Computational models of semantic memory

Timothy. T. Rogers

For inclusion in: *The Cambridge Handbook of Computational Cognitive Modeling*

1. Introduction

Consider the predicament of a young infant recently arrived in the world and trying to make sense of it. She has some resources at her disposal: sensory information about her environment, the ability to act on it, and in most cases a surrounding linguistic environment, family and culture that can help to teach her what she needs to know. Nevertheless the task is daunting. Suppose on one occasion that daddy gestures out the window and says "Look, a bunny!" To what is he referring? The field of green? The tall structures dotting the horizon? The brownish object streaking rapidly along the ground? Later in the evening mommy repeats the word, this time gesturing toward a white contour in a picture book—it is not moving, it is not brown, it is two-dimensional. At bedtime big brother says "Here's your bunny," this time handing her a soft pink fuzzy object. What on earth could they all be talking about!

And yet before she turns 10 she will know that the word "bunny" refers to a particular animal with long ears and a fluffy little tail, and what's more, she will know that bunnies have blood and bones inside; that they can reproduce and grow and die; that they can feel pain and get hungry; that they are warm to the touch; that they live in holes in the ground; that some people believe it brings good luck to wear a bunny-foot on a chain. When she gets a new bunny-rabbit as a pet, she will be able to infer that all of these things are true, even though she has never before encountered this <u>particular</u> bunny; and when she brings her new pet to show-and-tell, she will be able to communicate all of these facts to her classmates simply by talking. And, this knowledge about bunny-rabbits constitutes a tiny fraction of the general factual world-knowledge she will have accumulated. Understanding the basis of these human 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 memory.

Semantic memory is memory for meanings. In some disciplines (e.g. linguistics), the word semantics refers exclusively to the meanings of words and sentences. In cognitive science, however, the term typically encompasses knowledge of any kind of meaning, linguistic or non-linguistic, including knowledge about the meanings of words, sentences, objects and events, as well as general facts (Tulving, 1972). Accordingly, the terms "semantic memory" and "conceptual knowledge" are often used interchangeably in the literature. Semantic memory is usually differentiated from episodic memory (long-term declarative memory for particular episodes that are firmly rooted in a particular time and place; see the chapter by Norman, Detre and Polyn, this volume), procedural memory (long-term non-declarative memory for well-learned action sequences; see the chapters by Ohlsson and by Cleeremans in this volume), and working memory (short-term memory for retention and manipulation of task-relevant information; see chapter by De Pisapia, Repovs and Braver, this volume).

Semantic abilities are central to a broad swath of cognitive science, including language comprehension and production, object recognition, categorization, induction and inference, and reasoning. Each of these topics constitutes a domain of study in its own right, and many are covered in other chapters in the Handbook (see the chapter on exemplars models by Logan; the chapter on concepts and categorization by Kruschke; the chapter on induction and inference by Heit; and the chapter on Bayesian models by Griffiths, Kemp and Tenenbaum). This chapter focuses on three principal questions

motivating research in semantic memory: How do we come to know which items and events in the environment should be treated as "the same kind of thing" for purposes of communication, action, and induction; how do we learn to map language onto these kinds; and how are these cognitive abilities subserved by neural processes?

These questions have, of course, been the subject of philosophical inquiry for centuries, but the application of computational methods has considerably advanced our understanding of the cognitive and neural bases of semantic abilities. Indeed, semantic memory was the target of some of the earliest computer simulation work in cognitive science, and much contemporary research in the domain can be fruitfully viewed as a reaction to these early ideas. The next section of the chapter thus provides a brief overview of two theoretical frameworks that first came to prominence in the 1970's: spreading activation theories based on Collins and Quillian's (1969) influential computer model, and prototype theories deriving from the work of Eleanor Rosch (Rosch, 1978; Rosch & Mervis, 1975) and others. A consideration of the strengths and limitations of these basic ideas will highlight the most pressing questions guiding current research in semantic memory. The remaining sections then follow 3 parallel strands of modeling research that are beginning to offer leverage on this issues. Section 3 traces developments spurred by Hinton's (1981) Parallel Distributed Processing (PDP) model of semantics, culminating in the general approach to semantic cognition recently laid out by Rogers and McClelland (2004). Section 4 addresses how sensitivity to temporal structure in language and experience can shape conceptual representations, following a thread of research that begins with Elman's (1990) seminal work and culminates in Latent Semantic Analysis (LSA) and related approaches (Burgess & Lund, 1997; Landauer &

Dumais, 1997; Steyvers, Griffiths, & Dennis, 2006). Section 5 considers models targeted at understanding the neural basis of semantic abilities.

2. Hierarchies and prototypes

One of the earliest implemented computer models in cognitive science was the hierarchical spreading-activation model of semantic memory described by Collins and Quillian (Collins & Quillian, 1969). The model was predicated on the notion that semantic memory consists of a vast set of stored simple propositions (e.g. "cats have fur," "canaries can sing," and so on). Under the rules of logical inference, such a system of propositions can support new deductive inferences via the syllogism; for instance, given the propositions "Socrates is a man" and "all men are mortal," it is possible to infer that Socrates is mortal without requiring storage of a third proposition. Collins and Quillian's model effectively used the syllogism as a basis for organizing propositional knowledge in memory. In their model, concepts (mental representations of categories) are stored as nodes in a network, and predicates (specifying relationships between concepts) are stored as labeled links between nodes. Simple propositional beliefs are represented by linking two nodes with a particular predicate. For example, the belief that a robin is a kind of bird is represented by connecting the nodes robin and bird with a predicate that specifies class-inclusion (an ISA link, as in "a robin is a bird"); whereas the belief that birds can fly is represented by connecting the nodes bird and fly with a link labeled can, and so on.

The authors observed that, if concepts at different levels of specificity were linked with ISA predicates, the system could provide an economical means of knowledge storage and generalization. For instance, the knowledge that a canary is a kind of bird is represented by connecting the node for canary to the node for bird with an ISA link;

knowledge that birds are animals is stored by connecting the bird node to the animal node, and so on. To make inferences about the properties of a given concept such as canary, the model first retrieves all of the predicates stored directly with the corresponding node (e.g. can sing); but the search process then moves upward along the ISA links and searches properties at the next node, so that the predicates attached to more inclusive concepts also get attributed to the probe concept. For canary, activation first searches the bird node, supporting the inference that the canary can fly, and then the animal node, supporting the inference that the canary can move.

In addition to economy of storage, this system provided a simple mechanism of knowledge generalization; for example, to store the fact that all birds have a spleen, it is sufficient to create a node for spleen and connect it to the bird node with a link labeled has. The retrieval process will then ensure that has a spleen generalizes to all of the individual bird concepts residing beneath the bird node in the hierarchy. Similarly, if the system is "told" that there is something called a "Xxyzzyx" that is a kind of bird, it can store this information by creating a new node for Xxyzzyx and attaching it to the bird node. The retrieval mechanism will then ensure that all properties true of birds are attributed to the Xxyzzyx.

Early empirical assessments of the model appeared to lend some support to the notion that concepts were organized hierarchically in memory. Specifically, Collins and Quillian showed that the time taken to verify the truth of written propositions varied linearly with the number of nodes traversed in the hierarchy. Participants were fastest to verify propositions like "a canary can sing," which required searching a single node (ie canary), and slower to verify propositions like "a canary has skin," which required

searching three nodes in series (first canary, then bird, then animal). Later studies, however, seriously challenged the model as originally formulated, showing for instance that property- and category-verification times vary systematically with the prototypicality of the item probed—so that participants are faster to decide that a robin (a typical bird) has feathers than that a penguin (an atypical bird) has feathers. Since the nodes in the network were cast as non-compositional primitives, there was no way to represent "typicality" in the original model, and no process that would permit typicality to influence judgment speed. Moreover, the influence of typicality on property decision times was sufficiently strong as to produce results that directly contradicted the Collins and Quillian model. For example, participants were faster to decide that a chicken is an animal than that it is a bird, even though chicken and bird must be closer together in the hierarchy (Rips, Shoben, & Smith, 1973).

