robbins,** and others suggest that our training regime is unrealistic in ways they see as important; and **Kropff & Treves** argue that our feed-forward models, which do not have attractor dynamics, cannot explain important nonlinearities in children's conceptual development. None of these concerns, we believe, touch on criticisms of our actual theory – instead, they focus on less central details of particular pragmatic choices we made when devising simple model instantiations of the theory.

Of course, any such claim might justifiably be met with skepticism: It is very easy to attribute the strengths of a model to its core characteristics, and the limitations to superficial simplifications! How can the neutral observer determine whether any given aspect of a model really represents a critical theoretical claim or an unimportant implementational detail?

In answer to this question, we find it useful to invoke an analogy employed by Jorge Luis Borges (1998) in a short fiction entitled "On the Exactitude of Science." Borges tells of an empire in which cartographers strive for such perfect fidelity that the maps they create are built exactly to scale. The resulting products are a marvel of accuracy but, as maps, are essentially useless: The journey one would have to undertake to discern the distance between any two cities, for example, would be equally arduous in the map as in the real world. This satire illustrates an important and, we think, oft-neglected fact about the relationship between models in science and the systems they are modeling. Typically those systems are complex – perhaps too complex to understand in their entirety without aid. Faithful replication of the full system is, therefore, rarely if ever the goal of a modeling enterprise in science. Should such a program succeed, it might prove as difficult to understand the model as to understand the system itself! Rather, the function of a model is to simplify – to remove some of the complexity of the full system, and even to violate some of its known properties, so as to reveal more clearly other important characteristics.

This does not mean, of course, that all departures from the complex system are equally tolerable: It certainly seems fair to criticize the map that situates Paris closer to New York than to Marseilles. The question then arises, which departures from reality constitute helpful simplifications and which represent serious flaws? The answer to this question depends upon the model's intended function – that is, which aspects of the real, complex system the model is intended to illuminate, and how the theorist uses it to reason about the real system. The cartographer knows that the Earth is not small, or flat, or made of paper – he purposely sacrifices those elements of the real system so that his model better realizes its intended function (i.e., to facilitate an understanding of the spatial relationships among important landmarks in some region of the Earth). Because the mapmaker intends distances between points on the map to be proportional to distances among analogous points on the Earth, criticism of a map on these grounds is justified.

Thus, to evaluate whether criticisms of our models have important implications for our underlying theory, it is important to be clear about the intended function of the models we have presented. Often in cognitive science, it is understood that the primary function of a model is to fit data and to make predictions that can be tested empirically; but there are at least three less commonly acknowledged functions of models that have guided our work, which we believe to be equally important.

First, a model allows the theorist to investigate and better understand the implications of core theoretical principles. In our work, these include principles general to the parallel distributed processing (PDP) approach to cognition articulated in Rumelhart et al. (1986b; see especially Rumelhart & McClelland 1986) and developed extensively over the past two decades, as well as additional principles specific to our theory of semantic cognition which are described in Chapter 9 of *Semantic Cognition* and summarized in the précis. They hold, among other things, that mental representations are patterns of graded activation values across simple processing units, rather than discrete symbols; that these patterns influence each other through mappings encoded in connections among the elements that participate in the representations; that learning involves graded changes in these mappings; that the architecture of the semantic system fosters convergent influence of all kinds of information on a common underlying representation; that learning within this system must be gradual and interleaved; and so on. An important function of the simulation work is to provide a tool for better understanding the implications of these core principles. It goes without saying that these implications are complex and non-obvious. Even very simple models like the Rumelhart model can exhibit counterintuitive behaviors. Computer simulations with such models allow the theorist to figure out how and why these counterintuitive behaviors occur, and what implications they then have for the ability of the theory to explain particular aspects of human behavior.

Second, the models help to demonstrate the sufficiency of the theory to address the phenomena it purports to

explain. Suppose we had simply stated that a domain-general learning process based on the principles sketched in the preceding paragraph will give rise to a progressive differentiation of conceptual knowledge; that it acquires general before basic information about particular concepts but exhibits a basic-before-general trend in naming these same concepts; that it can explain why some categories are useful and informative and others not, why different properties become important for different semantic domains, how the semantic system comes to "ignore" irrelevant perceptual similarity and "hone in on" similarities and differences important for semantic generalization; and so on. None of this would have been believable, no matter how clearly articulated. The simulations provide a way of demonstrating that, in fact, an instantiation of the principles of the theory can lead to the consequences described. An actual model system really does exhibit the behavior we claim in the verbal articulation of the theory. Of course, extension of the model to more complex situations must also be addressed, but the initial demonstration in a simplified instantiation is important to counter claims that the approach cannot possibly work in principle.

Third, the model serves as a didactic tool for explaining to others how and why the relevant phenomena arise. Just as the simpler models make it easier for us, the researchers, to trace out the implications of certain assumptions, so too do these models aid our ability to communicate the important points to the reader. For instance, we might have begun our report of this work with a description of the most complicated model in *Semantic Cognition*, which (1) has a relatively large number of training items that vary in their category typicality, item frequency, and word frequency; (2) includes different contexts for naming items at different levels of specificity; (3) includes properties that are differentially "important" for different conceptual domains; and so on. All of the important effects in the book can be observed in this single model, which compared to the original Rumelhart model (Rumelhart & Todd 1993), is relatively complex. For the purposes of understanding why the model gives rise to these effects, and communicating this explanation to the reader, this model is less than ideal. Precisely because frequency, familiarity, typicality, contextual specificity, and so on, are all operating in the model simultaneously, it is difficult to understand which factors are responsible for which effects and why, or how the different factors interact. The simpler versions of the model (discussed in *Semantic Cognition*, Chs. 3 and 4) held many of these factors constant and so made it easier to understand, and hence to communicate, fundamental aspects of the model's behavior. After establishing these fundamental behaviors, it becomes easier to communicate how additional factors then operate in the more complex model to explain a broader range of phenomena. So, when **Marcus & Keil** suggest we use different models to explain different phenomena, we view this aspect of our work as a virtue: The different models allow us to more clearly illustrate why a particular model works, as opposed to just demonstrating that it does.

R1.1. Three ways our models simplify our underlying theory

With these points in mind, we now consider some of the ways our model departs from the underlying theory, and some of the critical reactions these departures have provoked. In each case, we view the departure in question as a simplification adopted to promote the goals of understanding, demonstrating, and communicating implications of the underlying theory.

R1.1.1. Simplified input and output representations. In most simulations, we employed a single input unit to represent each subordinate level concept in the model's environment (i.e., one unit for rose, another for daisy, another for pine, etc.); a single unit for each different context (e.g., *ISA*, is, can, has); and a single unit for each possible output attribute (e.g., grow, fly, feathers, gills, etc.). Given our theoretical commitment to distributed representations throughout the semantic system, this use of "localist" inputs and outputs is clearly a simplification used for convenience (see, e.g., *Semantic Cognition*, Chs. 4 and 9). Still, some commentators have wondered about this choice. **Snedeker** asks, for example, if the use of localist input representations was tantamount to accepting Fodor's claim that representations of all concepts are essentially built in. If the use of localist input and output representations was a necessary condition for the success of the model, this might be a valid concern, but as we demonstrated with simulations described in Chapter 4 of the book, this is not the case.

We adopted localist input representations partly for historical reasons: The simplest version of the model was, in fact, the very model used by Rumelhart (1990). Rumelhart was specifically interested in using a connectionist model to encode the contents of systems of propositions like "canaries can sing" or "trees have roots." We think the simplification here is natural and, essentially, uncontroversial. Although spoken words are, in our view, represented as distributed auditory patterns, the relationship between spoken word forms and their meanings is largely arbitrary, and it seems clear that our language-processing systems are capable of ignoring such similarities at the input level when mapping spoken word forms onto meanings (as distributed connectionist models also learn to do; see, e.g., Dilkina et al. 2008; Plaut & Shallice 1993). Note that the use of localist input units to represent distinct words does not amount to knowing the concepts in advance – it simply imbues the network with the ability to treat each possible word form as distinct from every other.

In *Semantic Cognition*, we were interested in understanding how people learn from experience with objects in the world, including statements about objects but also other aspects of experience, such as watching a canary and then seeing it fly away or hearing it sing. In this case the use of localist input representations is still a useful simplification for many purposes, but, like others, we were concerned that use of such representations might perhaps seem to presuppose too much "built-in" knowledge. In *Semantic Cognition*, Chapter 4, we showed that the model behaved essentially identically whether trained with a single unit to represent each subordinate concept or several different units to represent individual exemplars of each concept. We also replaced the localist input representations with distributed input patterns, and showed that this model behaved similarly to a model with localist inputs.

The distributed model could learn to overcome superficial input similarity (i.e., raw overlap in the input patterns), and to weight certain input features much more strongly than others, when mapping from inputs onto internal semantic representations. It also provided a more natural way of presenting novel items to the model: Rather than adding new units to represent new information – a procedure to which **Quinlan** particularly objects – new items could be presented as novel patterns of activation across existing units. The use of distributed input patterns, however, raised other issues that distracted from the main insights we wished to communicate. Specifically, such representations raise the question of what kind of "perceptual" similarity structure should be "built-in" to the inputs. Since we wished to show that the model's behavior does not depend upon building any similarity structure into the inputs, we continued to employ localist inputs for most simulations.

In short, our theory does not assume localist input representations. Our models used these for simplicity in many cases; but to demonstrate that the important effects did not depend upon this simplification, we also explored the use of distributed input representations. Similar issues arise regarding the context units in our model – as with the input units, we believe these representations to be distributed. Although we did not run simulations with distributed context representations in *Semantic Cognition*, we did consider the relationship between our framework and other models that learn distributed internal context representations in Chapter 9.

The situation with the attribute units is somewhat different: In this case, the outputs are in fact patterns with more than one active unit. For example, for the input *canary can*, the model learned to activate a pattern of activation across output units (specifically *grow, move, fly*, and *sing*). The pattern is coded over units that themselves are individually labeled with particular attributes; but what is important to the network is not the labels on the units, but the similarity relations among the patterns assigned to different items. In this regard, we think our models make quite defensible first approximations: For example, on the *can* units, canary overlapped the most with *robin*, somewhat less with *salmon* and *sunfish*, and very little with any of the plants. It is the pattern of overlap (both within and across contexts) and not the labels on the features that causes the network to learn and to behave as it does. These points are discussed in detail in Chapter 4 of our book.

R1.1.2. Simplified, unidirectional flow of activation. A second way in which our models simplify our theory concerns the flow of activation through the network. In *Semantic Cognition*, we used only feed-forward models, in which activation flows in only one direction, from inputs through internal representations toward outputs. In our theory, activation occurs as a recurrent, mutual-constraint satisfaction process, in which all of the units in the network participate in the activation of all other units (see e.g., Farah & McClelland 1991; Rogers et al. 2004). Among other things, this means that activation can flow both from internal representations toward representations of specific sensory, motor, and linguistic attributes, and from the attribute representations toward the internal representations. The use of a feed-forward network greatly speeds up processing and learning in the model, but it also raised two issues in the commentaries.

One set of issues that arises here pertains to the representation of new items in the network (e.g., **Snedeker**, **Quinlan**). In our feed-forward networks, the properties of individual items are represented across the network's output units. Thus, to teach the network about something new, we not only require a new input unit as previously discussed; we also must provide specific information about the item in the form of target outputs. From such targets we used the backpropagation-to-activation method to allow the model to derive internal representations of novel items when given information about their properties. Although we see this approach as a valuable extension of the connectionist framework (introduced long ago by Miikkulainen & Dyer [1987]) that can easily be motivated in conceptual terms ("Adjust your internal representation so that it accounts for the specified target properties") it might be felt that presenting "input" to the network over output units is unnatural and that backpropagation is implausible biologically.

Appreciation of the computational properties of connectionist models has often been diverted by concern over the biological plausibility of backpropagation. These models are not intended, however, to exactly reproduce the actual processing activity of real biological neurons, but rather to make explicit certain computations that real biological systems might carry out in slightly different ways. In fact, it is now quite well established that in recurrent networks, the difference between activation values at different time points can be used to compute the very same error signals that are explicitly computed by backpropagation (Hinton & McClelland 1988; Hinton & Salakhutdinov 2006; O'Reilly 1996). Thus, the abstract computational consequence we produce using backpropagation may be implemented through the activation process that occurs in a fully recurrent network.

Regarding the presentation of "input" to a network across its output units, this issue does not arise in fully recurrent networks because the distinction between input and output units disappears. Hinton's (1981) semantic network, the precursor of the feed-forward networks used later by Hinton (1986; 1989) and then further simplified by Rumelhart, was a fully recurrent network. The three elements of a *Item1-Relation-Item2* proposition like *canary can fly* are each represented by a pattern over a set of units that can serve as input and/or as output, and each has its own corresponding set of representation units (see Fig. R1). External inputs applied to any of the three input pools constrain the corresponding hidden unit representations through the propagation of ordinary activation signals. Thus, information such as the fact that a previously unknown item *can move* can be represented as a pattern over the *Relation* and *Item2* pools, and propagation of activation will then constrain the representation assigned to this item over the *Item1* internal representation units. In general, in recurrent models of semantic cognition, which we have employed extensively in other work (e.g., Farah & McClelland 1991; Lambon Ralph et al. 2007; Rogers et al. 2004), there is no need to treat some units exclusively as "inputs" and others as "outputs."

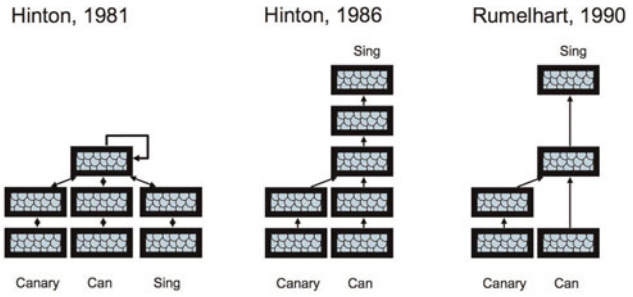

Figure R1. Schematic diagram illustrating the progressive simplification of Hinton's (1981) semantic representation model. The model was made strictly feed-forward in Hinton (1986; see also Hinton 1989), and then further simplified by Rumelhart (1990).

A separate issue concerns differences in processing dynamics between feed-forward and recurrent networks. Specifically, in feed-forward networks, a given input generates an essentially static pattern of activation across successive layers. In recurrent networks, a given unit can influence and be influenced by the other units with which it is connected. Consequently, a given input initiates a dynamic interaction among units, and the resulting patterns of activation across units can evolve over time in complex and nonlinear ways. In particular, they can exhibit attractor dynamics, that is, a tendency to settle from any of a range of inputs into the same final state or fixed point. Also, the use of fully recurrent networks may, as **Kropff & Treves** suggest, help explain abrupt transitions in the course of cognitive development; slight changes in connection weights can suddenly lead to a change in the attractor structure, causing relatively abrupt transitions even if the underlying connection changes are subtle and gradual.