These and other challenges led Collins and Loftus (1975) to elaborate the framework. Instead of a search process that begins at the bottom of the hierarchy and moves upward through class-inclusion links, the authors proposed a search mechanism by which the "activation" of a probe concept such as canary would "spread out" along all outgoing links, activating other nodes related to the probe, which in turn could pass activation via their own links. In this spreading activation framework, the strict hierarchical organization of the original model was abandoned, so that direct links could be established between any pair of concepts; and the authors further suggested that links between concept nodes could vary in their "strength," that is, the speed with which the spreading activation process could move from one node to the next. On this account, people are faster to retrieve the properties of typical items because these are more

strongly connected to more general concepts than are less typical items; and the system can rapidly determine that a chicken is an animal by storing a direct link between the corresponding nodes, rather than having to "deduce" that this is true by allowing activation to spread to animal via bird. These elaborations were, however, purchased at the cost of computational simplicity. One appeal of the original model was its specification of a search process in sufficient detail that it could be programmed on a computer. This precision and simplicity depended upon the strict hierarchical organization of concepts proposed by Collins and Quillian (1969). When all nodes can potentially be connected via links of varying strengths, it is not clear how to limit the search process—the spread of activation through the network—so as to retrieve only those properties true of the probe concept. For instance, if the proposition "all bikes have wheels" is stored by linking the nodes for bike and wheel, and the proposition "all wheels are round" is stored by linking wheel and round, how does the network avoid activating the predicate is round when probed with the concept bike?

A second limitation of the spreading-activation theory is that it was not clear how the propositional information encoded in the network should be "linked" to perceptual and motor systems. Spreading-activation theories seem intuitive when they are applied to purely propositional knowledge—that is, when the nodes in the network are understood as corresponding to individual words, and the links to individual predicates, so that the entire system of knowledge may be accurately characterized as a system of propositions. Under such a scheme, there are few questions about which concepts—which nodes and links—should inhabit the network. Very simply, each node and link corresponds to a word in the language, so that the contents of the network are determined by the lexicon,

and the structure of the network represents beliefs that can be explicitly stated by propositions (e.g. “All birds have feathers”). And, such a representational scheme seems most plausible when considering experiments of the kind conducted by Collins and Quillian (1969), where participants must make judgments about the truth of written propositions. When the stimuli to be comprehended are perceptual representations of objects, things get more complicated, because it is less clear which nodes in the network should be “activated” by a given stimulus. A particular dog might belong equally to the classes collie, dog, pet, animal and living thing, so which of these nodes should a visual depiction of the dog activate? More generally, it is unclear in propositional spreading-activation models how the nodes and links of the network relate to or communicate with the sensory and motor systems that provide input to and code output from the semantic system.

2.1 Prototype and similarity-based approaches

Around the same time, there was intensive research focusing directly on the question of how objects are categorized for purposes of naming and induction (see the chapters in the volume on concepts and categorization by Kruschke, on instance/exemplar models by Logan, and on induction and inference by Heit). Throughout the 50’s and 60’s, researchers appear to have assumed that membership in every-day categories could be determined with reference to necessary and sufficient criteria (Bruner, Goodnow, & Austin, 1956). Studies of category learning thus focused on understanding how people come to know which of an item’s properties are necessary and sufficient for membership in some category, and such studies typically employed simple stimuli with well-defined properties organized into artificial categories according to some

rule. For instance, participants might be shown a series of stimuli varying in shape, colour, and size, arbitrarily grouped by the experimenter into categories on the basis of one or more of these dimensions. The participant's goal was to determine the rule governing which items would fall into which categories, and the aim of the research was to determine which strategies participants employed to determine the rule, which kinds of rules were easy or difficult to learn, how easily participants could switch from one rule to another, and so on.

In the early 1970's, Rosch (Rosch & Mervis, 1975; Rosch, Simpson, & Miller, 1976), citing Wittgenstein (1953), observed that most every-day categories are not, in fact, defined by necessary and sufficient criteria; and that, instead, members of categories were best understood as sharing a set of family resemblances. So, for instance, most dogs tend to be hairy, four-legged friendly domesticated animals—even though none of these properties constitutes necessary nor sufficient grounds for concluding that something is a dog. Rosch further showed that the cognitive processes by which we categorize and make inferences about names and properties of objects appear to be influenced by family resemblance relationships (Mervis & Rosch, 1981; Rosch, 1978; Rosch et al., 1976). For instance:

1) Members of a given category can vary considerably in their typicality or representativeness, and members of a given language community show remarkable consistency in their judgments of typicality. For instance, people reliably judge robins to be good examples of the category <u>bird</u>, but judge penguins to be relatively poor examples.

2) Judgments of typicality appear to reflect the attribute structure of the environment. Items judged to be good or typical members have many properties in common with other category members and few distinguishing properties, whereas the reverse is true for items judged to be atypical.

3) Category typicality influences the speed and accuracy with which objects can be named and categorized: As previously mentioned, people are generally faster and more accurate to name and to categorize typical items than atypical items.

From these and other observations, Rosch proposed that semantic/conceptual knowledge about properties of common objects is stored in a set of category prototypes, that is, summary representations of categories that specify the properties most likely to be observed in category members. To retrieve information about a visually presented stimulus, the item is categorized by comparing its observed properties to those of stored category prototypes. The item is assigned to the prototype with the best match, and any properties stored with the matching prototype (including, for instance, its name, as well as other characteristics that may not be directly apparent in the stimulus itself) are then attributed to the object. On this view, category membership depends on similarity to a stored prototype and is therefore graded rather than all-or-nothing. People are faster to recognize typical category members because, by definition, they share more properties with their category prototype, so that the matching process completes more rapidly.

Rosch herself never proposed a computational implementation of prototype theory (Rosch, 1978). Her ideas did, however, spur a considerable volume of research into the computational mechanisms of categorization, which are the topic of another Chapter (see chapter by Krushke, this volume). For current purposes it is sufficient to

note that the similarity-based models deriving from Rosch's approach offer a quite different explanation of human semantic cognition than do spreading-activation theories. Specifically, generalization and induction occur as a consequence of similarity-based activation of stored representations in memory, and not through a process of implicit induction over stored propositions; representations in memory are not linked together in a propositional processing hierarchy or network; and items are treated as "the same kind of thing" when they activate the same prototype or similar sets of instance traces, and not because they connect to the same node in a processing hierarchy.

The two approaches have complementary strengths and weaknesses. Spreading activation models, because they propose that category representations are organized within a processing hierarchy, are economical and provide an explicit mechanism for induction across categories at different levels of specificity. They do not, however, offer much insight into the basis of graded category membership, typicality effects, and so on. Similarity-based theories provide intuitive accounts of such phenomena, but raise questions about the representation of concepts at different levels of specificity. Consider, for instance, the knowledge that both dogs and cats have eyes, DNA, the ability to move, and so on. In spreading-activation theories, such information can be stored just once with the animal representation, and then retrieved for particular individual animals through the spreading activation process. In similarity-based theories, it is not clear where such information resides. If it is stored separately with each category or instance representation, this raises questions of economy and capacity. On the other hand, if separate prototypes are stored for categories at different levels of specificity—one each for animal, bird and penguin, say—it is not clear whether or how these different levels of

representation constrain each other. If, for example, the bird prototype contains the attribute can fly but the penguin representation contains the attribute can not fly, how does the system "know" which attribution to make?

2.2 Challenges for current theories

It may seem that such issues are best resolved through some combination of spreading-activation and similarity-based approaches—and indeed Rosch (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976)and others (Jolicoeur, Gluck, & Kosslyn, 1984) do not seem to view the two frameworks as incompatible. Theoretical developments in semantic cognition have not, however, tended to move in this direction, partly because of serious critical reactions to similarity-based approaches raised in the 1980's that continue to shape research today. Five core issues arising from this criticism are summarized below; the remainder of the article will consider how computational models offer insight as to how to resolve these issues.

Category coherence. Some sets of items seem to form "good" or natural groupings whereas others do not (Murphy & Medin, 1985). For instance, the category dog encompasses a variety of items that seem intuitively to "go together" or to cohere, whereas a category such as grue—things that are currently blue but will turn green after the year 2010—does not. Moreover, the category dog supports induction; if you learn that a particular dog, say Lassie, has a certain kind of protein in her blood, you are likely to conclude that all or most other dogs have the same protein in their blood. Categories like grue do not support induction. What makes some categories, like dog, coherent and useful for induction, and other perfectly well-defined sets of items incoherent and useless? Put differently, how does the semantic system "know" for which groupings of

items it should form a category representation—a prototype or a node in the network—and which not? One possibility is that the system stores a category representation for each word (or at least each noun) in the lexicon; but this solution just pushes the question a step back: why should the language include a word for the concept dog but not for the concept grue? Any theory suggesting that semantic abilities depend upon a mediating categorization process without specifying how the system "knows" which category representations should be created has, in some sense, assumed what it is trying to explain.

Feature selection. Similarity-based models propose that retrieval of semantic information depends upon the degree of similarity between a probe stimulus and a set of stored representations. Any such assessment must specify which probe features or characteristics "count" toward the measure of similarity and how different features are weighted in the determination. As Murphy and Medin (1985) have noted, a zebra and a barber pole might be categorized as "the same kind of thing" if the property has stripes were given sufficient weight.

It is an empirical fact that people do selectively weight different properties in semantic tasks. In the first of many such experiments, Landau, Smith and Jones (1988) showed children a variety of blocks varying in shape, size and texture. After labeling one of the blocks by pointing at it and saying "see this, this is a dax," the authors asked children if they could find another "dax." Children could have used any of the salient features to generalize the new word; but the majority of children selected another object of a similar shape, largely ignoring its size, and texture. Thus the authors proposed that children are subject to a "shape bias" when learning new words—that is, they assume that

the word encompasses items with similar shapes, or equivalently, they weight shape heavily when constructing a representation of the word's meaning.

Moreover, although some properties are undoubtedly inherently more salient than others, this cannot be the sole explanation of such biases, because people will selectively weight the very same properties differently for items in different conceptual domains. For instance, Jones, Smith and Landau (June 1991) have shown that the shape-bias can be attenuated simply by sticking a pair of eyes on the various blocks. Specifically, children were much more likely to use common texture as a basis for generalizing a new name when the blocks had eyes than when they did not—suggesting that they believe texture to be more "important" for categorizing animals (most of which have eyes) than non-animals. Such domain-specific attribute weighting poses an interesting puzzle for similarity-based models: one cannot compute similarity to stored representations, and thus can't categorize, without knowing how different attributes should be weighted; but one cannot know which weightings to use until the item has been categorized, since different weightings are used for different kinds of things (Gelman & Williams, 1998).

Context sensitivity: Of the many things one knows about a common object such as a piano, only a small subset is ever important or relevant in any given situation. For instance, if you have arrived at a friend's house to help her move, the most important fact about the piano is that it is heavy; if however you have come to audition for a band, the most important fact is that it makes music. That is, the semantic information that "comes to mind" in any given situation depends upon the context. Meanings of words are also sensitive to both linguistic context and to real-world context—for instance, the referent of the phrase "Check out my hog" may be completely different depending on whether one is

speaking to a farmer or a biker. Contextual influences on semantic task performance have been robustly documented in a very wide variety of tasks (Yeh & Barsalou, 2006), yet the implications of such context-sensitivity seem not to have penetrated many models of semantics (Medin & Shaffer, 1978). Both spreading-activation models and prototype theories specify how individual concepts may be represented and activated, but an implicit assumption of such models is that contextual information is effectively discarded—neither approach specifies, for instance, what would differ in the retrieval process when one moves a piano as opposed to playing it. The default assumption seems to be that the very same representation (node or prototype) would be activated in both cases, and it is not clear how different information would come to the fore in the two situations.

Abstract concepts: What are the "properties" of concepts like justice, or alive, or beautiful, that would allow one to construct prototypes of these categories, or to connect them together with simple predicates in a spreading-activation network? Such questions may seem beyond the grasp of contemporary theories of semantic memory, which predominantly focus on knowledge about concrete objects with directly-observable characteristics; but in fact the same questions are pressing even for such theories. The reason is that the properties often invoked as being critical for representing concrete concepts are frequently quite abstract in and of themselves. Consider, for instance, the properties important for the concept animal, which might include self-initiated movement and action-at-a-distance (Mandler, 2000), contingent movement (Johnson, 2000), goal-directedness (Csibra, Gergely, Biro, Koos, & Brockbank, 1999), and "biological" patterns of motion (Bertenthal, 1993), among other things. It is difficult to see how these might be

directly available through perceptual mechanisms. For instance, different instances of self-initiated movement may be perceptually quite different—birds flap and glide, rabbits hop, snakes slither, people walk, and so on. To recognize that these different patterns of motion all have something important in common—"self-initiatedness," say—is to synthesize from them what is effectively an abstract feature. Even relatively concrete properties, such as having legs or a face, seem less and less concrete the more one considers the range of variability in the actual appearance of the legs or faces on, say, birds, dogs, fish and insects. So a relatively concrete concept such as animal depends upon the specification of properties that can be relatively abstract. Similarly, the most important properties for many manmade objects are often functions, which are also difficult or impossible to define with reference to purely perceptual characteristics. A hammer and a screwdriver, for example, have similar functions—they are used to fasten things together—and for this reason may be considered similar kinds of things, despite having quite different shapes and demanding quite different kinds of praxis. In general, theories of semantic memory must explain how people become sensitive to such abstract regularities, and are able to use them to constrain property generalization. The suggestion that such regularities are directly apparent in the environment is not transparently true for many such properties.

Representing multiple objects, relationships, and events. Finally, it should be clear that both spreading-activation and similarity-based theories are targeted predominantly at explaining knowledge about individual concepts, corresponding roughly to the meanings of single words. But semantic abilities extend considerably beyond knowledge about the meanings of individual words and objects: it encompasses knowledge about events and

situations (e.g. how to order in a restaurant) as well as knowledge about various relationships between and within individual objects, including associative relationships (e.g. hammers are used with nails), and causal relationships among object properties (e.g. having hollow bones causes a bird to be light) and between objects (e.g. having a certain scent causes the flower to attract bees). In many cases, the single object's meaning seems to rely on its relationships to these other objects—for instance, it makes no sense to conceive of the hammer as a "decontextualized pounder" (Wilson & Keil, 2000); rather the hammer's meaning depends partly on the fact that it is used specifically to pound nails, usually with the intent of attaching two separate objects. Without some account of how multiple objects and their relationships to one another combine to form representations of events and scenes, it is difficult to understand how such knowledge arises even for the meanings of single words and objects.