While we share **Kropff & Treves**'s enthusiasm for attractor networks, we would note that the abruptness of developmental transitions may often be overstated. In fact, some developmental psychologists who used to argue for very discrete jumps from stage to stage in performance of certain tasks now advocate an "overlapping waves" model (Siegler & Chen 1998), and there are many signs of a fairly gradual change accompanied by a gradually increasing sensitivity to previously ignored information (McClelland 1995; Schapiro & McClelland, in press). Thus, while there may be good reasons to favor models with attractor dynamics, we are not sure that purported discontinuities in development are among them. The patterns of acceleration and deceleration seen even in feed-forward networks, discussed extensively in *Semantic Cognition*, Chapter 3, may be sufficient to address most aspects of these data (Schapiro & McClelland, in press).

In summary, we view the feed-forward networks that we have used as simplifications of the dynamically more complex, highly recurrent networks that are likely to be at work in the brain. Analysis of learning and processing in fully recurrent models is more difficult, though, and the use of backpropagation-to-activation in a feed-forward model captures much of what we might otherwise rely on recurrent networks for. We are interested in further exploring differences between these types of networks in their ability to explain patterns of change in development.

R1.1.3. Simplified environment. The third major simplification adopted in our work is in the training "environment," that is, the set of patterns used to train the network. This is, of course, true of essentially all modeling work in this domain, connectionist or not, and as in all such work, the key question is whether the simplifications we have adopted capture enough of the characteristics of real-world structure to be interesting, or whether they are unrealistic in ways that critically influence the reported results.

According to **ter Meulen** and several others (**Kemp & Tenenbaum**, **Rakison & Lupyan**, **Quinlan**), children do not experience all of an item's properties every time they encounter the object. As a criticism of our model, this point is off target, since the model receives information at the same time only about those properties that are relevant to the current context. In the *ISA* (naming) context, for example, the model learns about an item's names; whereas in the *has* context, it learns about the item's parts; in the *can* context, it learns about behaviors; and so on. Context-sensitivity in our models is specifically intended to address precisely the concerns that the authors raise, that children only get exposed to partial information about an item's properties in any given situation or context.

This point aside, it is certainly the case that, for all of our models, both the training patterns themselves and the range of contexts in which items are encountered are hardly realistic, as **Robbins** notes, for instance. Does this simplicity undermine our arguments in favor of the overall approach exemplified by our simulations? For one thing, even the most complex model only learns about 21 different items, all living things. We think it is fair to ask whether the theory can scale up well to much larger training environments based on naturalistic corpora. The application of connectionist models to such corpora has been a key goal of work by McRae and colleagues (Cree et al. 1999; McRae & Cree 2002; McRae et al. 1997), who have successfully used their models to understand a variety of phenomena in priming and speeded property verification. We would be deeply gratified if others would join us in exploring this extension of our work. In this connection it may be worth noting that recent advances on the machine learning side of connectionist modeling have greatly increased the efficiency of learning in connectionist networks (Bengio et al. 2006; Hinton & Salakhutdinov 2006; Ranzato et al. 2007), and we are actively planning to work on such extensions drawing on these developments.

Specific criticisms can also be leveled at the particular contexts and attributes used in our networks. For example, the network is trained with attributes labeled *can grow* and *ISA living thing*. It seems unlikely that children have access to such information at early phases of their development, but it is important to recognize once again that the model's behavior is determined, not by the specific labels given to the *Item, Context*, and *Attribute* units, but by the relationships among the patterns with which it is trained. The important point is that the *can grow* and *ISA living thing* attributes are shared by all of the items in our training environments. We think it

defensible that, in fact, plants and animals do share some attributes in common that are accessible to young children (again, see *Semantic Cognition*, Ch.4, for discussion).

What is important for the model is that, in the patterns used to train our networks, the attributes show systematic coherent covariation across the different items and contexts. Our theory explains many of the key phenomena in development (and adult semantic cognition) with reference to the ways that this assumed structure influences learning and processing. For these explanations to be correct explanations of the phenomena as they occur in humans, it must be the case that the actual experiences of children exhibit the structure that our simulations suggest is important. For example, we found that basic-level naming advantages arise in the model when basic-level labels pick out sets of things that share many properties with each other and few properties with other sets of things (as Rosch et al. [1976] originally proposed). If it is not the case that members of basic-level categories share many properties with one another and few with members of contrasting categories, then this would provide evidence against our account of basic-level effects. In this particular case, empirical evidence supports the assumptions that guided the construction of the training corpus: Basic level categories do appear to maximize distinctiveness and informativeness, as assessed in a variety of norming and experimental studies over the years (McRae et al. 2005; Murphy 2000; Rosch et al. 1976; Tanaka & Taylor 1991). Thus, whether or not the specific patterns used in our simulations are fully realistic in the sense that they capture information about the properties of a large set of actual objects that people are likely to know about, they may be realistic in the sense that they capture the essential aspects of structure that are important, under the theory, for producing the effects of interest.

In our view, the place where the theory is in most need of development is in the enrichment of its *Context* representations. We relied on a very small and somewhat arbitrary set of possible contexts, which came initially from the small set of relations used in Collins and Quillian's (1969) propositional hierarchy and were in turn adopted by Rumelhart. But as **Hampton** points out, the range of variation of object properties across contexts is extremely rich, with important, and as yet unexplored, implications for the nature of conceptual representations. We look forward to extensions of our theory into a fuller investigation of these issues.

In summary, some of the features of the models used in *Semantic Cognition* are deliberate simplifications that should not be confused with assumptions of our underlying theory. It is important to be clear about where we have simplified, and we hope the preceding discussion clarifies these issues. It is also legitimate to ask: If we had not employed these simplifications, would a more complex model based on the theory still exhibit the phenomena we have illustrated with the simple models? If the answer is "yes," then the simpler model can be viewed as a useful tool for understanding the behavior of the more complex reality that fully conforms to the principles of our theory. It would be possible to construct a more complex model that is truer to the theory, and there are good reasons why this should be pursued, but there will still be a place for the simpler models of the kind we have used, that exhibit many of the fundamental characteristics of our approach very clearly.

R2. Are there aspects of semantic cognition our approach cannot address?

Several commentaries focused on phenomena that we have not yet addressed in our work. We agree with **Marcus & Keil** when they state that we have not shown that "intercorrelations alone suffice for human reasoning." A full account of all aspects of human reasoning is surely a task that will require far more than one book, and certainly new insights will be required before these books can be written. We do hope and expect that some of the ideas in *Semantic Cognition* will play a role in the gradual emergence of these insights.

For the most part, we found the focus in many of the commentaries on the yet-to-be-explained very positive, since it points to new opportunities to extend the coverage of the theory. We found that these comments fell into three types: Phenomena outside the current scope of the theory, either because we do not aspire to extend our approach to them, or because they lie beyond its current reach; phenomena to which the approach might relatively easily be extended; and phenomena that seem to challenge fundamental aspects of our approach. We consider the first type briefly before turning more detailed attention to the last two.

R2.1. Phenomena outside the theory's current scope

There are aspects of human cognition that we do not plan to address within our framework. For example, as **Feeney, Crisp, & Wilburn** (**Feeney et al.**) note, we have not tried to address explicit deductive inference – that is, the process of reasoning from premises to conclusions via the rules of logic. As we discussed in Chapter 9 of our book, we do believe these processes are deeply influenced by the kinds of implicit knowledge addressed in our book, and this separates us from cognitive scientists who treat the processing of logical syllogisms and other formal structures as indicative of the fundamental nature of human thought (e.g., Fodor & Pylyshyn 1988; Marcus 2001). We take the alternative view that such processes, though they can be mastered with practice, are not the natural basis for human cognition. Instead they are acquired skills that depend like other aspects of cognition on extensive relevant experience.

Just outside our model's current scope lies an important direction for future development of our approach: to address the relationship between the kinds of implicit processes examined in our model and processes that control or regulate these processes. **Feeney et al.** review evidence that control processes play a role even in fairly simple inductive inferences of a kind we think our model should address. We think this role occurs through the internal manipulation of context representations and the internal re-use of outputs of implicit cognitive processes as re-entrant inputs to the system, rather than through the use of some completely separate type of processing system. The PDP model of control of automatic processes in the Stroop task (Cohen et al. 1990) and the latter part of the chapter in the PDP books on schemata and sequential

thought processes (Rumelhart et al. 1986c) discuss these ideas. We agree with Feeney et al. that the fleshing out of these ideas will require another book.

A successor to our approach may someday begin to address the phenomena of insight and creativity – crucial features of human cognition that are not easily captured, we believe, in systems like those envisioned by, e.g., Fodor & Pylyshyn (1988) and Marcus (2001), and which a complete theory of human cognition should ultimately address. It is our belief that these characteristics of human thought arise from graded constraint-satisfaction-based processes dependent on implicit knowledge acquired by mechanisms of the sort that we explore in our book (as these are guided by control processes) rather than from the application of structure-sensitive or "algebraic" processes of the kind favored by Fodor and Pylyshyn (1988) and Marcus (2001). Actually developing a model that addresses these topics will be a challenge that is certainly worth undertaking.

R2.2. Phenomena compatible with the approach

The next group of phenomena are those we believe our framework could readily address, but which have not been a focus of our work to date.

R2.2.1. Initial saliency. Mandler, **Quinn**, and **Rakison & Lupyan** point out that although all properties are given equal initial weight in our simulations, in fact some properties are likely to be initially more salient to infants and so are likely to strongly shape early concept learning. As we indicated in *Semantic Cognition*, Chapter 4, saliency variations across different kinds of information can easily be incorporated into our framework. Salient features can be given more "weight" in a connectionist network in a number of ways: by scaling error derivatives accruing on the units that represent such properties; by scaling the learning rate associated with these properties; or by using more units to represent them. Any or all of these strategies would be available to evolution as a simple way of adjusting the relative salience of particular stimulus dimensions or attributes. We are not opposed to the possibility that development may reflect, in part, differential salience of some types of information relative to other types. Our effort has been to show that, in models with a convergent architecture, saliency scaling can arise automatically for properties that exhibit coherent covariation with one another. This kind of emergent saliency may be sufficient to account for many developmental phenomena, and where there is a convincing case to be made for differential salience of certain types of information, it is possible to incorporate this into the framework.

R2.2.2. Cross-linguistic influences on concepts. Majid & Huettig note evidence that people in different linguistic communities may organize their concepts somewhat differently, suggesting that the way we refer to objects and events in speech may influence the concepts we acquire. We agree with these points, and they are completely compatible with our framework, where producing and comprehending language is one of many abilities supported by the semantic system. In our models these abilities are supported in a small way through the *ISA* relation and attribute units, which capture information conveyed primarily through explicit labeling of objects by others in the environment. Consequently, the structure of a given language will influence the concepts we acquire, just as will other kinds of culture-specific experiences, including different degrees of exposure to different kinds of information, differential emphasis on certain distinctions in some languages or cultures relative to others, and so on. Some such influences were examined in our simulation of phenomena in expertise (see *Semantic Cognition*, Ch. 5), and other recent work is extending these ideas (Dilkina et al. 2007).

R2.2.3. Different kinds of situations and contexts. Hampton notes that the different contexts adopted by our models – the situations in which different items are encountered that determine which attributes are of current relevance – are extremely simplistic, and wonders whether they could be expanded, perhaps by adopting a more extended context-coding scheme such as that employed by McRae and colleagues (e.g., McRae et al. 2005). We agree that a richer variety of contexts is necessary for further development of the theory. In addition to considering "semantically based" kinds of contexts, as Hampton suggests, we believe that it will be important to consider the temporal context in which items are encountered – that is, the sequence of events surrounding a particular encounter with an item – as well as other aspects of context, such as the particular settings in which various items are encountered, the items with which they tend to interact, and the linguistic contexts in which objects are designated by category labels at different levels of specificity. These and related issues are discussed in *Semantic Cognition*, Chapter 8.

R2.2.4. Accounting for delusions and other forms of semantic dysfunction. A major part of our research effort has been to use our models to better understand disorders of semantic cognition occurring as a consequence of neurodegenerative illnesses and other forms of brain damage (e.g., Lambon Ralph et al. 2007; Rogers & Patterson 2007; Rogers et al. 2004). **MacDonald** inquires whether our framework could further aid in understanding neuropsychiatric disorders in which delusions are a significant feature. For instance, some theorists have proposed that Capgras syndrome – where the sufferer believes that close family and friends have been replaced with impostors – arises from a functional disconnection between the systems that mediate person-recognition and the limbic system (Ellis & Lewis 2001). Loved ones are recognized but fail to evoke the expected emotional response, a failure subsequently interpreted as evidence that that the individual is not "really" who she claims to be. This theory has a straightforward interpretation under the fully recurrent model implementations of our theory: Disruption of connections between the integrative semantic pool and a separate "limbic" pool that encodes emotional responses, for example, could produce partial recognition of a person without the associated emotional response.

In summary, the four phenomena considered here exemplify ways in which the theory might be fruitfully extended. The fact that the theory could likely be extended to address them testifies to its breadth and utility.

R2.3. Phenomena that challenge our framework

The third group consists of phenomena raised by several commentators (**Opfer & Doumas**, **Marcus & Keil**, **Quinlan**, **Kemp & Tenenbaum**) as fundamental challenges to our theory. These are aspects of human cognition that the commentators believe to be (1) a critical part of semantic cognition and (2) beyond the ability of our theory, in principle, to capture.

These comments almost all focused on the common idea that "deeper" aspects of human thinking involve setting aside superficial similarity relationships (e.g., physical similarity) among items, and taking note instead of similarities among the relations between items. On this view, two structured entities (domains, situations, etc.) are thought to be analogous if the items within them enter into a similar pattern of interrelations with one another. And, to the extent that two structures are analogous, alignment of the structures can promote detection of similarity (and subsequently inductive inference) between items that otherwise may have little in common (Gentner 1983). As one simple example, foxes prey on hens just as eagles prey on mice. By virtue of the shared *preys on* relation in these propositions, it is possible to recognize that eagles and foxes are in some sense similar kinds of things – more similar, in this respect, than are eagles and chickens, even though eagles and chickens may share more properties. The ability to detect such relational similarities and to use these to reason about object properties is thought by many to require knowledge representations that explicitly mark relations among different items – such as, for example, directed graphs or propositional hierarchies.

Accordingly, several commentators have raised questions about our framework's ability to deal with relations effectively. As one example, **Opfer & Doumas** state that:

> children, unlike a PDP network, can represent these relations and their fillers in a manner that preserves relation-filler independence (i.e., relations and their fillers are represented independently), while simultaneously representing the bindings between roles and fillers in an explicit and dynamic fashion. Thus, children can appreciate how 'Fido chases Felix' is like 'Felix chases Fido' (same elements involved in the same relation) and how they differ (role-bindings are reversed).