Summary

In summary, two different computational frameworks informed research in semantic memory throughout the 1970's and 80's: spreading activation models and prototype (and other similarity-based) theories. These two frameworks still form the theoretical background to much empirical research in cognitive psychology, and are the most likely to be covered in cognitive psychology textbooks. And, both approaches continue to foster ongoing research, especially in the domains of categorization (Smith, 2002; Zaki & Nosofsky, 2004) and in artificial intelligence (Crestani, 1997) and aspects of lexical processing and speech production (Bodner & Masson, 2003; Dell, 1986). These frameworks raise challenging questions, however, about the computational basis of human semantic abilities, specifically:

1. Why do some sets of items form more “coherent” categories than others, and how does the semantic system “know” which category representations to form?
2. Why are some properties more “important” for governing semantic generalization and induction than others, and how does the system “know” which properties are important for which concepts?
3. How is context represented, and how does it work to constrain which information “comes to mind” in a given situation?
4. How are abstract concepts and properties acquired?
5. How can the system combine multiple concepts together to represent events, scenes, and relationships among objects?

Some of these questions have been addressed, with varying degrees of success, by computational modeling efforts that fall outside the scope of this article, since they mostly pertain to domains addressed by other chapters in the Handbook. Much of this work is specifically focused on understanding categorization phenomena. Anderson’s Rational model of categorization (Anderson, 1991) provides an explanation of how a categorization-based semantic system can “decide” which category representations should be created in memory. Models proposed by Kruschke (1992) and (Nosofsky, 1986) provide hypotheses about how certain feature dimensions are selectively weighted when making categorization judgments; and the “context” models of categorization (Medin & Shaffer, 1978; Nosofsky, 1984) provide some suggestions as to how different kinds of information about a given concept may be retrieved in different situations or contexts. As previously noted, computational models focused on these questions are discussed at length in other Handbook chapters—specifically, the chapter on

categorization and concepts by John Krushke, the chapter on induction and inference by Evan Heit, the chapter on instance/exemplar models by Gordon Logan, and the chapter on Bayesian approaches to cognition by Tom Griffiths, Charles Kemp, and Josh Tenenbaum. The work discussed in sections 3-5 below will follow three threads of research in semantic cognition that derive from the Parallel Distributed Processing approach to cognition.

3. Distributed semantic models.

3.1 Hinton's (1981) distributed model.

The first important thread begins with Hinton's (1981) proposal for storing propositional knowledge (of the kind described in a Quillian-like semantic network) in a Parallel Distributed Processing (PDP) network. As in most information-processing frameworks, PDP models typically have "Inputs" that respond to direct stimulation from the environment, and "Outputs" that correspond to potential actions or behaviors. In such models, information is represented as a pattern of activation across a pool of simple, neuron-like processing units; and information processing involves the flow of activation within and between such pools by means of weighted, synapse-like connections. The activation of any given unit depends upon the sum of the activations of the sending units from which it receives inputs, multiplied by the value of the intermediating weights. This net input is then transformed to an activation value according to some transfer function (often a logistic function bounded at 0 and 1). A network's ability to complete some input-output mapping depends upon the values of the intermediating weights; in this sense, a network's "knowledge" is often said to be "stored" in the weights. To store new information in a network, it is not necessary to add new architectural elements; instead,

the weights in the existing network must be adjusted to accommodate the new information. So, learning in a connectionist framework does not involve the addition of new data structures, prototypes, or propositions, but instead involves the adjustment of connection weights to promote some new mapping between input and output (see the chapter on connectionist approaches to cognition by Michael Thomas and Jay McClelland, this volume).

--Figure 1 about here--

Hinton (1981) was interested to show how a body of propositional information might be stored in such a network, without any explicit proposition-like data structures in the system. The architecture of his model, shown in Figure 1, reflects the structure of a simple proposition of the form Item-Relation-Attribute; there is a single bank of neuron-like processing units for each part of the proposition. Different fillers for each slot are represented as different patterns of activation across the corresponding pool of units. For example, the representation of the proposition Clyde is gray would correspond to one pattern of activity across each of the three groups of units: one for Clyde, one for is, and one for gray.

All three banks are send and receive weighted connections to a fourth layer (labelled Prop in the illustration). When a pattern of activation is applied across the three input layers, Prop units compute their inputs as the sum of the activations across input units weighted by the magnitude of the interconnecting weights. Each input thus produces a pattern of activity across the Prop units, which in turn send new signals back to the Item, Relation and Attribute units, which update their states accordingly in reaction to the new inputs. The process iterates until the unit states stop changing, at which point

the network is said to have settled into a steady state. Hinton demonstrated that individual propositions could be stored in the network, by adjusting the interconnecting weights to make the patterns representing the proposition stable. To achieve this, Hinton trained the model with a variant of the delta-rule learning algorithm, which is explained in detail in the chapter by Thomas and McClelland (this volume). After training, each stored proposition would be represented in the network by a unique pattern of activity across the Prop units, which simultaneously activated and received support from the input patterns.

This early model had several interesting properties and implications. First, it was capable of completing stored propositions when given two of its terms as inputs. For example, when provided with the inputs Clyde and is, the network settled into a steady state in which the pattern representing the correct completion of the proposition (gray) was observed across the Attribute units. Second, several such propositions could be stored in the network, in the same finite set of weights. Thus, in contrast to spreading-activation and similarity-based models, new information could be stored in memory without adding representational elements to the system. Third, when appropriate representations were chosen, the network provided a natural mechanism for generalization. If related objects (such as various individual elephants) were represented by overlapping patterns of activity across the Item units, they would contribute similar inputs to the Prop units. Thus, the entire network would tend to settle into an appropriate steady state (corresponding to the most similar stored proposition) when given a novel input that overlapped with familiar, stored patterns. For example, if the network had stored the proposition Clyde is gray, and was then given the inputs Elmer is in the Item and Relation units, it would settle to a state in which the pattern corresponding to gray

was observed across Attribute units—provided that the representations of Clyde and Elmer were sufficiently similar. Thus, the model exhibited the two characteristics most fundamental to both spreading-activation and prototype theories: an economical means of storing information, and a mechanism for generalizing stored information to new stimuli.

The model also offered some leverage on two of the questions posed above from our consideration of prototype and spreading-activation theories. Specifically, the first question—how does the system "know" for which categories it should create representations—becomes moot in this framework. There are no discrete category representations in Hinton's model. Individual items—which in spreading activation theories would correspond to individual nodes, and in prototype theories to individual category prototypes—are represented as distributed patterns of activity across the same set of processing units. The same is true of different predicates, different attributes, and full propositions—all are represented as distributed patterns across processing elements. Generalization is governed, not by a categorization process nor by the search of an explicit processing hierarchy, but by the similarities captured by these various distributed representations. This scheme does not address the important question of category coherence—why some sets of items form good categories that support induction whereas others do not—but it no longer requires an answer to the question of which categories are stored in memory and which not.

Second, Hinton pointed out that, when many propositions are stored in the network, neither the Item nor the Relation inputs alone are sufficient to uniquely determine a correct pattern of activation in the Attribute units. For instance, suppose the

model has stored the following propositions about Clyde the Elephant and Frank the Flamingo:

1. Clyde is gray
2. Frank is pink
3. Clyde has a trunk
4. Frank has a beak

Here the output generated by a given item (Clyde or Frank) depends upon the particular relation, is or has, with which it occurs. Similarly, the response generate for a given relation depends upon which item is being probed. Both the Item and Relation representations provide constraints on the ultimate interpretation of the inputs into which the network settles (the Prop representation), and jointly these determine the completion of the proposition. Put differently, the model generates different internal representations and hence different outputs for the very same item depending on the context in which the item is encountered. This early model thus provides some tools for understanding influences of context on semantic representation and processing.

Hinton's model also raised many questions, of course. Most obviously, the model's capacity to learn without interference and to generalize appropriately depends entirely on the particular patterns of activity chosen to represent various items, relations, and attributes. Hinton simply hand-selected certain patterns to illustrate the appeal of the basic framework. How are appropriate internal representations acquired under this framework?

3.2 The Rumelhart model

This question was explicitly addressed by Rumelhart (Rumelhart, 1990; Rumelhart & Todd, 1993), who showed how the same propositional content stored in the Collins and Quillian (1969) hierarchical model can be learned by a simple connectionist network trained with backpropagation (see the chapter on connectionist models for a detailed explanation of the backpropagation learning algorithm). An adaptation of Rumelhart's model is shown in Figure 2; it can be viewed as a feed-forward instantiation of a model similar to Hinton's, in which the network is provided with the Item and Relation terms of a simple proposition as input, and must generated all appropriate completions of the proposition as output.

--Figure 2 about here--

The model consists of a series of nonlinear processing units, organized into layers, and connected in a feed-forward manner as shown in the illustration. 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, 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, that is:

$$net_j = \sum_i a_i w_{ij}$$

…where net_j is the net input of the receiving unit *j*, *i* indexes units sending connections to *j*, *a* indicates activation of each sending unit, and w_{ij} indicates the value of the weight projecting from sending unit *i* to receiving unit *j*. The net input is then transformed to an activation *a* according to the logistic function, which bounds activation at 0 and 1:

$$a = \frac{1}{1 + e^{-net}}$$

To find an appropriate set of weights, the model is trained with the backpropagation learning algorithm (Rumelhart, Hinton, & Williams, 1986). First, an Item and Relation are presented to the network by setting the activations of the corresponding input units to 1 and all other inputs to 0, and activation is propagated forward to the output units, with each unit computing its net input and activation according to the equations above. The observed output states are then compared to the desired or target values, and the difference is converted to a measure of error. In this case, the error is the sum over output units of the squared difference between the actual output activations and the target values:

$$err_p = \sum_i (a_{p_i} - t_{p_i})^2$$

…where err_p indicates the total error for a given pattern *p*, *i* indexes each output unit, *a* indicates the activation of each output unit given the input pattern for *p*, and *t* indicates the target value for each output unit for pattern *p*. The partial derivative of this error with respect to each weight in the network is computed in a backward pass, and each weight is adjusted by a small amount to reduce the error (see the chapter on connectionist models by Michael Thomas and Jay McClelland for further information on the backpropagation learning rule).

Although the model's inputs are localist, each individual Item unit projects to all of the units in the layer labelled 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 input and target states 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 the case of the Rumelhart network, for reasons elaborated below, the learned representations turn out to capture the semantic similarity relations that exist among the items in the network's training environment. These learned similarity relations provided a basis for generalization and property inheritance, just as did the assigned similarities in Hinton's (1981) model. For instance, after the model had learned about the 8 items shown in Figure 2, the authors could teach the model a single fact about a new item—say, that a sparrow is a kind of bird—and then could query the model about other properties of the sparrow.[1] In order to learn that the new item (the sparrow) is a kind of bird, the model must represent it with a pattern of activation similar to the previously-learned robin and canary, since these are the only items to which the label “bird” applies. Consequently, the model tends to attribute to the sparrow other properties common to both the robin and the canary: it “infers” that the sparrow can move and fly but can not swim; has feathers, wings and skin but not roots or gills; and so on. That is, the key function of semantic memory that, in Hinton’s (1981) model, was achieved by hand-crafted representations—generalization of previously-learned information to new items—was accomplished in Rumelhart’s model by internal representations that were “discovered” by the backpropagation learning rule.

3.3 Feature weighting and category coherence.

Rogers and McClelland (2004) have suggested that Rumelhart's model provides a simple theoretical framework for explaining many of the important phenomena motivating current research in semantic cognition. On this construal, the two input layers of the model represent a perceived object and a context provided by other information available together with the perceived object. For instance, the situation may be one in which a young child is looking at a robin on a branch of a tree, and, as a cat approaches, sees it suddenly fly 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. That is, an object and a situation afford the basis for implicit predictions (which may initially be null or weak), and observed events then provide the basis for adjusting the connection weights underlying these predictions, thereby allowing the experience to drive change in both underlying representations and predictions of observable outcomes. The range of contexts in which the child might encounter an object may vary widely: the child may observe the object and what others are doing with it (pick it up, eat it, use it to sweep the floor, etc); some encounters may involve watching what an object does in different situations; others may involve naming and other kinds of linguistic interactions. Semantic/conceptual abilities arise from the learning that occurs across many such situations, as the system comes to make increasingly accurate predictions about the consequences of observing different kinds of items in different situations and contexts.

The Rumelhart model provides a simplified implementation of this view of semantic abilities: The presentation of an "object" corresponds to the activation of the

appropriate pattern of activity over the input units in the Rumelhart model; the context can be represented via the activation of an appropriate pattern over 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.

--Figure 3 about here--

The authors suggested that this framework is appealing partly because it provides answers to some of the puzzling questions about the acquisition of semantic knowledge discussed previously. To show this, Rogers and McClelland (2004) trained a variant of the model shown in Figure 2, and investigated its behavior at several different points during the learning process.

The first important observation was that model's internal representations underwent a "coarse-to-fine" process of differentiation, such that items from broadly different semantic domains (the plants and animals) were differentiated earliest in learning; whereas closely related items (e.g. the rose and daisy) were differentiated latest. Figure 3 shows a multidimensional scaling of the internal representations generated by the model across the Representation layer for all 8 items at 10 different points during training. The lines trace the trajectory of each item throughout learning in the 2-dimensional compression of the representation state space. The labelled end-points represent the final learned internal representations after 1500 epochs of training. These end-points recapitulate the semantic similarity relations among the 8 items: the robin and canary are quite similar, for instance, and both are more similar to the two fish than they are to the 4 plants. The lines tracing the developmental trajectory leading to these end-

points show that the 8 items, initially bunched together in the middle of the space, soon divide into two clusters (plant or animal) based on animacy. Within these clusters, there is little differentiation of items. Next, the global categories split into smaller intermediate clusters (e.g. birds and fish) with little differentiation of the individual items within each cluster, and finally the individual items are pulled apart. In short, the network's representations appear to differentiate in relatively discrete stages, first completing differentiation of at the most general level before progressing to successively more fine-grained levels of differentiation.

The basis for this nonlinear, stage-like process of coarse-to-fine differentiation in the model proved key to explaining several critical phenomena in the study of human semantic abilities. To see why the model behaves in this fashion, 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 meaningless pattern of activity throughout the network. Since oaks and pines share many output properties, this pattern results in a similar error signal for the two items, and the weights leaving the oak and pine units 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 it tends to move in a similar direction as the other plants. Similarly, the rose tends to be pushed in the same direction as all of the other plants, and the other animals 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 in the early stages of learning in the model, 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?

To understand the stage-like pattern of differentiation, 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 the same representation. When this is so, if the weights going forward from the representation layer “work” to capture these shared properties for one of the animals, they must simultaneously work to capture them for all of the others. Similarly, any weight change that is made to capture the shared properties for one of the items will produce the same benefit in capturing these properties for all of the other items: If the representations of all of the items are the same, then changes applied to the forward-projecting weights for one of the items will affect all of the others items equally, and so the changes made when processing each individual item will tend to accumulate with those made in processing the others. On the other hand, weight changes made to capture a property of an item that is not shared by others with the same representation will tend to be detrimental for the other items, and when these other items are processed the changes will actually be reversed. For example, two of the animals (canary and robin) can fly but not swim, and the other two (the salmon and the

sunfish) can swim but not fly. If the four animals all have the same representation, what is right for half of the animals is wrong for the other half, and the weight changes across different patterns will tend to cancel each other out. 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 representational similarity structure at a given point in time influences the speed with which various different kinds of attributes are learned in the model, in the weights projecting forward from the Representation layer. But what about the weights from the input units to the representation layer? These determine the representational similarity structure between items in the first place. As previously stated, items with similar outputs will have their representations pushed in the same direction, while 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 answer 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 particular configuration of weight projecting forward from the Representation layer. For instance, if the network activated “has wings” and “has scales” to an equal degree for all animals (since half the animals have wings and the other half have scales) then there is no way of adjusting the representation of, say, the canary that will simultaneously reduce error on both the “has wings” and “has scales” units. Consequently, these properties will not exert much

influence on the weights projecting into the Representation layer, and will not affect how the representation of canary changes. In other words, error propagates much more strongly from properties that the network has begun to master.