A similar statement is made by **Marcus & Keil**, who suggest that such role-filler bindings cannot be achieved by connectionist networks without a "combinatorial explosion of input nodes." **Kemp & Tenenbaum** reiterate these points, and further indicate skepticism that our framework could ever cope with the indefinite variability in the number of roles associated with different relations (e.g., two in chase (dog, cat) versus three in eat (boy, cereal, spoon) – what they call different "arities"), nested relations (e.g., eating food causes the canary to grow), or the ability to discern similarities among different kinds of relations (e.g., the fact that the relation terms *has* and *possesses* have near-identical meanings).

It is true that the simple Rumelhart network used in *Semantic Cognition* has some of these limitations, but these do not reflect a fundamental limitation of the framework. Indeed, three of these issues – the issue of relation-filler independence and binding raised by **Opfer & Doumas**, the problem of achieving role-filler binding without "a combinatorial explosion of elements" raised by **Marcus & Keil**, and the issue of different "arities" and nested relationships noted by **Kemp & Tenenbaum** – were addressed by the Sentence Gestalt model (St. John & McClelland 1990), which employed a simple recurrent network architecture in which a series of inputs are integrated over time into a single learned internal representation (Fig. R2). In this case, sentences involving any number of relations were presented to the model one word at a time, and the model was trained to answer queries about the meaning of the sentence. After training, the model would generate a distributed internal representation for any given sentence – the "sentence gestalt" – from which it could produce the answer to any question about the fillers of every role. Each sentence included a subject and several other optional arguments; and word order, prepositions, and semantic information determined the assignment of fillers in the sentences to roles including agent, patient, recipient, and instrument. Only the relation role appeared obligatorily in all sentences (the surface subject role was always filled as well, but its underlying role could vary as in, e.g., "The boy broke the window," "the window broke," "the hammer broke the window," "the car was parked," etc.). An extension of this model by Rohde (2002) further addressed the issue of embedded relational structures.

Both of these models can be probed specifically for the filler of each of the roles in a sentence, including, crucially, the filler of the relation role. This would allow specification of the exact ways in sentences like "Fido chases Felix" and "Felix chases Fido" (or "lectures about movies" and

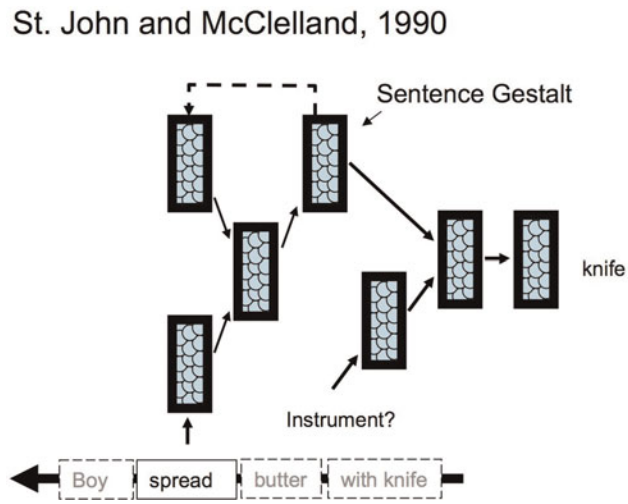

Figure R2. Sentence Gestalt model of St. John and McClelland (1990). In this model, words making up declarative sentences are presented one by one over the input layer of the network (pool of units at lower left). Each word is used, together with the sentence gestalt representation derived from the previous word, to create an updated sentence gestalt. Probes querying the filler of one of the roles in the sentence can be presented to the probe units. In the example shown, the words "Boy" and "spread" from the sentence "Boy spread butter on bread" have so far been presented. The network, when queried for the instrument, can already respond "knife" since in its experience knives are the only instruments used for spreading. Note that the sentences may contain an indefinite number of roles, one of which is the "action" role. A later model by Rohde (2002) extended the probe syntax to allow querying specific arguments of specific clauses of multi-clause sentences.

"movies about lectures"; see **Marcus & Keil**) are similar (they involve the same relation) and how they are different (the fillers of the roles associated with these relations are reversed). The Sentence Gestalt model also learned appropriate similarity structure among the relations appearing in its training corpus (as did Elman's [1990] simple recurrent network model and Hinton's [1989] family trees model). Because the Sentence Gestalt model learned internal representations of sentences from inputs arriving one word at a time, it could capture role-filler bindings without encountering the "combinatorial explosion" of input nodes alluded to by **Marcus & Keil**. Rohde's (2002) extension of the model to embedded relations showed that the framework has no difficulty with the nested relations raised by **Kemp & Tenenbaum**.

In sum, many commentators appear to hold the intuition that our framework is incapable of dealing with knowledge about relations and view this as a critical flaw. The Rumelhart network's simplicity perhaps invites this reaction, but precisely these issues have been addressed in prior work. In *Semantic Cognition*, Chapter 9, we discuss alternative network structures, based in part on the Sentence Gestalt model and Rohde's (2002) extension of it, that would address these limitations while still preserving all of the positive features of our approach.

R2.4. Are connectionist networks merely statistical engines of a familiar kind?

We have not yet addressed three important points raised in the commentaries concerning the putative limitations of PDP approaches to cognition generally, and to semantic cognition in particular. Two were previously noted: The first of these is **Kemp & Tenenbaum**'s claim that our theory cannot explain how people learn that different relation terms, such as *has* and *possesses*, have similar meanings. The second is the claim that our model will have difficulty capturing very abstract semantic relationships and concepts, such as the concept predator, which appears to group items, not on the basis of property overlap, but on the basis of a given item's relationship to other items – so that the eagle and fox are considered similar "kinds," whereas the eagle and chicken are considered different "kinds," despite the fact that eagles and chickens certainly share many properties with one another. We suspect that both of these issues arise from the underlying concern that connectionist models can only capture similarity structure that arises from direct property overlap. If our models only come to represent items as similar when they share sets of observable properties, then it is difficult to understand how we come to know that very dissimilar items are, in some ways, similar "kinds" of things; or how very abstract terms, such as relationship terms, might have quite similar meanings.

The third point is related and concerns statements by several commentators to the effect that connectionist models are merely statistical learning machines. For instance, **Borsboom & Visser** state that they only perform "statistical categorization procedures," which they take to include hierarchical clustering analysis, factor analysis, principle components analysis, regression, and multidimensional scaling; **Marcus & Keil** call them "finely tuned statistical engines"; and **Quinlan** describes them as "connectionist counting machines." It is difficult not to see such comments as dismissive of the general approach. Borsboom & Visser, for example, recommend that – since we use a variety of statistical methods to understand the model's behavior anyway – we simply dispense with the model and apply the same more standard statistical methods directly to the model training patterns. The assumption appears to be that the connectionist learning procedures themselves contribute essentially nothing – that standard multivariate analyses of the training patterns would reveal the same structure that emerges in our model; and that the various phenomena we account for in the book, such as the progressive differentiation of concepts, domain-specific patterns of attribute weighting, basic-level advantages in naming, and so on, would be fairly transparently related to the kinds of results yielded by these more standard analyses. A similar sentiment may underlie Quinlan's statement that the model only discovers structure that is obvious given the training examples, and **Snedeker**'s suggestion that the "concepts" the model acquires are essentially built in.

The logic underlying such conclusions is not clear to us. It seems to be something like the following: Connectionist models are a kind of statistical categorization procedure; therefore they must have the same properties as other statistical categorization methods; therefore one can apply other methods to the same patterns to equal effect. But this is a fallacy, akin to concluding that bats, because they are mammals, must not be able to fly, since most more-familiar mammals cannot fly. Connectionist models may be a kind of "statistical categorization procedure" under some definitions, but this does not mean that they have exactly all and only the same properties as other statistical procedures.

We believe that connectionist models differ from many familiar statistical procedures in important ways. In the current section, we wish to demonstrate that this is so by doing just as **Borsboom & Visser** and **Quinlan** suggest: by showing that our models can learn structure that is not discovered by two commonly used unsupervised analysis methods, namely, hierarchical cluster analysis and principal components analysis. Furthermore, we will show that this structure is precisely the very abstract kind of structure that **Kemp & Tenenbaum** and others suggest is beyond the capability of connectionist approaches: Our model learns to treat items as similar when they enter into similar relations with other items even if they have no direct property overlap (as in the case of the predator concept); and it learns to treat different relation terms as similar when they capture comparable patterns of similarity among different items, even if they are associated with completely non-overlapping sets of properties.

Our demonstration relies on a corpus of patterns similar to those used in the simulations reported in *Semantic Cognition*, but coming in four completely non-overlapping sets, which we will call domains. Each domain contains eight items, and within each domain, the items can be encountered in four different contexts. The contexts do not overlap across domains – each domain has its own set of four contexts. Finally, each item has a set of attributes pertaining to it in each context, and as with items and contexts, each domain has its own completely distinct set. Thus, across domains, there is no overlap of any kind in any of the *Item, Context*, or *Attribute* layers.

Within a given domain there exists similarity structure in the overlap of output attributes across items; and this similarity structure among the eight items within a domain is identical across the four domains. To see this, consider the similarity relations in Figure R3, which shows a hierarchical cluster analysis of the output patterns describing all 32 individuals. The cluster algorithm strongly differentiates the four domains, since individuals in different domains share no properties. The plot also shows that the within-domain similarity structure is the same for every domain: the shape of the tree beneath each superordinate node is identical. In this sense, the domains are "analogous" or "alignable." Each domain contains two very similar individuals (squares), four individuals that are very similar to one another but distant from their other domain members (circles), and two "oddball" individuals (stars). It is important to note that this cross-domain structure is completely absent from the network's inputs (which are all localist) and from its outputs (see top panel of Fig. R3).

In other words, although items from different domains share no properties in this data set, there exists a second-order isomorphism across the domains: The individuals represented as circles of varying shades, for instance, have no properties in common, but each relates to the other individuals in its own domain in precisely the same way.

The architecture we employed for this simulation is shown in Figure R4. It is similar to that used in our book, except that we have added an additional hidden layer between the *Context* input units and the *Hidden* layer, so that the model must learn to represent each of the different contexts with a distributed pattern of activation across the units labeled *Context Representation*.

What does this version of the Rumelhart network learn about the 32 items in this corpus? To answer this question, we trained the model in the usual way. The connection weights in the network were initialized with very small random values, and on each trial, an *Individual* and *Context* unit was activated in the input. Activations were computed in a forward pass; the difference between the actual and desired outputs was computed and transformed to a measure of error; and the weights were adjusted in a backward sweep so as to reduce the error. The model was trained for 30,000 epochs without momentum and with a learning rate of 0.05, at which point 99% of the output units were activated to within 0.1 of their target values. We then stepped through the 32 individual inputs, recorded the learned pattern of activation across *Representation* units for each item, and computed the Euclidean distance matrix describing the dissimilarities among these learned representations. This simulation was run five times with different initial starting weights, and the resulting distance matrices were averaged across simulation runs, to ensure that the results do not reflect chance findings from a particular set of starting weights.

Figure R5 shows a hierarchical cluster analysis of the average distance matrix from these simulations. The model clearly finds a different organization of its internal representations than that expressed in the output vectors directly: In the representations learned by the network, individuals with corresponding similarity relations to other items within each domain are now treated as similar. The second-order similarity is now the dominant organizing structure in the representations assigned to these items in the network. This occurs, even though these individuals have no properties in common.

Again, this result is not a strange artifact of the clustering algorithm, but can be observed directly in the pairwise distance matrix itself (Fig. R5 top): The model clearly captures cross-domain similarity structure based on an individual's role within the domain, while representing as dissimilar individuals who actually share some overt properties in common but who play very different roles within the same domain.

It looks as though the "statistical categorization procedure" embedded in connectionist models is different in some important ways from the procedures some of its critics (especially, **Borsboom & Visser**) equate it with. Multi-layer connectionist networks can certainly be seen as a kind of statistical learning machine, but they differ from other such procedures in their ability to discover interesting internal representations of their inputs as a consequence of the learning procedure and the particular network architecture. These differences make such networks especially relevant for understanding human semantic cognition. In this case the simulation demonstrates that there need not exist any direct overlap of properties in order for the model to learn to represent sets of items as similar to one another. That is, the model is capable of learning a kind of analogical similarity structure – representing as similar the items that occupy similar positions within alignable structures.

This simulation builds on the early "family trees" simulations of Hinton (1986; 1989), and there are other related demonstrations of transfer between non-overlapping patterns in simple recurrent network simulations (e.g., Dienes et al. 1999). There is one important difference between our simulation and Hinton's, however. In Hinton's simulations, items from different domains (i.e., members of different families) could appear in the same context (i.e., with the same "relation" unit activated in the input), so that in fact input patterns did overlap somewhat across domains. In contrast, we have used completely distinct context input units in each of the four domains so that items from different domains never overlap in any way in the input or output. So, we can also ask: Has the network learned anything interesting about the contexts themselves? Has it, in fact, been able to discover the cross-domain similarities in these contexts?

To answer this question, we used principal components analysis (PCA) to examine the similarities among (1) the output patterns associated with each context, and (2) the similarities among the 16 context representations learned by the model. Figure R6A shows plots of the first two components of a matrix encoding the average of the output patterns associated with each of the 16 individual contexts. Because each context is associated with a completely non-overlapping set of output properties, there is very little structure for the analysis to discover. The first component thus weakly differentiates the four individual contexts from domain 1; and the second component weakly differentiates domain 1 from all other domains. Subsequent components similarly serve to differentiate the individual contexts from one another without revealing any interesting substructure. As shown in Figure R6B, the same analysis conducted on the model's learned internal representations of contexts yields quite different results.

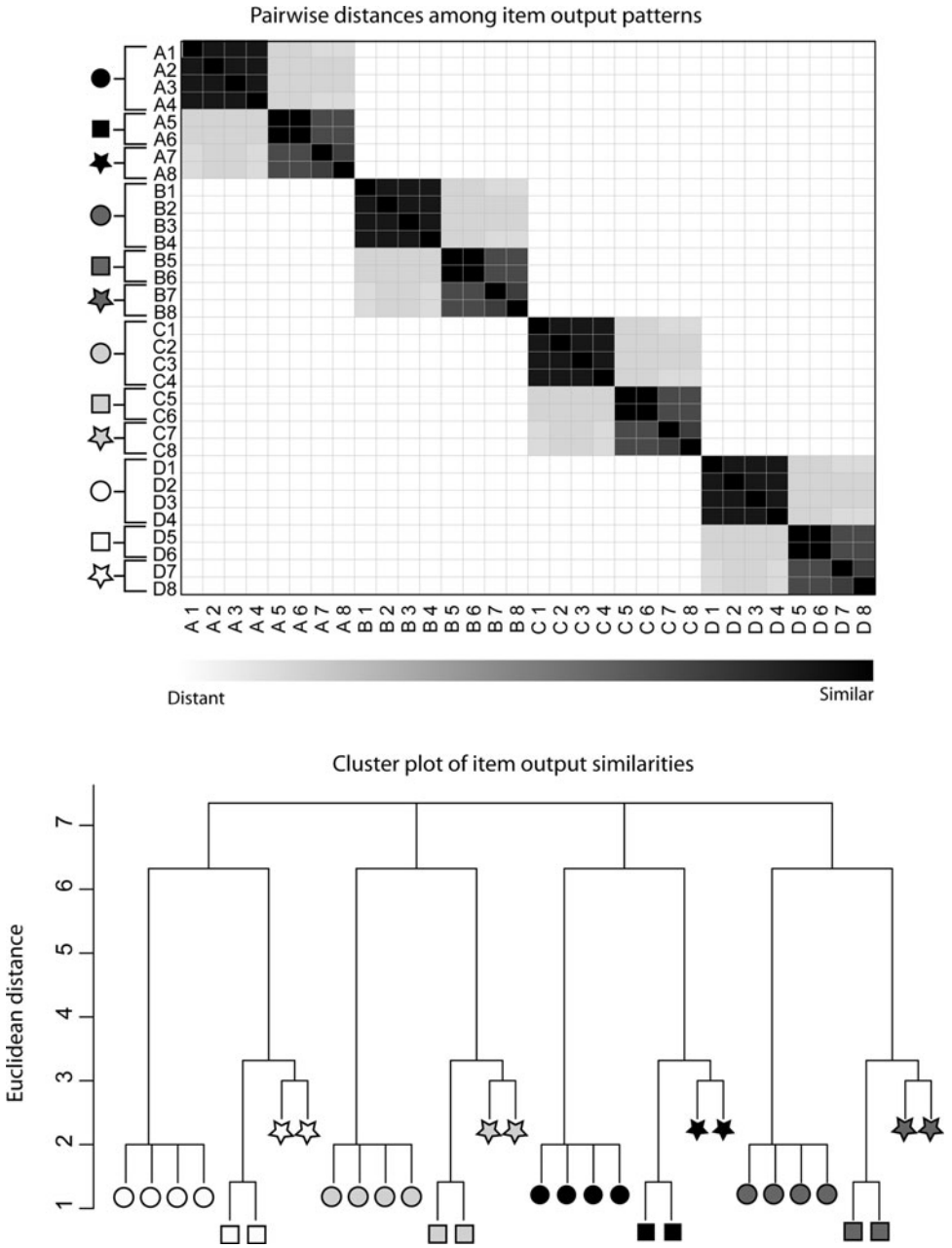

Figure R3. Similarities among the output training vectors for 32 items (collapsing across context) as captured by the actual pairwise distance matrix with cells shaded to reflect Euclidean distances between patterns (top) and by a hierarchical cluster analysis of these distances (bottom). The symbols to the left of the distance matrix indicate the branch labels shown on the cluster plot. Because there is no feature-overlap across domains, both snapshots of the output data suggest that items from different domains should be treated as distant from one another. Thus the distance plot shows no overlap at all between items in different domains; and in the cluster analysis places items from the same domain beneath a common superordinate node. Both plots also show that, within each domain, the eight items enter into precisely the same set of similarity relationships with one another: the patterns of pairwise distances within domain are identical in the distance plot; and the shape of the tree beneath each superordinate branch is identical across domains.

In this case, the first principal component neatly organizes the contexts in a cross-domain fashion: context 1 in domain A is represented as similar to context 1 from domains B, C, and D, but different from other contexts; context 2 from domains A–D are represented as similar to one another and different from other contexts; and so on. The second component serves to differentiate the individual contexts, but with little additional apparent structure.

Why does the model learn to treat some contexts as similar to one another? The reason is that any given context captures a certain set of similarity relations among the items in the associated domain; and these similarity relations are also expressed in the corresponding contexts from the other domains. That is, context A1 captures the same similarity relations among domain A members as does context B1 among domain B members, context C1 among domain C members, and context D1

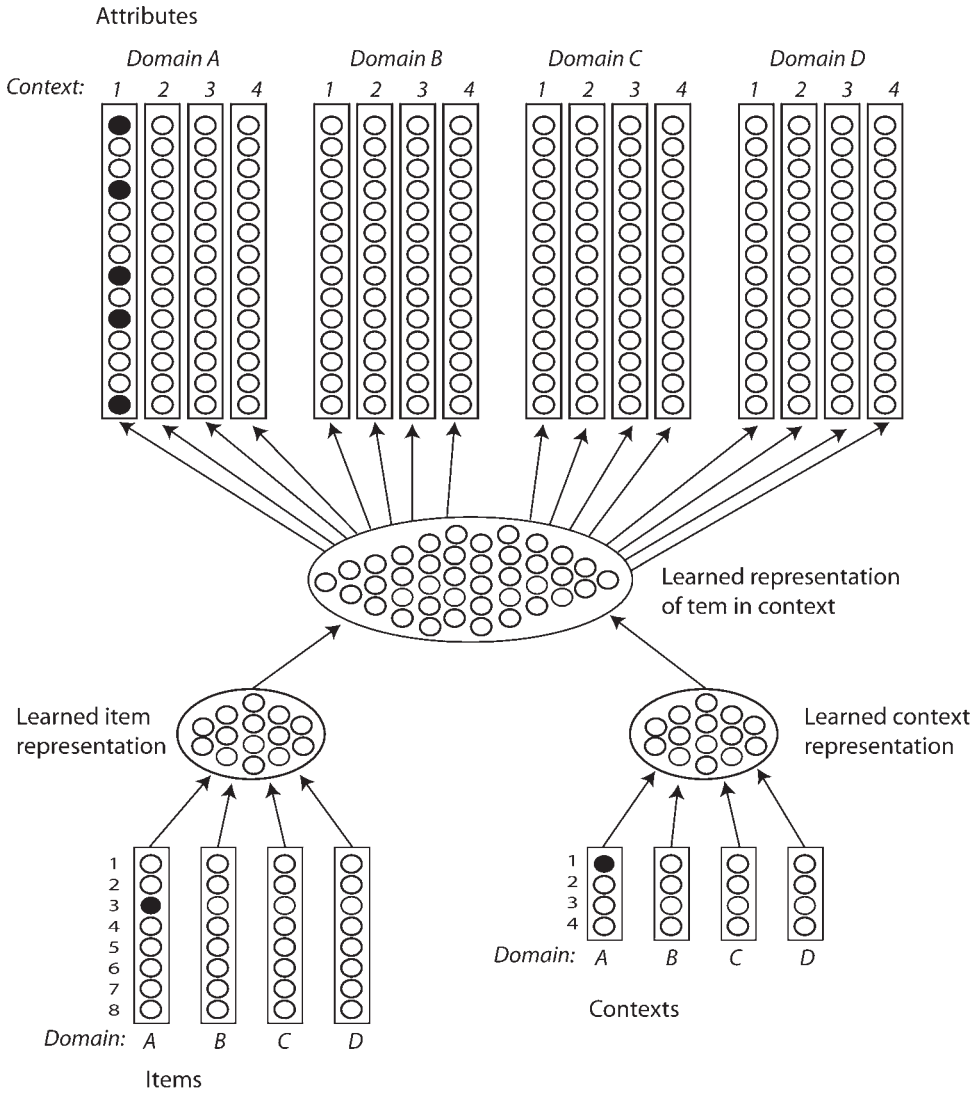

Figure R4. The architecture employed in our new simulations. As in the Rumelhart architecture, different individual items and contexts are represented locally in the input, and representations of attributes are distributed across output units. In this case, input items and contexts are organized into domains so that items from different domains never overlap in their inputs or outputs. In addition to learning a distributed internal representation of each individual, this model must also learn a distributed internal representation of each context (Context-rep) that is independent of the particular item being processed.

among domain D members. So, contexts A1, B1, C1, and D1 get represented as similar to one another. Context A2, in contrast, captures a different set of similarities among domain A members, and so gets represented as quite different from context A1. Though this simulation is abstract, it is easy to see that two different relations with near-identical meanings (whether in the same language or in different languages) – such as *have, possess*, and *avoir* – will come to be treated as quite similar by our model, contrary to the intuitions expressed by **Kemp & Tenenbaum** in their commentary.

In short, the model is capable of learning correspondences across domains between (1) items that share no input or output properties but relate to other items in their domains in similar ways and (2) relations/contexts that share no input or output properties, but which organize concepts in similar ways. These correspondences were not captured by other multivariate techniques applied to the output patterns themselves.

R2.5. They said it couldn't be done

In his commentary, **Quinlan** urges a focus on what connectionist networks like ours cannot do, and comes to a pessimistic assessment. Although his list of "can'ts" is short, others provide their own lists, and, if models like ours could do none of these things, it would indeed be worth discarding them for alternatives. But let us reconsider some of the key items on the list.

Connectionist networks can't learn relational similarity (**Feeney et al**; **Marcus & Keil**; **Opfer & Doumas**; **Quinlan**). In fact, they can. Indeed, ongoing work is beginning to show how this sensitivity to relational similarity can be exploited to explain specific phenomena in the empirical study of abstract concepts, analogy, metaphor, and so on. Clearly, dismissal of the research program on the basis these perceived limitations is not warranted.

Connectionist "counting-machines" are no different from "statistical categorization procedures" (**Borsboom**

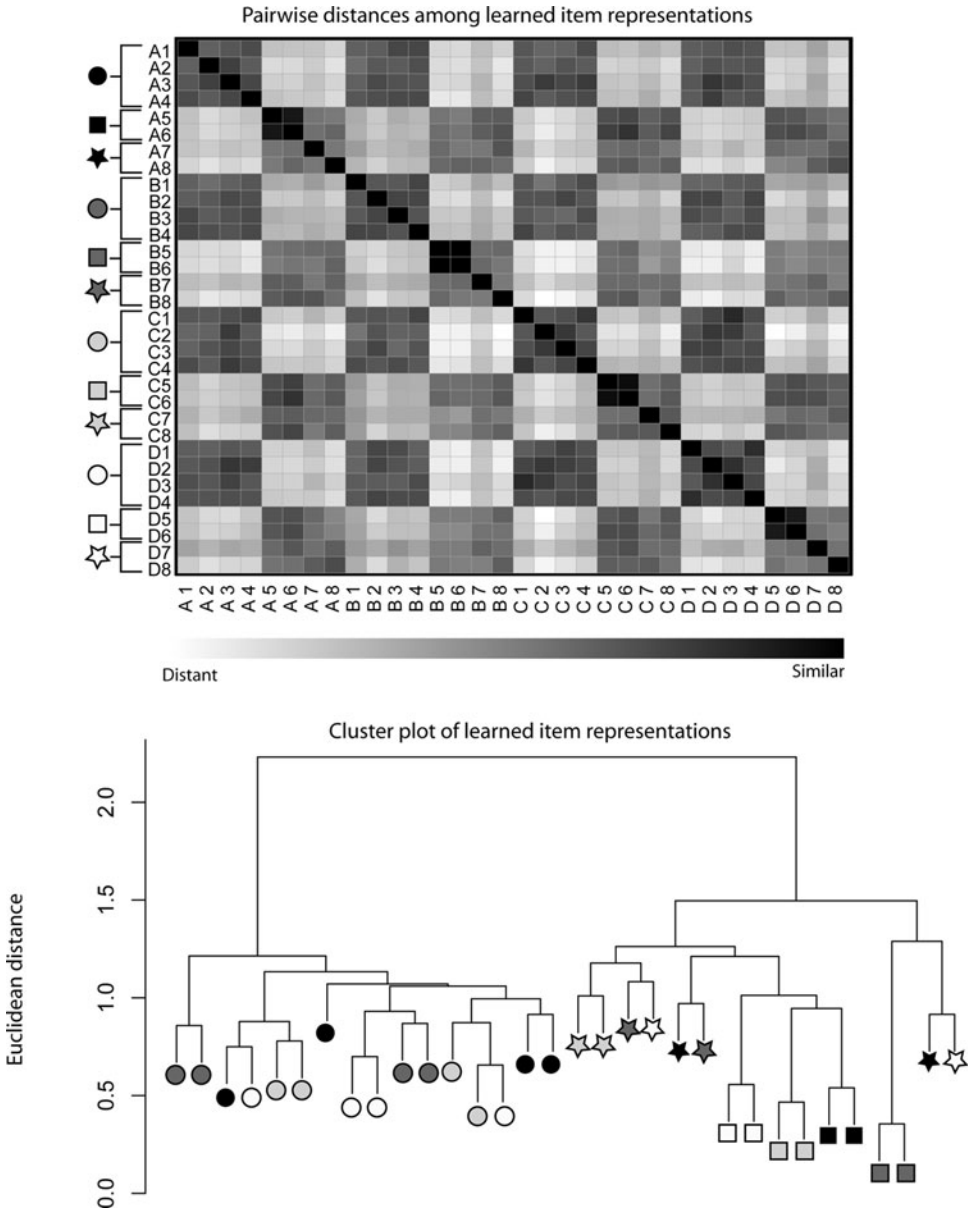

Figure R5. Similarities among the representations of 32 items learned by the model, as captured by the pairwise Euclidean distance matrix and by as hierarchical cluster analysis. In contrast to the analysis of the attribute vectors themselves (Fig. R3), the learned representations capture substantial cross-domain structure: items that play similar roles within their respective domains get represented as more similar to one another than do items that play quite different "roles" within the same domain.

& Visser; **Quinlan**). We have shown that they are different. Principal components analysis and hierarchical cluster analysis both fail to discover structure that our network discovers.

All the interesting structure is "built in" to the inputs and outputs by the experimenters, so any old structure-sensitive statistical method will find it (**Snedeker**; **Quinlan**). While the structure is in some sense in the input, it is not true that just any statistical procedure will find it. In fact, as we have repeatedly stressed, the ability of our models to learn about cross-context and cross-domain structure in its training environment depends critically upon the convergent architecture used in our models. Models trained on exactly the same patterns with a different architecture would not acquire the same internal representations and would exhibit completely different patterns of learning and generalization. The convergent architecture encourages the discovery of cross-domain structure, a property we view as crucial for our domain-general structure-sensitive learning procedure.

It is impossible to learn new concepts in a connectionist system (or any other learning system), according to **Snedeker**, citing Fodor (1998). Our model learns to group together items that have nothing in common in the input or the output, but that all relate to their respective domains in similar ways. If one considers the input and output units to correspond to "innate concepts," as Snedeker suggests, then the model has essentially learned a new set of concepts – concepts that, like prey and predator, transcend the particular domain (e.g., of land animals or sea creatures) but that relate corresponding items across these domains.