--Figure 4 about here--

Rogers and McClelland (2004) illustrated this phenomenon by observing the derivative of the error signal propagated back to the Representation units for the canary item. Specifically, this derivative was calculated across 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). Since weights projecting into the Representation units are adjusted in proportion to these error derivatives, the calculation indicates to what extent these three different kinds of features are influencing representational change at different points in time. Figure 4 shows how the error derivatives from these three kinds of properties change throughout training when the model is given the canary (middle plot). This is graphed 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 sing, fly and move (top plot). The Figure shows that there comes a point at which the network is beginning to differentiate the plants and the animals, and is beginning to activate move correctly for all of the animals. At this time properties like can move (reliably differentiating plants from animals) are producing a much stronger error derivative at the Representation units than are properties like can fly or can sing. As a consequence, these properties are contributing much more strongly to

changing the representation weights than are the properties that reliably differentiate birds from fish, or the canary from the robin. Put differently, the knowledge that the canary can move is more "important" for determining how it should be represented than the information that it can fly and sing, at this stage of learning. (The error signal for move eventually dies out as the correct activation reaches asymptote, since there is no longer any error signal to propagate once the model has learned to produce the correct activation).

The overall situation can be summarized as follows. Initially, the network assigns virtually the same representation to all items, and the only properties that vary systematically with these representations are those that are shared by all items (e.g. can grow, is living). All other properties have their influence on the weights almost completely cancelled out, since changes that favor one item will hinder another. Since there are many properties common to the animals and not shared by plants (and vice versa), however, weak error signals from these properties begin to move the various animal representations away from the plant representations. When this happens, the shared animal representation can begin to drive learning (in the forward weights) for properties that the animals have in common; and the shared plant representation can begin to drive learning for properties common to plants. These properties thus begin to exert a much stronger influence on the network's internal representations than do, for instance, the properties that differentiate birds from fish. The result is that the individual animal representations remain similar to one another, but are rapidly propelled away from the individual plant representations. Gradually the weak error signals propagated from the properties that discriminate more fine-grained categories begin to accumulate, causing

these subgroups to differentiate slightly, and providing the basis for another "wave" of differentiation. This process eventually propagates down to the subordinate level, where individual items are differentiated from one another.

The network's tendency to differentiate its internal representations in this way does not arise from some general bias toward discovering superordinate category structure per se. Instead it comes from patterns of higher-order covariation exhibited amongst the output properties themselves. The first wave of differentiation in the model will distinguish those subgroups whose shared properties show the strongest tendency to consistently covary together across the corpus (corresponding to those with the highest eigenvalues in the property covariance matrix; see Rogers and McClelland, 2004, Chapter 3 for further detail)—that is, properties that show the strongest tendency to covary coherently. In the model corpus, and perhaps in real experience, such subgroups will correspond to very general semantic domains. For instance, animals share many properties—self-initiated and biological movement, biological contours and textures, facial features, and so on—that are not observed in plants or manmade objects. The system will not be pressured, however, to differentiate superordinate groups that do not have cohesive structure (e.g. toys versus tools). Further waves of differentiation will then distinguish groupings whose shared properties show the next strongest patterns of coherent covariation.

It is worth noting that these interesting phenomena depend upon three aspects of the network architecture. First, semantic representations for all different kinds of objects must be processed through the same weights and units at some point in the network, so that learning about one item influences representations for all items. This convergence in

the architecture forces the network to find weights that work for all items in its experience, which in turn promotes sensitivity to high-order covariation amongst item properties. Second, the network must begin with very similar representations for all items, so that learning generalizes across all items until they are differentiated from one another. Third, learning must be slow and interleaved, so that new learning does not destroy traces of previous learning. These architectural elements are critical to the theory and are taken as important design constraints on the actual cortical semantic system (see the last section of this chapter, and the chapter on episodic memory by Ken Norman, Greg Detre and Sean Polyn, this volume). It is also worth noting that the effects do not depend upon the use of the backpropagation learning algorithm per se. Any learning algorithm that serves the function of reducing error at the output (e.g. contrastive Hebbian learning, GeneRec, leabra, etc…) could potentially yield similar results—so long as they permit new learning to generalize relatively broadly. For instance, learning algorithms that promote representational sparsity (e.g. some parameterizations of leabra) will diminish the degree to which learning generalizes across different items, and so may not show the same sensitivity to higher-order covariation.

Rogers and McClelland's (2004) analysis of learning in the Rumelhart model provides a basis for understanding two of the pressing questions summarized earlier:

i) Category coherence. Why do some groupings of items seem to form "good" categories that support induction whereas others do not? The model suggests that "good" categories consist of items that share sets of properties that vary coherently together across many situations and contexts. Because these properties strongly influence representational change early in learning, they strongly constrain the degree to which

different items are represented as similar/dissimilar to one another, which in turn constrains how newly-learned information will generalize from one item to another. Rogers and McClelland (2004) showed how this property of the model can address phenomena as diverse as the progressive differentiation of semantic representations in infancy (Mandler, 2000), basic-level advantages in word-learning in later childhood (Mervis, 1987), "illusory correlations" in induction tasks (Gelman, 1990), and sensitivity to higher-order covariation in category-learning experiments (Billman & Knutson, 1996).

ii) Selective feature weighting. Why are certain properties "important" for representing some categories and not others? The PDP account suggests that a given property becomes "important" for a given category when it covaries coherently with many other properties. This "importance" is reflected in two aspects of the system's behavior. First, coherently-covarying properties are the main force organizing the system's internal representations—so that items with a few such properties in common are represented as similar even if they have many incoherent properties that differ. Second, coherent properties are learned much more rapidly: because items that share such properties are represented as similar, learning for one item tends to generalize well to all other items that share the property. In simulation experiments, Rogers and McClelland showed that this emergent "feature weighting" provided a natural account of several phenomena sometimes thought to require innate knowledge structures. These include sensitivity to "conceptual" over perceptual similarity structure in infancy (Pauen, 2002), domain-specific patterns of feature-weighting (Keil, 1989; Macario, 1991), and the strong weighting of "causal" properties in determine conceptual similarity relations (Ahn, 1998; Gopnik & Sobel, 2000).

3.4 Context sensitivity

It is worth touching on one further aspect of the Rumelhart model because it relates to issues central to the next two sections. The analyses summarized above pertain to the item representations that arise across the Representation units in the Rumelhart model. These units receive input from the localist input units corresponding to individual items, but they do not receive input from the Context input units. Instead the distributed item representations feed forward to the Hidden units, which also receive inputs from the Context inputs, and then pass activation forward to the output units. The pattern of activation arising across Hidden units may thus be viewed as a learned internal representation of an item occurring in a particular context. That is, in addition to learning context-independent representations (across the Representation units), the Rumelhart network also learns how these representations should be adapted to suit the particular context in which the item is encountered. These context-sensitive representations allow the network to produce different outputs in response to the same item—a key aspect of semantic cognition discussed in the introduction.

It turns out that this context-sensitivity also explains a puzzling aspect of human cognition—the tendency to generalize different kinds of newly-learned information in different ways. For instance, Carey (1985) showed that older children inductively generalize biological facts (such as “eats” or “breathes”) to a much broader range of living things than they do psychological facts (“thinks,” “feels”). Because the Rumelhart model suggests that the same items get represented differently in different contexts, it provides a way of understanding why different “kinds” of properties might generalize in different ways.