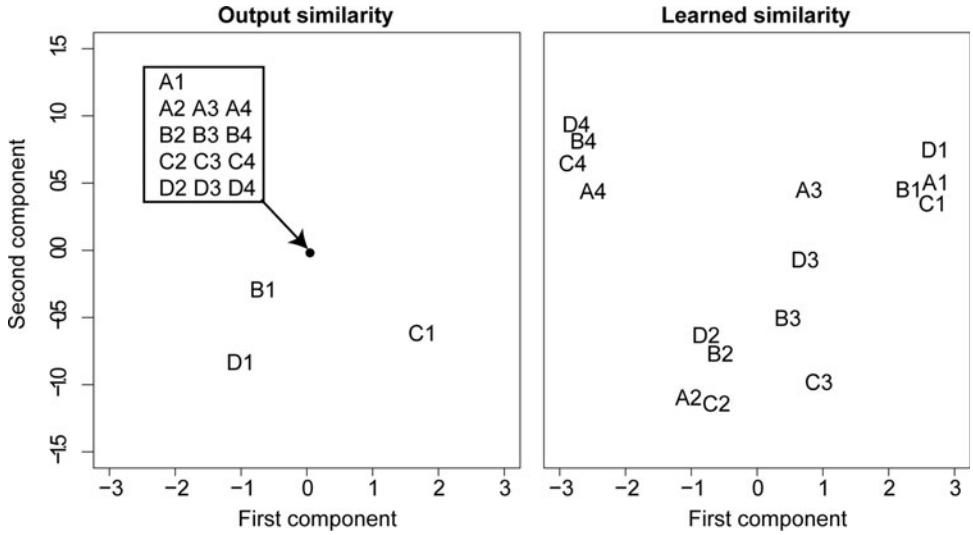

Figure R6. Principal components analyses of the 16 different contexts in the model training environment. The left panel (R6A) shows the first two components extracted from a matrix containing, for each context, the associated pattern of output activations averaged across all items. The right panel (R6B) shows the first two components extracted from the model's learned internal representations of the different contexts. Even though different contexts are associated with completely non-overlapping sets of output attributes, the model finds interesting structure that is not apparent from analysis of the training patterns alone. Specifically, it treats contexts as similar if their associated attributes express similar patterns of overlap among the eight items in each domain.

Connectionist networks can't learn about how different contexts are similar to or different from one another. **Kemp & Tenenbaum** suggest that, because our model uses different units to represent different relations, it will not be able to learn that different relations can be, in some respects, similar. We have previously seen that they were wrong about this through the example of the Sentence Gestalt model, and our new simulation reinforces this demonstration.

In summary, our simulation illustrates that the capabilities and behaviors of even very simple feed-forward neural networks continue to be widely misunderstood and counter to the intuitions of many researchers in the field. Several of the critiques of our work appear to be based on mistaken claims about in-principle limitations of the framework. The very fact that these misconceptions are so ubiquitous provides evidence supporting our earlier argument regarding an important function of modeling work: It allows the theorist to uncover, demonstrate, and communicate insights about possible cognitive mechanisms that run counter to intuition.

R3. Different levels of analysis or alternative frameworks for capturing structure in cognition?

One remaining issue arising in the commentaries concerns levels of analysis in cognitive science. This issue was most directly raised by **Kemp & Tenenbaum**, who take the position that PDP approaches and structured approaches are compatible, aiming at explanations at different levels of analysis, as also suggested by Smolensky (1988). Though other commentaries were less explicit about levels of analysis *per se*, several commentators argued that full accounts of semantic cognition will require some form of structured symbolic representation (**Opfer & Doumas**, **Marcus & Keil**) and/or the exploitation of "rational analysis" (**Quinlan**, **Feeney et al.**). Thus, while our following comments directly address Kemp & Tenenbaum's remarks, they may be relevant to the points of some of these other commentators as well.

We agree that there are different ways of thinking about levels of analysis and the place of connectionist models in these taxonomies. We do not agree, however, that the difference between our approach and structured probabilistic approaches adopted by **Kemp & Tenenbaum** is primarily one of levels. Connectionist approaches may seem to be cast at a different level from some other approaches, but we believe this is a misperception. To us, the difference between our approach and Kemp & Tenenbaum's is one of overall framework, and cuts across Marr's (1982) computational, representational/algorithmic, and implementational levels. In what follows, we briefly contrast two such frameworks in cognitive science. Both take seriously the issue of levels – the idea that there is a higher level of fundamental principles, above the level of algorithms and implementation details. They differ, however, in their claims about what these fundamental principles are.

On one approach, which we will call the structuralist framework, theorists explain cognitive phenomena with reference to explicitly stipulated domain-specific relationships among structured symbolic objects – as Chomsky did, for example, in *Syntactic Structures* (Chomsky 1957). The fundamental principles are those that govern the relationships, and these are taken as the ultimate explanatory basis for the characteristics of the domain itself. Different cognitive domains behave according to different principles, and the principles themselves specify both the structure of items in the domain and the conditions that determine which items are part of the domain (e.g., grammatical) and which items are not (e.g., ungrammatical).

For its advocates (which appear to include **Opfer & Doumas** and **Marcus & Keil**, along with **Kemp & Tenenbaum)** the structuralist approach appears to

address a key challenge in language: On one hand, languages clearly have structure; but on the other, the actual structure of real languages appears to be quite complex and often somewhat arbitrary. The structuralist approach appeals to some because it suggests that the complex and arbitrary-seeming structure of real languages might arise from much simpler and more regular underlying structures that behave according to a relatively constrained set of fundamental principles. Other domains of cognition also exhibit their own characteristic structure, and so may similarly be amenable to examination through the structuralist framework (as proposed, e.g., by Keil 1981).

The alternative, emergentist perspective, with which we associate ourselves, begins with the assumption that domain structure arises from the non-obvious interplay of domain-general principles and processes operating under constraints imposed by the domain-specific cognitive tasks and environments with which humans and other intelligent beings are faced. In language, for instance, many emergentist theorists treat the idealized characteristics of sentences, as elucidated by linguists, not as the starting point for an explanation of language, but as one aspect of the empirical facts about language that need to be explained. For example, Joan Bybee (2001), a linguist whose vision of language places her outside the Chomskyian framework, has argued that the regularities found in language arise from domain-general principles of cognition as these come into play when faced with the task of communicating using a low-bandwidth communication channel (spoken language). A similar perspective has been proposed by MacWhinney (2006). Christiansen and Ellefson (2002), Hare and Elman (1995), and Lupyan and McClelland (2003) have used simple connectionist models to show how characteristics of sentential and morphological structure could arise over historical time from models of individual speaker/hearers that embody many of the same domain-general principles that we have articulated in *Semantic Cognition.*

To appreciate the basic intuition underlying the emergentist view, consider Hofstadter's (1995) example of the sand dune: a structure whose particular characteristics at a given moment emerge from the operation of multiple factors, including the mutual influences exerted on each other by particles of sand operating in concert with the effects of external forces such as winds, tides, and gravity. Sand dunes clearly have structure, but this structure is an emergent consequence of the elements and forces in play – a consequence that is to be explained by general principles that govern these elements and forces. Emergentists like Bybee argue that the structures of languages are like sand dunes, continually shaped by an ongoing interplay of forces that can produce both the systematic tendencies found in language, and arbitrary-seeming variations that are otherwise seen as distracting flies in the ointment.

With these perspectives in mind, we now turn to a brief consideration of the structured probabilistic approach advocated by **Kemp & Tenenbaum** in comparison to our framework. According to their approach (see also e.g., Tenenbaum et al. 2006), cognitive phenomena are best understood with reference to Bayesian inference processes operating over probabilistic graphical models which directly represent structure (e.g., they use a mutation hierarchy to represent generic properties of animals). So stated, our approaches seem similar in two senses: both appeal to completely domain-general underlying principles, and both are grounded in an optimal inference framework.

The point that connectionist models are grounded in optimization is important. It was not a main focus of our book, but it was and remains a crucial element of the framework, anticipated in important precursors of the PDP volumes, and subsequently explored in considerable depth (see e.g., Ackley et al. 1985; Hinton & Sejnowski 1983; MacKay 1992; McClelland 1998; Rumelhart 1977). While we do not see optimality as an easy guide to explanation of cognitive phenomena, we certainly believe that it is worthwhile to consider what is optimal in a given context, and to compare what is optimal to what is actually observed in human behavior.

There are, however, several important differences between the approaches, and these cut across Marr's levels. We focus here on differences at the level of the overall theory of the computation – in particular, the goal of the computations being performed – and differences at the level of representations and algorithms.

R3.1. Goal of the computation

For **Kemp & Tenenbaum**, and perhaps for other structuralists, it appears that the goal of learning is to discover the optimal representation of the underlying structure of the domain *per se*. The emergentist point of view entails a different goal for learning – namely, the goal of correctly capturing the statistical relationships between inputs and outputs experienced in an environment. This same goal underlies both backpropagation and Boltzmann machine learning (Ackley et al. 1985). Additional constraints on complexity are often incorporated and serve, in a domain-general way, to promote generalization to unseen inputs. Of course, the statistics that drive learning in these models ultimately depend on complex, abstract, nonlinear relationships and are not simply direct first-order relationships between observable variables – but in contrast to the structuralist view, the discovery of such structured representations is not itself the goal of the computation.

R3.2. Representations

According to the structured probabilistic approach of **Kemp & Tenenbaum**, the learner selects a specific structural form of representation from among a specified set of alternative forms or types. For example, Kemp & Tenenbaum suggest that learners select a probabilistic mutation hierarchy to represent the generic properties of animals, a continuous two-dimensional space to represent hues, and a one-dimensional space to represent the stances of Supreme Court justices. The approach does not require the theorist to pre-specify which structure is appropriate for a given domain, and so in this sense is domain general. It does appear to require, however, an enumerated catalog of structures available for selection, among which there must exist a structure suited to the domain which in turn must be imbued with

a sufficiently high prior probability of selection (see also Perfors et al. [2006] for related work on language).

Connectionist modelers like Hinton and ourselves, on the other hand, do not suggest that any such structures are available to the learner for selection; instead, learning operates in a completely continuous representational space. When the experiences provided by the environment actually arise from a specific form of structure (like a mutation hierarchy), a connectionist model will come to behave as though it has learned that structure, but the structure itself is not explicitly represented as such, nor is it selected from a set of possible alternative structures. As a consequence, learning is not constrained to be fully consistent with one of a set of pre-specified alternatives: under our approach, real structure in the environment need not exactly match some particular structure type. In our view, this continuous representational potential makes the connectionist framework ultimately more appropriate for new semantic learning.

R3.3. Algorithms

Kemp & Tenenbaum might argue that their approach makes no commitment at the algorithmic level: The goal is only to find the structure that best fits available data subject to priors and simplicity constraints. Yet in practice, they proceed differently than we do, in ways that seem consistent with other differences between the approaches. Specifically, they rely on comparative evaluation of alternative candidate structures and selection among these, a common approach used within the graphical models framework. This may be an importantly different kind of algorithm than the gradient-based algorithms employed in fully continuous neural networks. Apart from any consideration of their relative psychological or neural plausibility, there is reason to believe that fundamental issues of computational complexity favor gradient-based approaches (Bengio & Bengio 2000). Fully continuous neural network models now represent the state of the art in several areas of machine learning and ameliorate the curse of dimensionality that plagues many other approaches (Bengio et al. 2006; Hinton & Salakhutdinov 2006; Rajat et al. 2007; Ranzato et al. 2007).

Kemp & Tenenbaum suggest that structured probabilistic approaches are to be preferred because "they make more direct contact with previous psychological research on semantic cognition." To us, this reflects the fact that these models are but a small step away from the commitments to innate domain-specific structure that have been articulated by many of the researchers whose work we addressed in our book. Our work instead makes contact with a newer body of connectionist research that applies a common set of domain general principles to address phenomena in perception, attention, memory, language, development of naive physics concepts, and now the fundamental structure of semantic cognition.

R4. The future of cognitive science

In this response we have tried to clarify the nature of our approach and to address misconceptions about its potential to address semantic cognition and other phenomena in cognitive science. In concluding, we consider the two predictions ventured by **Kemp & Tenenbaum** about the future of cognitive science: First, that "researchers will eventually understand how structured approaches to semantic cognition are implemented in the brain," and second, that "deep insights at the neural level will only be possible once we have a deep understanding of the computations supported by structured semantic representations." Although we agree with their statement that such predictions are "notoriously unreliable," we offer alternative predictions of our own: That structured approaches will have an important role in our field, but that researchers will come to understand that this role is to provide useful approximate descriptions of emergent properties of cognitive systems. These descriptions will be understood to capture structure that emerges from the interplay of domain-general principles and mechanisms, interacting with the particular constraints that arise in domain-specific contexts.

References

The letters "a" and "r" before author's initials stand for target article and response references, respectively.

Ackley, D. H., Hinton, G. E. & Sejnowski, T. J. (1985) A learning algorithm for Boltzmann machines. *Cognitive Science* 9:147–69. [rTTR]

Ahn, W. (1998) Why are different features central for natural kinds and artifacts? The role of causal status in determining feature centrality. *Cognition* 69:135–78. [aTTR]

Ahn, W., Marsh, J. K. & Luhmann, C. C. (2002) Effect of theory-based feature correlations on typicality judgments. *Memory and Cognition* 30(1):107–18. [aTTR]

Akrami, A., Liu, Y., Treves, A. & Jagadeesh, B. (2006) Dynamics of neural response in inferotemporal cortex during categorical processing of natural images. *Society for Neuroscience Abstracts* 504.9. [EK]

Anderson, J. R. (1990) *The adaptive character of thought*. Erlbaum. [aTTR]
(1991) The adaptive nature of human categorization. *Psychological Review* 98(3):409–29. [CK, rTTR]

Anderson, J. R. & Betz, J. (2001) A hybrid model of categorization. *Psychonomic Bulletin and Review* 8(4):629–47. [GFM]

Bahrick, L. E., Gogate, L. J. & Ruiz, I. (2002) Attention and memory for faces and actions in infancy: The salience of actions over faces in dynamic events. *Child Development* 73:1629–43. [JMM]

Balaban, M. T. & Waxman, S. R. (1997) Do words facilitate object categorization in 9-month-old infants? *Journal of Experimental Child Psychology* 64:3–26. [DHR]

Barsalou, L., Simmons, W., Barbey, A. & Wilson, C. D. (2003) Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences* 7(2):84–91. [aTTR]

Barsalou, L. W. (1987) The instability of graded structure: Implications for the nature of concepts. In: *Concepts and conceptual development: Ecological and intellectual factors in categorization*, ed. U. Neisser, pp. 101–40. Cambridge University Press. [JAH]
(2003) Situated simulation in the human conceptual system. In: *Conceptual representation*, ed. H. E. Moss & J. A. Hampton, pp. 513–62. Psychology Press. [JAH]

Bengio, S. & Bengio, Y. (2000) Taking on the curse of dimensionality in joint distributions using neural networks. *IEEE Transactions on Neural Networks* (special issue on data mining and knowledge discovery) 11:550–57. [rTTR]

Bengio, Y., Lamblin, P., Popovici, D. & Larochelle, H. (2006) Greedy layer-wise training of deep networks. In: *Advances in neural information processing systems (NIPS)*. MIT Press. [rTTR]

Bomba, P. C. & Siqueland, E. R. (1983) The nature and structure of infant form categories. *Journal of Experimental Child Psychology* 35:294–328. [aTTR]

Borges, J. L. (1998) On the exactitude of science. In: *Collected fictions*, trans. A. Hurley, p. 325. Penguin. [rTTR]

Boyd, R. (1986) Natural kinds, homeostasis, and the limits of essentialism. Unpublished manuscript, Cornell University. [aTTR]

Braitenberg, V. & Schüz, A. (1991) *Anatomy of the cortex: Statistics and geometry*. Springer-Verlag. [EK]

Brentano, F. C. (1874/1995) *Psychology from an empirical standpoint*. Routledge. [DB]

Broadbent, D. (1985) A question of levels: Comment on McClelland and Rumelhart. *Journal of Experimental Psychology: General* 114(2):189–92. [CK]

Brown, P. (2007) "She had just cut/broken off her head": Cutting and breaking verbs in Tzeltal. *Cognitive Linguistics* 18:319–30. [AM]

Brown, R. (1958) How shall a thing be called? *Psychological Review* 65:14–21. [aTTR]

Bybee, J. (2001) *Phonology and language use*. Cambridge University Press. [rTTR]

Carey, S. (1985) *Conceptual change in childhood*. MIT Press. [JEO, PCQ, aTTR]

(2000) The origin of concepts. *Journal of Cognition and Development* 1:37–42. [PCQ]

Carey, S. & Spelke, E. (1994) Domain-specific knowledge and conceptual change. In: *Mapping the mind: Domain specificity in cognition and culture*, ed. L. A. Hirschfeld & S. Gelman, pp. 169–200. Cambridge University Press. [aTTR]

Carlson, G. & Pelletier, J., eds. (1995) *The generic book*. University of Chicago Press. [AGBtM]

Chomsky, N. (1957) *Syntactic structure*. Mouton. [rTTR]

(1980) Rules and representations. *Behavioral and Brain Sciences* 3:1–61. [aTTR]

Christiansen, M. H. & Ellefson, M. (2002) Linguistic adaptation without linguistic constraints: The role of sequential learning in language evolution. In: *The transition to language*, ed. A. Wray, pp. 335–58. Oxford University Press. [rTTR]

Cleeremans, A. (1993) *Mechanisms of implicit learning: Connectionist models of sequence processing*. MIT Press. [aTTR]

Cleeremans, A. & McClelland, J. L. (1991) Learning the structure of event sequences. *Journal of Experimental Psychology: General* 120:235–53. [aTTR]

Cohen, J. D., Dunbar, K. & McClelland, J. L. (1990) On the control of automatic processes: A parallel distributed processing model of the Stroop effect. *Psychological Review* 97:332–61. [rTTR]

Cohen, J. D. & Servan-Schreiber, D. (1992) Context, cortex and dopamine: A connectionist approach to behavior and biology in schizophrenia. *Psychological Review* 99(1):45–77. [AWM]

Collins, A. M. & Loftus, E. F. (1975) A spreading-activation theory of semantic processing. *Psychological Review* 82:407–28. [aTTR]

Collins, A. M. & Quillian, M. R. (1969) Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior* 8:240–47. [DB, arTTR]

Cree, G., McRae, K. & McNorgan, C. (1999) An attractor model of lexical conceptual processing: Simulating semantic priming. *Cognitive Science* 23(4):371–414. [rTTR]

Crisp, A. K., Feeney, A. & Shafto, P. (under review) Testing dual process accounts of category-based conjunction fallacy: When is decontextualised reasoning necessary for logical responding? [AF]

Damasio, A. R. (1989) The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation* 1:123–32. [aTTR]

Dienes, Z., Altmann, G. & Gao, S. (1999) Mapping across domains without feedback: A neural-network model of transfer of implicit knowledge. *Cognitive Science* 23(1):53–82. [rTTR]

Dilkina, K., McClelland, J. L. & Boroditsky, L. (2007) How language affects thought in a connectionist model. In: *Proceedings of the 29th Annual Cognitive Science Society Conference*, pp. 215. Erlbaum. [rTTR]

Dilkina, K., McClelland, J. L. & Plaut, D. C. (2008) A single-system account of semantic and lexical deficits in five semantic dementia patients. *Cognitive Neuropsychology* 25(2):136–64. [rTTR]

Doumas, L. A. A., Hummel, J. E. & Sandhofer, C. M. (2008) A theory of the discovery and predication of relational concepts. *Psychological Review* 115(1):1–43. [CK, JEO]

Eimas, P. D. & Quinn, P. C. (1994) Studies on the formation of perceptually based basic-level categories in young infants. *Child-Development* 65(3):903–17. [JMM, aTTR]

Ellis, H. D. & Lewis, M. B. (2001) Capgras delusion: A window on face recognition. *Trends in Cognitive Science* 5(4):149–56. [rTTR]

Ellis, H. D. & Young, A. W. (1990) Accounting for delusional misidentifications. *British Journal of Psychiatry* 157:239–48. [AWM]

Ellis, H. D., Young, A. W., Quayle, A. H. & De Pauw, K. W. (1997) Reduced autonomic responses to faces in Capgras delusion. *Proceedings of the Royal Society of London B* 264:1085–92. [AWM]

Elman, J. L. (1990) Finding structure in time. *Cognitive Science* 14:179–211. [arTTR]

(1991) Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning* 7:194–220. [aTTR]

Evans, J. St. B. T. (2006) The heuristic-analytic theory of reasoning: Extension and evaluation. *Psychonomic Bulletin and Review* 13:378–95. [AF]

Evans, J. St. B. T. & Over, D. E. (2004) *If*. Oxford University Press. [AF]

Farah, M. & McClelland, J. L. (1991) A computational model of semantic memory impairment: Modality-specificity and emergent category-specificity. *Journal of Experimental Psychology: General* 120:339–57. [rTTR]

Feeney, A. (2007) How many processes underlie category-based induction? Effects of conclusion specificity and cognitive ability. *Memory and Cognition* 35:1830–39. [AF]

Feeney, A., Shafto, P. & Dunning, D. (2007) Who is susceptible to conjunction fallacies in category-based induction? *Psychonomic Bulletin and Review* 14:884–89. [AF]

Fodor, J. (1998) *Concepts: Where cognitive science went wrong*. Oxford University Press. [rTTR, JS]

(2000) *The mind doesn't work that way: The scope and limits of computational psychology*. MIT Press/Bradford Books. [aTTR]

Fodor, J. A. & McLaughlin, B. P. (1990) Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition* 35:183–204. [PQ]

Fodor, J. A. & Pylyshyn, Z. W. (1988) Connectionism and cognitive architecture: A critical analysis. *Cognition* 28:3–71. [CK, rTTR]

French, R. M. (1990) Sub-cognition and the limits of the Turing test. *Mind* 99:53–65. [SER]

(1999) When coffee cups are like old elephants, or why representation modules don't make sense. In: *Understanding representation in the cognitive sciences*, ed. A. Riegler, M. Peshl & A. von Stein, pp. 158–63. Plenum. [SER]

Frith, C. D. (2004) The pathology of experience. *Brain* 127:239–42. [AWM]

Fulkerson, A. L. & Haaf, R. A. (2003) The influence of labels, non-labeling sounds, and source of auditory input on 9- and 15-month-olds' object categorization. *Infancy* 4:349–69. [DHR]

Garcia, J. & Koelling, R. A. (1966) Relation of cue to consequence in avoidance learning. *Psychonomic Science* 4(3):123–24. [aTTR]

Gauthier, I. & Tarr, M. J. (2002) Unraveling mechanisms for expert object recognition: Bridging brain activity and behavior. *Journal of Experimental Psychology: Human Perception and Performance* 28:431–46. [PCQ]

Gelman, R. & Williams, E. M. (1998) Enabling constraints for cognitive development and learning: A domain-specific epigenetic theory. In: *Handbook of child psychology, Vol. II: Cognition, perception and development, 5th edition*, ed. D. Kuhn & R. Siegler, pp. 575–630. Wiley. [aTTR]

Gelman, S. A. (2003) *The essential child: Origins of essentialism in everyday thought*. Oxford University Press. [PCQ]

Gelman, S. A. & Wellman, H. M. (1991) Insides and essences: Early understandings of the nonobvious. *Cognition* 38:213–44. [aTTR]

Gentner, D. (1983) Structure-mapping: A theoretical framework for analogy. *Cognitive Science* 7(2):155–70. [SER, rTTR]

(1989) The mechanisms of analogical learning. In: *Similarity and analogical reasoning*, ed. S. Vosniadou & A. Ortony, pp. 199–241. Cambridge University Press. [JEO]

Gentner, D. & Markman, A. B. (1993) Analogy – Watershed or Waterloo? Structural alignment and the development of connectionist models of analogy. In: *Advances in neural information processing systems, vol. 5*, ed. S. J. Hanson, J. D. Cowan & C. L. Giles, pp. 855–62. Morgan Kaufmann. [GFM]

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Schulz, T. & Danks, D. (2004) A theory of causal learning in children: Causal maps and Bayes nets. *Psychological Review* 111(1):131. [aTTR]

Gopnik, A. & Meltzoff, A. N. (1997) *Words, thoughts, and theories*. MIT Press. [aTTR]

Gopnik, A. & Sobel, D. M. (2000) Detecting blickets: How young children use information about novel causal powers in categorization and induction. *Child Development* 71(5):1205–22. [aTTR]

Gopnik, A. & Wellman, H. M. (1994) The theory theory. In: *Mapping the mind: Domain specificity in cognition and culture*, ed. L. A. Hirschfeld & S. A. Gelman. Cambridge University Press. [aTTR]

Greif, M. L., Kemler Nelson, D. G., Keil, F. C. & Gutierrez, F. (2006) What do children want to know about animals and artifacts? Domain-specific requests for information. *Psychological Science* 17(6):455–59. [GFM]

Hadley, R. F. (2000) Cognition and the computational power of connectionist networks. *Connection Science* 12:95–110. [DB]

Hampton, J. A. (1982) A demonstration of intransitivity in natural categories. *Cognition* 12:151–64. [JAH]

(1984) The verification of category and property statements. *Memory and Cognition* 12:345–54. [JAH]

(1993) Prototype models of concept representation. In: *Categories and concepts: Theoretical views and inductive data analysis*, ed. I. Van Mechelen, J. A. Hampton, R. S. Michalski & P. Theuns, pp. 64–83. Academic Press. [aTTR]

Hampton, J. A., Estes, Z. & Simmons, S. (2007) Metamorphosis: Essence, appearance, and behavior in the categorization of natural kinds. *Memory and Cognition* 35:1785–1800. [PCQ]

Handley, S. J., Capon, A., Beveridge, M., Dennis, I. & Evans, J. St. B. T. (2004) Working memory and inhibitory control in the development of children's reasoning. *Thinking and Reasoning* 10:175–96. [AF]

Hare, M. & Elman, J. (1995) Learning and morphological change. *Cognition* 56(1):61–98. [rTTR]

Hinton, G. E. (1981) Implementing semantic networks in parallel hardware. In: *Parallel models of associative memory*, ed. G. E. Hinton & J. A. Anderson, pp. 161–87. Erlbaum. [arTTR]

(1984) *Distributed representations*. (Technical Report CMU-CS-84–157). Department of Computer Science, Carnegie Mellon University. [PQ]

(1986) Learning distributed representations of concepts. In: *Proceedings of the 8th Annual Conference of the Cognitive Science Society*, pp. 1–12. Erlbaum. [arTTR]

(1989) Learning distributed representations of concepts. In: *Parallel distributed processing: Implications for psychology and neurobiology*, ed. R. G. M. Morris, pp. 46–61. Clarendon Press. [rTTR]

Hinton, G. E. & McClelland, J. L. (1988) Learning representations by recirculation. In: *Neural information processing systems*, ed. D. Z. Anderson, pp. 358–66. American Institute of Physics. [rTTR]

Hinton, G. E. & Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks. *Science* 313:504–507. [rTTR]

Hinton, G. E. & Sejnowski, T. J. (1983) Optimal perceptual inference. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, June 1983*, pp. 448–53. Computer Society of the IEEE/IEEE Press. [rTTR]

(1986) Learning and relearning in Boltzmann machines. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1*, ed. D. E. Rumelhart & J. L. McClelland, pp. 282–317. MIT Press. [aTTR]

Hofstadter, D. (1995) *Fluid concepts and creative analogies*. Basic Books. [rTTR]

Holyoak, K. J. & Thagard, P. (1995) *Mental leaps: Analogy in creative thought*. MIT Press. [JEO]

Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA* 79:2554–58. [EK]

Hummel, J. E. & Holyoak, K. J. (2003) A symbolic-connectionist theory of relational inference and generalization. *Psychological Review* 110:220–63. [JEO]

James, W. (1890) *Principles of psychology, vol. 1*. Holt. [DHR]

Johnson-Laird, P. N. (2006) *How we reason*. Oxford University Press. [AF]

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

Jönsson, M. L. & Hampton, J. A. (2006) The inverse conjunction fallacy. *Journal of Memory and Language* 55:317–34. [JAH]

Kamp, H. & Reyle, U. (1993) *From discourse to logic*. Kluwer. [AGBtM]

Kay, P. & Regier, T. (2003) Resolving the question of color naming universals. *Proceedings of the National Academy of Sciences USA* 100:9085–89. [AM]

Keil, F. C. (1979) *Semantic and conceptual development: An ontological perspective*. Harvard University Press. [CK, aTTR]

(1981) Constraints on knowledge and cognitive development. *Psychological Review* 88(3):197–227. [rTTR]

(1989) *Concepts, kinds, and cognitive development*. MIT Press. [GFM, JEO, PCQ, aTTR]

(1991a) The emergence of theoretical beliefs as constraints on concepts. In: *The epigenesis of mind: Essays on biology and cognition*, ed. S. Carey & R. Gelman. Erlbaum. [aTTR]

(1991b) Theories, concepts, and the acquisition of word meaning. In: *Perspectives on language and cognition: Interrelations in development*, ed. J. P. Byrnes & S. A. Gelman. Cambridge University Press. [GFM]

(1994) The birth and nurturance of concepts by domains: The origins of concepts of living things. In: *Mapping the mind: Domain specificity in cognition and culture*, ed. L. A. Hirschfeld & S. A. Gelman, pp. 234–54. Cambridge University Press. [aTTR]