--Figure 5 about here--

Rogers and McClelland (2004) trained a variant of the Rumelhart model with a corpus of 16 items from the same 4 categories as the original (birds, fish, trees and flowers), and examined the patterns of activation that arose across the Representation and Hidden units for these 16 items in different contexts. Figure 5 shows a multidimensional scaling of these patterns. The middle plot shows the learned similarities between item representations in the context-independent layer; the top plot shows the similarities across Hidden units for the same items in the is context; and 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. By contrast, in the is context, there are few properties shared among objects of the same kind, so that the network is pressured to strongly differentiate items in this context. The context-weighted similarities illustrated in the Figure determine how newly-learned properties will generalize in different contexts. If the network is taught, for instance, that the maple tree "can queem" (where "queem" is some novel property), this fact will tend to generalize strongly to all of the plants, since these are represented as very similar in the "can" context. If it is taught that the maple tree "is queem," the new fact will not generalize strongly to all plants, but will weakly generalize to other items that, in the "is" context, are somewhat similar to the maple. In short, because the model's internal representations are sensitive to contextual constraints, the "base" representations learned in the context-independent Representation can be reconfigured to capture similarity

relationships better suited to a given context. This reshaping can then influence how newly-learned information will generalize.

Summary

In summary, this thread of research offers a promising theoretical framework for semantic cognition that addresses some of the core issues discussed in the introduction. The framework suggests that the semantic system allows us, when presented with a perceptual or linguistic stimulus in some particular situation, to make context-appropriate inferences about properties of the item denoted by the stimulus. It suggests that these inferences are supported by distributed internal representations that capture semantic similarity relations; and that these relations can be adapted to suit particular contexts. It further suggests that the internal representations are learned through experience, and shows how the learning dynamics that arise within the framework provide an explanation of category coherence, feature selection, and context-sensitivity in semantics. The framework does not explicitly address other key challenges for a theory of semantics—specifically, the representation of abstract concepts, events, and multiple objects and relationships. These are the main focus of the next section.

4. Temporal structure, events, and abstract concepts

4.1 Simple recurrent networks

The second important thread of research derives in part from the seminal work of Elman (1990). Elman was interested, not only in semantics, but in several different aspects of language, including the ability to segment the auditory stream into words, to organize words into different syntactic classes, and to use information about word order to constrain the interpretation of sentences. The key insight of this work was that all of

these different abilities may derive from a similar underlying learning and processing mechanism—one that is sensitive to statistical structure existing in events that unfold over time. The catalyst for this insight was the invention of neural network architecture that permitted sensitivity to temporal structure—the "simple recurrent network" (SRN) or "Elman net" shown in Figure 6.

--Figure 6 about here--

The three leftmost layers of the SRN shown in the Figure constitute a feed-forward connectionist network similar to that used by Rumelhart: units in the input layer are set directly by the environment; activation feeds forward through weighted connections to the Hidden layer, and from there to the output layer. What makes the model "recurrent" is the Context layer shown on the right of the Figure. Activation of these units feeds forward through weighted connections to influence the Hidden units, just as do the input units. The activations of the Context units are not set by inputs from the environment however. Instead, they contain a direct copy of the activation of the Hidden units from the previous time-step. It is this "memory" of the previous hidden-unit state that allows the network to detect and respond to temporal structure.

As a simple example, suppose that the network's inputs code the perception of a spoken phoneme, and that the network's task is to predict in its outputs what the next phoneme will be. Each individual input and output unit might, for instance, be stipulated to represent a different syllable in English. To process a statement such as "pretty baby," the network would first be presented with the initial syllable (/pre/). Activation would spread forward to the Hidden units and then to the output units through the interconnecting weights. On the next step of the sequence, the Hidden unit pattern

representing /pre/ would be copied to the Context layer, and the network would be given the next syllable in the phrase (/ti/). On this step, the Hidden unit activations will be influence both by this new input, and by the activations of the Context units which contain the trace of the preceding hidden representation. In other words, the new hidden representation will code a representation of /ti/ in the context of having previously seen /pre/. Activation again feeds forward to the outputs, which code the network's "best guess" as to the likely next phoneme. On the third step, the Hidden unit representation is again copied to the Context layer, and the next syllable (/ba/) is presented as input. Again the Hidden representations are influenced both by the input and by the Context unit activations; but this time the Context representation has been influenced by two previous steps (/pre/ followed by /ti/). In other words, the new Hidden representation now codes /ba/ in the context of previously encountering /pre/ followed by /ti/. In this manner, new inputs are successively "folded in" to the Context representation, so that this representation constitutes a distributed internal representation of the sequence up to the present point in time. As a consequence of this "holding on" to previously presented information, the model can produce different outputs for exactly the same input, depending upon previously-occurring inputs. That is, it is sensitive to the temporal context in which a given input is encountered.

SRNs can be trained with backpropagation just like a standard feed-forward network. In the syllable-prediction example, the presentation of each input syllable would provoke a pattern of activation across output units (via hidden units) that can be compared to a target pattern to generate a measure of error. Since the task is prediction in this example, the target is simply the next-occurring syllable in the speech stream.

Weights throughout the network can then be adjusted to reduce the error. These weight adjustments are typically applied to all forward-going weights, including those projecting from the Context to Hidden layers. Learning on the weights projecting from Context to Hidden layers allows the network to adjust exactly how the sequence history coded in the Context influences the Hidden representation on the current time-step; and this in turn influences which steps of the sequence are robustly preserved in the Context representation itself. If there is no temporal structure, so that the sequence history has no implication for how a current input is processed, the weights from Context->Hidden will never grow large, and the Hidden representation will be driven almost exclusively by the Input and not by the Context. As a consequence, the Context representation itself will only reflect the representation of the preceding item, and will not "build up" a representation of the sequence preceding that item. On the other hand, if there is temporal structure—so that predictions derived from a given input can be improved by "taking into account" the preceding items in the sequence, then the weights projecting from Context to Hidden units will be structured by the learning algorithm to captialize on these relationships, so that the Hidden states come to be more strongly influenced by the Context, and preceding states get "folded in" to the new context representation.

The SRN turned out to be a valuable tool for understanding a variety of linguistic phenomena precisely because language has temporal structure at many different timescales. At a relatively small timescale, for instance, it is the case that syllable-to-syllable transitions that occur within words tend to be much more predictable than the transitions that occur between words. From the earlier example, the transition from /pre/->/ti/ is much more frequent in English than the transition /ti/-> /ba/ (Saffran, Aslin, &

Newport, 1996). Because this is true, a syllable-prediction network like the one sketched out above can provide a strategy for detecting word-boundaries in a continuous speech stream: simply place the boundaries wherever prediction error is high. At broader timescales, SRNs provide a way of thinking about processing of syntactic information in languages like English where such information is often carried by word order. And, it turns out, SRNs and related approaches offer important insights into the acquisition and representation of semantic information.

Here again, the critical insight was offered by Elman (1990), who trained an SRN in which the input and output units, instead of corresponding to individual syllables, instead represented individual words. The network's task, just as before, was prediction—in this case, prediction of the next word in a sentence, given the current word as input. Elman trained the model with a sequence of simple 2- and 3-word sentences (e.g. "Woman smashes plate," "Cat moves," and so on) presented to the network in a long series. He then examined the internal representations arising across Hidden units in response to the activation of each individual word. The interesting observation was that words with similar meanings tended to be represented with similar patterns of activation across these units—even though input and outputs in the model were all localist, so that there was no pattern overlap between different words in either the input or output. Somehow the network had acquired information about semantic relatedness solely by trying to predict what word would come next in a sentence!