Kemp, C., Goodman, N. D. & Tenenbaum, J. B. (2007a) Learning causal schemata. In: *Proceedings of the 29th Annual Conference of the Cognitive Science Society*, ed. D. S. McNamara & J. G. Trafton, pp. 389–94. Erlbaum. [CK]

(2008) Learning and using relational theories. In: *Advances in neural information processing systems, vol. 20*, ed. J. C. Platt, D. Koller, Y. Singer & S. Roweis, pp. 753–60. MIT Press. [CK]

Kemp, C., Perfors, A. & Tenenbaum, J. B. (2004) Learning domain structures. In: *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, ed. K. Forbus, D. Gentner & T. Regier, pp. 672–78. Erlbaum. [CK]

Kemp, C., Shafto, P., Berke, A. & Tenenbaum, J. B. (2007b) Combining causal and similarity-based reasoning. In: *Advances in neural information processing systems, vol. 19*, ed B. Schölkopf, J. Platt & T. Hoffman, pp 681–88. MIT Press. [CK]

Kemp, C. & Tenenbaum, J. B. (2003) Theory-based induction. In: *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, ed. R. Alterman & D. Kirsh, pp. 658–63. Erlbaum. [CK]

Kropff, E. (forthcoming) Full solution for the storage of correlated memories in an autoassociative memory. In: *Proceedings of the International Meeting on "Closing the Gap Between Neurophysiology and Behaviour: A Computational Modelling Approach," Birmingham, U.K., May 2007*. Available at: http://arxiv.org/abs/0707.3066. [EK]

Kropff, E. & Treves, A. (2007) Uninformative memories will prevail: The storage of correlated representations and its consequences. *Human Frontier Science Program Journal* 1(4):249–62. [EK]

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

Kugler, P. & Turvey, M. (1987) *Information, natural law, and the self-assembly of rhythmic movement*. Erlbaum. [SER]

Lambon Ralph, M. A., Lowe, C. & Rogers, T. T. (2007) The neural basis of category-specific semantic deficits for living things: Evidence from semantic dementia, HSVE and a neural network model. *Brain* 130:1127–37. [rTTR]

Lascarides, A. & Asher, N. (1993) Temporal interpretation, discourse relations and commonsense entailment. *Linguistics and Philosophy* 16(5):437–93. [AGBtM]

Laurence, S. & Margolis, E. (1999) Concepts and cognitive science. In: *Concepts: Core readings*, ed. E. Margolis & S. Laurence, pp. 3–81. Bradford Books/MIT Press. [JS]

Levinson, S. C. (2007) Cut and break verbs in Yélî Dnye, the Papuan language of Rossel Island. *Cognitive Linguistics* 18:207–18. [AM]

Love, B. C., Medin, D. L. & Gureckis, T. M. (2004) SUSTAIN: A network model of category learning. *Psychological Review* 111(2):309–32. [GFM]

Lupyan, G. (2005) Carving nature at its joints and carving joints into nature: How labels augment category representations. In: *Modelling language, cognition and action: Proceedings of the 9th Neural Computation and Psychology Workshop*, ed. A. Cangelosi, G. Bugmann & R. Borisyuk, pp. 87–96. World Scientific. [DHR]

(in press) From chair to "chair": A representational shift account of object labeling effects on memory. *Journal of Experimental Psychology: General*. [DHR]

Lupyan, G. & McClelland, J. L. (2003) Did, made, had, said: Capturing quasi-regularity in exceptions. In: *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, ed. R. Alterman & D. Kirsh, pp. 740–45. Erlbaum. [rTTR]

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

MacKay, D. J. (1992) A practical Bayesian framework for backpropagation networks. *Neural Computation* 4:448–72. [arTTR]

MacWhinney, B. (2006) Emergentism: Use often and with care. *Applied Linguistics* 27:729–40. [rTTR]

Madole, K. L. & Oakes, L. M. (1999) Making sense of infant categorization: Stable processes and changing representations. *Developmental Review* 19:263–96. [DHR]

Majid, A., Boster, J. S. & Bowerman, M. (forthcoming) The crosslinguistic categorization of everyday events: A study of cutting and breaking. [AM]

Majid, A., Bowerman, M., van Staden, M. & Boster, J. S. (2007a) The semantic categories of cutting and breaking: A crosslinguistic perspective. *Cognitive Linguistics* 18:133–52. [AM]

Majid, A., Gullberg, M., van Staden, M. & Bowerman, M. (2007b) How similar are semantic categories in closely related languages? A comparison of cutting and breaking in four Germanic languages. *Cognitive Linguistics* 18:179–94. [AM]

Malt, B. C. (1995) Category coherence in cross-cultural perspective. *Cognitive Psychology* 29:85–148. [AM]

Malt, B. C., Sloman, S. A., Gennari, S., Shi, M. & Wang, Y. (1999) Knowing versus naming: Similarity and the linguistic categorization of artifacts. *Journal of Memory and Language* 40:230–62. [AM]

Mandler, J. M. (2000) Perceptual and conceptual processes in infancy. *Journal of Cognition and Development* 1:3–36. [PCQ]

(2004) *The foundations of mind: Origins of conceptual thought*. Oxford University Press. [JMM]

Mandler, J. M., Bauer, P. J. & McDonough, L. (1991) Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology* 23:263–98. [aTTR]

Mandler, J. M. & McDonough, L. (1993) Concept formation in infancy. *Cognitive Development* 8:291–318. [JMM, PCQ, aTTR]

(1996) Drinking and driving don't mix: Inductive generalization in infancy. *Cognition* 59:307–55. [aTTR]

(1998) On developing a knowledge base in infancy. *Developmental Psychology* 34:1274–88. [JMM]

Marcus, G. F. (1998a) Can connectionism save constructivism? *Cognition* 66:153–82. [GFM]

(1998b) Rethinking eliminative connectionism. *Cognitive Psychology* 37(3):243–82. [GFM]

(2000) Children's overregularization and its implications for cognition. In: *Cognitive models of language acquisition*, ed. P. Broeder & J. Murre. Oxford University Press. [GFM]

(2001) *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press. [CK, GFM, rTTR]

Mareschal, D. (2000) Infant object knowledge: Current trends and controversies. *Trends in Cognitive Science* 4:408–16. [aTTR]

Marr, D. (1971) Simple memory: A theory for archicortex. *Philosophical Transactions of the Royal Society of London, B* 262:23–81. [EK, aTTR]

(1982) *Vision*. Freeman. [arTTR]

Martin, A. & Chao, L. L. (2001) Semantic memory in the brain: Structure and processes. *Current Opinion in Neurobiology* 11:194–201. [aTTR]

Massaro, D. (1988) Some criticisms of connectionist models of human performance. *Journal of Memory and Language* 27:213–34. [DHR]

Massey, C. M. & Gelman, R. (1988) Preschooler's ability to decide whether a photographed unfamiliar object can move by itself. *Developmental Psychology* 24(3):307–17. [aTTR]

McClelland, J. L. (1989) Parallel distributed processing: Implications for cognition and development. In: *Parallel distributed processing: Implications for psychology and neurobiology*, ed. R. G. M. Morris, pp. 8–45. Oxford University Press. [aTTR]

(1991) Stochastic interactive activation and the effect of context on perception. *Cognitive Psychology* 23:1–44. [aTTR]

(1994) Learning the general but not the specific. *Current Biology* 4:357–58. [aTTR]

(1995) A connectionist perspective on knowledge and development. In: *Developing cognitive competence: New approaches to process modeling*, ed. T. J. Simon & G. S. Halford, pp. 157–204. Erlbaum. [rTTR]

(1998) Connectionist models and Bayesian inference. In: *Rational models of cognition*, ed. M. Oaksford & N. Chater, pp. 21–53. Oxford University Press. [arTTR]

McClelland, J. L., McNaughton, B. L. & O'Reilly, 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* 102:419–57. [JMM, aTTR, SER]

McClelland, J. L. & Rumelhart, D. E. (1986) A distributed model of human learning and memory. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group, pp. 170–215. MIT Press. [aTTR]

McClelland, J. L., St. John, M. F. & Taraban, R. (1989) Sentence comprehension: A parallel distributed processing approach. *Language and Cognitive Processes* 4:287–335. [aTTR]

McGeoch, J. A. (1942) *The psychology of human learning*. Longman, Greene. [SER]

McRae, K. & Cree, G. (2002) Factors underlying category-specific deficits. In: *Category specificity in brain and mind*, ed. E. M. E. Forde & G. Humphreys, pp. 211–49. Psychology Press. [rTTR]

McRae, K., Cree, G., Seidenberg, M. & McNorgan, C. (2005) Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, Instruments, and Computers* 37(4):547–59. [EK. rTTR]

McRae, K., De Sa, V. & Seidenberg, M. (1997) On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General* 126(2):99–130. [rTTR]

Medin, D., Coley, J. D., Storms, G. & Hayes, B. (2003) A relevance theory of induction. *Psychonomic Bulletin and Review* 10:517–32. [AF]

Mermillod, M., French, R. M., Quinn, P. C. & Mareschal, D. (2004) The importance of long-term memory in infant perceptual categorization. In: *Proceedings of the 25th Annual Conference of the Cognitive Science Society*, ed. R. Alterman & D. Kirsh, pp. 804–809. Erlbaum. [PCQ]

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

Miikkulainen, R. & Dyer, M. G. (1987) *Building distributed representations without microfeatures*. (Technical Report No. UCLA-AI-87–17). University of California–Los Angeles, Department of Computer Science. [rTTR]

Morris, J. S., Ohman, A. & Dolan, R. J. (1999) A subcortical pathway to the right amygdala mediating "unseen" fear. *Proceedings of the National Academy of Sciences USA* 96:1680–85. [AWM]

Movellan, J. & McClelland, J. L. (2001) The Morton-Massaro law of information integration: Implications for models of perception. *Psychological Review* 108:113–48. [aTTR]

Mummery, C. J., Patterson, K., Price, C. J., Ashburner, J., Frackowiak, R. S. J. & Hodges, J. (2000) A voxel-based morphometry study of semantic dementia: Relationship between temporal lobe atrophy and semantic memory. *Annals of Neurology* 47(1):36–45. [aTTR]

Munakata, Y. & McClelland, J. L. (2003) Connectionist models of development. *Developmental Science* 6(4):413–29. [aTTR]

Munakata, Y., McClelland, J. L., Johnson, M. H. & Siegler, R. (1997) Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review* 104:686–713. [aTTR]

Murphy, G. L. (2000) Explanatory concepts. In: *Explanation and cognition*, ed. F. Keil & R. A. Wilson, pp. 361–92. MIT Press. [rTTR]

(2002) *The big book of concepts*. MIT Press. [PCQ, aTTR]

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

Murphy, G. L. & Wright, J. C. (1984) Changes in conceptual structure with expertise: Differences between real-world experts and novices. *Journal of Experimental Psychology/Learning, Memory, and Cognition* 10:144–55. [PCQ]

Nazzi, T. & Gopnik, A. (2001) Linguistic and cognitive abilities in infancy: When does language become a tool for categorization? *Cognition* 80:B11–B20. [DHR]

Nisbett, R. E. & Wilson, T. D. (1977) Telling more than we can know: Verbal reports on mental processes. *Psychological Review* 84(3):231–59. [aTTR]

Nosofsky, R. M. (1984) Choice, similarity, and the context theory of classification. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 10:104–10. [aTTR]

(1986) Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General* 115(1):39–61. [GFM, aTTR]

O'Reilly, R. (1996) The LEABRA model of neural interactions and learning in the neocortex. Unpublished doctoral dissertation, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA. [rTTR]

Oaksford, M. & Chater, N., eds. (1998) *Rational models of cognition*. Oxford University Press. [aTTR]

Opfer, J. E. (2002) Identifying living and sentient kinds from dynamic information: The case of goal-directed versus aimless autonomous movement in conceptual change. *Cognition* 86:97–122. [JEO]

Opfer, J. E. & Bulloch, M. J. (2007) Causal relations drive young children's induction, naming, and categorization. *Cognition* 105:207–17. [JEO]

Opfer, J. E. & Siegler, R. S. (2004) Revisiting preschoolers' living things concept: A microgenetic analysis of conceptual change in basic biology. *Cognitive Psychology* 49:301–32. [JEO]

Osherson, D. N., Smith, E. E., Wilkie, O., López, A. & Shafir, E. (1990) Category-based induction. *Psychological Review* 97:185–200. [AF]

Paivio, A. (1971) *Imagery and verbal processes*. Holt, Rinehart, and Winston. [SER]

Pauen, S. (2002a) Evidence for knowledge-based category discrimination in infancy. *Child Development* 73(4):1016–33. [PCQ, aTTR]

(2002b) The global-to-basic shift in infants' categorical thinking: First evidence from a longitudinal study. *International Journal of Behavioural Development* 26(6):492–99. [aTTR]

Perfors, A., Tenenbaum, J. & Regier, T. (2006) Poverty of the stimulus? A rational approach. Paper presented at the 28th Annual Conference of the Cognitive Science Society, Vancouver, British Columbia. [rTTR]

Pinker, S. & Prince, A. (1988) On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28:73–193. [GFM]

Plaut, D. C. & Shallice, T. (1993) Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology* 10(5):377–500. [rTTR]

Quinn, P. C. (2004a) Is the asymmetry in young infants' categorization of humans versus nonhuman animals based on head, body, or global Gestalt Information? *Psychonomic Bulletin and Review* 11:92–97. [PCQ]

(2004b) Multiple sources of information and their integration, not dissociation, as an organizing framework for understanding infant concept formation. *Developmental Science* 7:511–13. [PCQ]

(2005) Young infants' categorization of humans versus nonhuman animals: Roles for knowledge access and perceptual process. In: *Building object categories in developmental time: 32nd Carnegie Symposium on cognition, vol. 32*, ed. L. Gershkoff-Stowe & D. Rakison, pp. 107–30. Erlbaum. [PCQ]

Quinn, P. C. & Eimas, P. D. (1998) Evidence for a global categorical representation of humans by young infants. *Journal of Experimental Child Psychology* 69:151–74. [PCQ]

Quinn, P. C. & Johnson, M. H. (2000) Global-before-basic object categorization in connectionist networks and 2-month-old infants. *Infancy* 1:31–46. [JMM, aTTR]

Quinn, P. C., Lee, K., Pascalis, O. & Slater, A. M. (2007) In support of an expert-novice difference in the representation of humans versus non-human animals by infants: Generalization from persons to cats occurs only with upright whole

images. *Cognition, Brain, and Behavior (Special Issue on the Development of Categorization)* 11:679–94. [PCQ]
Rajat, R., Battle, A., Lee, H., Packer, B. & Ng, A. Y. (2007) Self-taught learning: Transfer learning from unlabeled data. In: *Proceedings of the 24th ACM International Conference on Machine Learning, 2007. ACM International Conference Proceedings Series, vol. 227*, pp. 759–66. [rTTR]
Rakison, D. H. (2005) Developing knowledge of motion properties in infancy. *Cognition* 96:183–214. [DHR]
Rakison, D. H. & Lupyan, G. (in press) Developing object concepts in infancy: An associative learning perspective. *Monographs of SRCD*. [DHR]
Ranzato, M., Poultney, C., Chopra, A. & LeCun, Y. (2007) Efficient learning of sparse representations with an energy-based model. In: *Advances in neural information processing systems (NIPS)*, ed. B. Schölkopf, J. Platt & T. Hoffman, pp. 1137–44. MIT Press. [rTTR]
Rehder, B. (2003) A causal-model theory of conceptual representation and categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29(6):1141–59. [GFM]
Rehder, B. & Kim, S. (2006) How causal knowledge affects classification: A generative theory of categorization. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 32(4):659–83. [GFM]
Richland, L. E., Morrison, R. G. & Holyoak, K. J. (2006) Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology* 94:249–73. [JEO]
Ripley, B. D. (1996) *Pattern recognition and neural networks*. Cambridge University Press. [DB]
Rips, L. J. (1994) *The psychology of proof*. MIT Press. [AF]
(2001) Necessity and natural categories. *Psychological Bulletin* 127:827–52. [JAH]
Rips, L. J., Shoben, E. J. & Smith, E. E. (1973) Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior* 12:1–20. [aTTR]
Robbins, S. E. (2002) Semantics, experience and time. *Cognitive Systems Research* 3:301–35. [SER]
(2004) On time, memory and dynamic form. *Consciousness and Cognition* 13:762–88. [SER]
(2006a) On the possibility of direct memory. In: *New developments in consciousness research*, ed. V. W. Fallio, pp. 1–64. Nova Science. [SER]
(2006b) Bergson and the holographic theory of mind. *Phenomenology and the Cognitive Sciences* 5:365–94. [SER]
(2007) Time, form and the limits of qualia. *Journal of Mind and Behavior* 28:19–43. [SER]
Roberts, S. & Pashler, H. (2000) How persuasive is a good fit? A comment on theory testing. *Psychological Review* 107:358–67. [DHR]
Rogers, T. T., Lambon Ralph, M., Garrard, P., Bozeat, S., McClelland, J. L., Hodges, J. R. & Patterson, K. (2004) The structure and deterioration of semantic memory: A computational and neuropsychological investigation. *Psychological Review* 111 (1):205–35. [arTTR]
Rogers, T. T. & McClelland, J. L. (2004) *Semantic cognition: A parallel distributed processing approach*. MIT Press. [DB, AF, JAH, CK, EK, AM, AWM, JMM, GFM, JEO, PQ, PCQ, DHR, SER, arTTR, JS, AGBtM]
Rogers, T. T. & Patterson, K. (2007) Object categorization: Reversals and explanations of the basic-level advantage. *Journal of Experimental Psychology: General* 163(3):451–69. [rTTR]
Rohde, D. L. T. (2002) A connectionist model of sentence comprehension and production. Unpublished doctoral dissertation, Computer Science Department, Carnegie Mellon University, Pittsburgh, PA. (Available as Technical Report CMU-CS-02–105.) [rTTR]
Rohde, D. L. T. & Plaut, D. C. (1999) Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition* 72(1):67–109. [aTTR]
Rosch, E. R. (1978) Principles of categorization. In: *Cognition and categorization*, ed. E. R. Rosch & B. B. Lloyd, pp. 27–48. Erlbaum. [JAH]
Rosch, E. R. & Mervis, C. B. (1975) Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology* 7:573–605. [aTTR]
Rosch, E. R., Mervis, C. B., Gray, W., Johnson, D. & Boyes-Braem, P. (1976) Basic objects in natural categories. *Cognitive Psychology* 8:382–439. [arTTR]
Roth, E. M. & Shoben, E. J. (1983) The effect of context on the structure of categories. *Cognitive Psychology* 15:346–78. [JAH]
Rumelhart, D. E. (1977) Toward an interactive model of reading. In: *Attention and performance, vol. VI*, ed. S. Dornic, pp. 573–603. Erlbaum. [rTTR]
(1990) Brain style computation: Learning and generalization. In: *An introduction to neural and electronic networks*, ed. S. F. Zornetzer, J. L. Davis & C. Lau, pp. 405–20. Academic Press. [arTTR]
Rumelhart, D. E., Durbin, R., Golden, R. & Chauvin, Y. (1995) Backpropagation: The basic theory. In: *Back-propagation: Theory, architectures, and applications*, ed. Y. Chauvin & D. E. Rumelhart, pp. 1–34. Erlbaum. [aTTR]
Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986a) Learning representations by back-propagating errors. *Nature* 323(9):533–36. [aTTR]
Rumelhart, D. E. & McClelland, J. L. (1985) Levels indeed! A response to Broadbent. *Journal of Experimental Psychology: General* 114(2):193–97. [CK]
(1986) PDP models and general issues in cognitive science. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1*, ed. D. E. Rumelhart, J. L. McClelland & the PDP Research Group, pp. 110–46. MIT Press. [rTTR]
Rumelhart D. E., McClelland, J. L. & the PDP Research Group, eds. (1986b) *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2*. MIT Press. [rTTR]
Rumelhart, D. E., Smolensky, P., McClelland, J. L. & Hinton, G. E. (1986c) Schemata and sequential thought processes in PDP models. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2*, ed. J. L. McClelland, D. E. Rumelhart & the PDP Research Group, pp. 7–57. MIT Press. [arTTR]
Rumelhart, D. E. & Todd, P. M. (1993) Learning and connectionist representations. In: *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, ed. D. E. Meyer & S. Kornblum, pp. 3–30. MIT Press. [arTTR]
Sarle, W. S. (1994) Neural networks and statistical models. In: *Proceedings of the Nineteenth Annual SAS Users Group International Conference*. pp. 1538–50. SAS Institute. [DB]
Savelsbergh, G. J. P., Whiting, H. T. & Bootsma, R. J. (1991) Grasping tau. *Journal of Experimental Psychology: Human Perception and Performance* 17:315–22. [SER]
Schapiro, A. C. & McClelland, J. L. (in press) A connectionist model of a continuous developmental transition in the balance scale task. *Cognition*. [rTTR]
Schmidt, L. A., Kemp, C. & Tenenbaum, J. B. (2006) Nonsense and sensibility: Discovering unseen possibilities. In: *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, ed. R. Sun & N. Miyake, pp. 744–49. Erlbaum. [CK]
Schmittmann, V. D., Visser, I. & Raijmakers, M. E. J. (2006) Multiple learning modes in the development of rule-based category-learning task performance. *Neuropsychologia* 44:2079–91. [DB]
Schultz, J., Imamizu, H., Kawato, M. & Frith, C. (2004) Activation of the human superior temporal gyrus during observation of goal attribution by intentional objects. *Journal of Cognitive Neuroscience* 16:1695–705. [JEO]
Shafto, P., Kemp, C., Baraff, E., Coley, J. & Tenenbaum, J. B. (2005) Context-sensitive induction. In: *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, ed. B. G. Bara, L. Barsalou & M. Bucciarelli, pp. 2003–2008. Erlbaum. [CK]
Shafto, P., Kemp, C., Mansinghka, V., Gordon, M. & Tenenbaum, J. B. (2006) Learning cross-cutting systems of categories. In: *Proceedings of the 28th Annual Conference of the Cognitive Science Society* ed. R. Sun & N. Miyake, pp. 2146–51. Erlbaum. [CK]
Shultz, T. R. & Vogel, A. (2004) A connectionist model of the development of transitivity. In: *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, ed. K. Forbus, D. Gentner & T. Regier, pp. 1243–48. Erlbaum. [CK]
Siegler, R. S. (2005) Models of categorization: What are the limits? In: *Building object categories in developmental time*, ed. L. Gershkoff-Stowe & D. H. Rakison, pp. 433–39. Erlbaum. [DHR]
Siegler, R. S. & Chen, Z. (1998) Developmental differences in rule learning: A microgenetic analysis. *Cognitive Psychology* 36(3):273–310. [rTTR]
Sloman, S. A. (1993) Feature based induction. *Cognitive Psychology* 25:231–80. [AF]
(1996) The empirical case for two systems of reasoning. *Psychological Bulletin* 119:3–22. [AF]
Smith, E. E. & Medin, D. L. (1981) *Categories and concepts*. Harvard University Press. [aTTR]
Smith, L. B. (2000) From knowledge to knowing: Real progress in the study of infant categorization. *Infancy* 1(1):91–97. [aTTR]
Smith, L. B. & Heise, D. (1992) Perceptual similarity and conceptual structure. In: *Percepts, concepts, and categories*, ed. B. Burns, pp. 233–72. North Holland. [PCQ]
Smolensky, P. (1986) Information processing in dynamical systems: Foundations of harmony theory. In: *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 1*, ed. D. E. Rumelhart & J. L. McClelland, pp. 194–281. MIT Press. [aTTR]
(1988) On the proper treatment of connectionism. *Behavioral and Brain Sciences* 11(1):1–23; discussion 23–74. [CK, rTTR]
Spelke, E. S., Breinlinger, K., Macomber, J. & Jacobson, K. (1992) Origins of knowledge. *Psychological Review* 99(4):605–32. [aTTR]

St. John, M. F. (1992) The story gestalt: A model of knowledge-intensive processes in text comprehension. *Cognitive Science* 16:271–306. [aTTR]

St. John, M. F. & McClelland, J. L. (1990) Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence* 46:217–57. [rTTR]

Stanovich, K. E. (1999) *Who is rational: Studies of individual differences in reasoning*. Erlbaum. [AF]

Steyvers, M., Griffiths, T. L. & Dennis, S. (2006) Probabilistic inference in human semantic memory. *Trends in Cognitive Science* 10:327–34. [PQ]

Tanaka, J. & Taylor, M. (1991) Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology* 23:457–82. [rTTR]

Tanaka, J. W. (2001) The entry point of face recognition: Evidence for face expertise. *Journal of Experimental Psychology: General* 130:534–43. [PCQ]

Tenenbaum, J. B. (2000) Rules and similarity in concept learning. In: *Advances in neural information processing systems, vol. 12*, ed. S. A. Solla, T. K. Leen & K. R. Muller, pp. 59–65. MIT Press. [CK]

Tenenbaum, J. B., Griffiths, T. & Kemp, C. (2006) Theory-based bayesian models of inductive learning and reasoning. *Trends in Cognitive Science* 10(7):309–18. [rTTR]

ter Meulen, A. (1995) *Representing time in natural language: The dynamic interpretation of tense and aspect*. MIT Press. [AGBtM]

(2000) Chronoscopes: The dynamic representation of facts and events. In: *Speaking of events*, ed. J. Higginbotham, F. Pianesi & A. Varzi, pp. 151–68. Oxford University Press. [AGBtM]

(2004) Dynamic definite descriptions, implicit arguments and familiarity. In: *Descriptions and beyond*, ed. M. Reimer & A. Bezuidenhout, pp. 344–57. Oxford University Press. [AGBtM]

(2006) Cohesion in temporal context: Aspectual adverbs as dynamic indexicals. In: *Comparative and cross-linguistic research in syntax and semantics: Negation, tense and clausal architecture*, ed. R. Zanuttini, H. Campos, E. Herburger & P. Portner, pp. 362–77. Georgetown University Press. [AGBtM]

Treves, A. (2005) Frontal latching networks: A possible neural basis for infinite recursion. *Cognitive Neuropsychology* 22:276–91. [EK]

Turvey, M. & Carello, C. (1995) Dynamic touch. In: *Perception of space and motion*, ed. W. Epstein & S. Rogers, pp. 401–90. Academic Press. [SER]

van Benthem, J. & ter Meulen, A. (1997) *Handbook of logic and language*. Elsevier/North-Holland/MIT Press. [AGBtM]

van Zaanen, M. (2001) Bootstrapping structure using similarity. Doctoral dissertation, University of Leeds. Available at: citeseer.ist.psu.edu/324452.html [AGBtM]

(2002) Implementing alignment-based learning. In: *Proceedings of the International Colloquium on Grammatical Inference (ICGI), vol. 2482*, ed. P. Adriaans, H. Fernau & M. van Zaanen, pp. 312–14. Springer-Verlag. Available at: http://www.ics.mq.edu.au/~menno/research/publications/ [AGBtM]

(2002) Implementing alignment-based learning. *Proceedings of the International Colloquium on Grammatical Inference (ICGI), Amsterdam*, pp. (2482)312–14. Springer Verlag. [AGBtM]

Warrington, E. & Shallice, T. (1984) Category specific semantic impairments. *Brain* 107(3):829–54. [EK]

Waxman, S. R. & Markow, D. B. (1995) Words as invitations to form categories: Evidence from 12- to 13-month-old infants. *Cognitive Psychology* 29:257–302. [DHR]

Wellman, H. M. & Gelman, S. A. (1997) Knowledge acquisition in foundational domains. In: *Cognition, perception and development*, 5th ed., vol. 2, ed. D. Kuhn & R. Siegler, pp. 523–73. Wiley. [aTTR]

Wigner, E. P. (1970) *Symmetries and reflections*. MIT Press. [SER]

Wilburn, C. J. & Feeney, A. (2007) What develops when, and does it all develop together? The development of sensitivity to amount and diversity of evidence in inductive reasoning. Paper presented at the Annual Conference of the British Psychological Society Developmental Section, Plymouth, UK, August 29–31, 2007. [AF]

Wills, T. J., Lever, C., Cacucci, F., Burgess, N. & O'Keefe, J. (2005) Attractor dynamics in the hippocampal representation of the local environment. *Science* 308:873–76. [EK]

Wilson, R. A. & Keil, F. C. (2000) The shadows and shallows of explanation. In: *Explanation and cognition*, ed. F. C. Keil & R. A. Wilson, pp. 87–114. MIT Press. [aTTR]

Woit, P. (2006) *Not even wrong: The failure of string theory and the search for unity in physical law*. Basic Books. [SER]

Woodward, J. (2000) Explanation and invariance in the special sciences. *British Journal for the Philosophy of Science* 51:197–214. [SER]

(2001) Law and explanation in biology: Invariance is the kind of stability that matters. *Philosophy of Science* 68:1–20. [SER]

Xu, F. (2002) The role of language in acquiring object kind concepts in infancy. *Cognition* 85:223–50. [DHR]

Xu, F. & Tenenbaum, J. B. (2007a) Word learning as Bayesian inference. *Psychological Review* 114:245–72. [PQ]

(2007b) Sensitivity to sampling in Bayesian word learning. *Developmental Science* 10:288–97. [PQ]

Younger, B. A. & Cohen, L. B. (1986) Developmental change in infants' perception of correlations among attributes. *Child Development* 57:803–15. [DHR]