Why should this be? The answer is that words with similar meanings, precisely <u>because</u> they have similar meanings, tend to occur in similar linguistic contexts. For instance, because dogs and cats are both kinds of pet, we tend to use similar words when

referring to them in speech: we say things like, "I have to feed the dog/cat," "Don't worry, the dog/cat doesn't bite," "Please let the dog/cat outside," and so on. Elman's simulations suggested that the more similar two words are in meaning, the more similar are the range of linguistic contexts in which they are encountered. Because the representation of a given item is, in the SRN, influenced by the temporal contexts in which it is encountered, then items that occur in similar contexts tend to receive similar representations. Just as the Rumelhart network learns to represent items as similar when they overlap in their output properties, so the SRN learns to represent items as similar when they overlap in the distribution of items that precede and follow them. Because items with similar meanings tend to be preceded and followed by similar distributions of words in speech, this suggests that the acquisition of semantic similarity relations may be at least partially supported by a learning mechanism that is sensitive to the context in which the words occur.

There are three aspects of this research that offer leverage on the theoretical issues listed previously. First, the internal representations acquired by an SRN are, like the representations that arise across the Hidden layer of the Rumelhart network, context-sensitive. In both cases, the distributed patterns that promote the correct output capture, not just the current input, but the current input encountered in some context. As a consequence, both kinds of network can produce different responses to the same item, depending on the context. In an SRN, the context needn't be represented as a separate input from the environment (as it is in the Rumelhart network), but can consist solely of a learned internal representation of the sequence of previously-encountered inputs.

Second, Elman's approach suggests one way of thinking about representation of meanings for abstract concepts. Because semantic similarity relations are apparent (at least to some degree) from overlap in the linguistic contexts in which words tend to appear in meaningful speech, then such relations might be derived even for words with abstract meanings. Words like "fair" and "just" may not be associated with obvious perceptual-motor attributes in the environment, but they likely occur within similar linguistic contexts ("The decision was just," "The decision was fair"). The insight that word-meanings may partially inhere in the set of contexts in which the word is encountered may therefore provide some explanation as to how learning of such meanings is possible.

Third, the representations arising in the Context layer of an SRN capture information, not just about a single input, but about a series of inputs encountered over time. That is, these representations are inherently representations of whole events rather than individual items. The SRN thus offers a tool for understanding how the semantic system might construct internal representations that capture the meaning of a whole event, instead of just the meaning of a single object or word.

The remainder of this section discusses some of the implications of these ideas for theories of semantic memory as they have been cashed out in two influential modelling approaches: Latent Semantic Analysis (Landauer & Dumais, 1997) and related approaches (Burgess & Lund, 1997; Steyvers et al., 2006), and the "Sentence Gestalt" models described by St. John and McClelland (McClelland, St. John, & Taraban, 1989; St. John & McClelland, 1990; St. John, 1992).

4.2 Latent Semantic Analysis

Latent Semantic Analysis (LSA) is an approach to understanding the semantic representation of words (and larger samples of text) that capitalizes on the previously-mentioned observation that words with similar meanings tend to occur in similar linguistic contexts (Landauer, Foltz, & Laham, 1998). Elman had illustrated the face validity of the idea by training an SRN with a small corpus of sentences constructed from a limited set of words. The pioneers of LSA and related approaches established the power of the idea by investigating precisely how much information about semantic relatedness among words can be extracted from linguistic context in large corpora of written text.

The basic computations behind LSA are fairly straightforward. The process begins with a large set of samples of text, such as an encyclopaedia in which each article is considered a separate sample in the set. From this set, a matrix is constructed. Each row of the matrix corresponds to a single word appearing at least once in the set, and each column corresponds to one of the text samples in the set. The elements of the matrix indicate the frequency with which a word was encountered within the sample. For instance, the word "date" might occur 10 times in an encyclopeadia article on calendars; once in an article on Egypt; 3 times in an article on dried fruit; 0 times in an article on lasers; and so on. So, each word is associated with a row vector of frequencies across text samples; and each text sample is associated with a column vector of frequencies across words. If it is true that words with similar meanings occur in similar contexts, then the vectors for words with similar meanings should point in similar directions. The similarity structure of the word-to-text co-occurrence matrix thus captures information about the semantic relatedness of the individual words. To get at this structure, the elements of the matrix are usually transformed to minimize variation due to overall word frequencies (for

instance, by taking the log of the frequencies in each cell); the co-occurrence matrix is converted to a similarity matrix by computing the pairwise correlation between all rows; and the similarity matrix is then subject to a singular value decomposition (a computationally efficient means of estimating eigenvectors in a very large similarity matrix). The singular-value decomposition returns a large set of orthogonal vectors that re-describe the similarity matrix (one for each word in the corpus); typically all but the first 300 or so of these vectors are then discarded. The resulting representation contains a description of each word in the corpus as a vector in a ~300-dimensional space.

What is remarkable about this process is that the similarity structure of the resulting vectors appears to parallel, sometimes with surprising accuracy, the semantic similarities discerned by human subjects amongst the words in the corpus. Semantic distances yielded by LSA and similar measures correlate with the magnitude of contextual semantic priming effects in lexical decisions tasks (Landauer & Dumais, 1997); with normative estimates of the semantic relatedness between pairs of words and with word-sorting (Landauer et al., 1998); with the likelihood of confusing two items in free-recall list-learning tasks (Howard & Kahana, 2002); and so on. Such correspondences would seem to suggest some non-arbitrary relationship between the representations computed by LSA-like methods and the word-meaning representations existing in our minds. But what, specifically, is the nature of this relationship?

At the very least, LSA demonstrates that overlap in linguistic context can convey considerable information about the degree to which different words have similar meanings. In short, Elman's speculation—that words with similar meanings appear in similar contexts—appears to be true in actual language. So a learning mechanism that is

sensitive to the temporal context in which words occur may help to promote the learning of semantic similarity relationships. Nevertheless the skeptic might justly question the conclusion that semantic representations can be derived solely from a "dumb" word- or phrase-prediction algorithm (Glenberg & Robertson, 2000). Surely there is more to meaning than simply being able to anticipate which words are likely to follow one another in speech—you can't learn a language just by listening to the radio. And indeed, as a theory of semantic processing, LSA raises many questions. How are the semantic representations it computes—abstract vectors in a high-dimensional space—accessed by perceptual and linguistic input; how do they support naming, action, and other behaviors; and how are they influenced by these non-linguistic aspects of experience? What content do they have?

One response to these criticisms is as follows. LSA shows that sensitivity to high-order temporal structure in language can yield important information about semantic similarity structure. Empirical studies show that the human semantic system is sensitive to the similarity structure computed by LSA-like measures; so it is possible that the human semantic system is also sensitive in some degree to high-order temporal structure in language. But this is not to say that the semantic system is not also sensitive to structure in other aspects of experience: the same learning mechanisms that extract information from statistical structure in speech may also operate on non-linguistic perceptual information to support predictions about future events or appropriate actions; and such a mechanism might even assimilate high-order patterns of covariation between linguistic and non-linguistic sources of information, so that the resulting representations support predictions, not just about what words are likely to follow a given statement, but

also about upcoming perceptual experiences as well. That is, LSA demonstrates that the human mind is sensitive to temporal structure in one aspect of experience (linguistic experience), and the same mechanism that gives rise this sensitivity may also mediate learning in other perceptual and motor domains.

4.3 The Sentence Gestalt model.

The notion that the semantic system might capitalize on statistical structure both within language and between language and other aspects of experience is apparent in many current theories of conceptual knowledge. This idea has not been very directly implemented in any computational model, for obvious reasons—it requires fairly explicit theories about perception, action, speech production and comprehension, all of which constitute broad and controversial domains of study in their own right! Important progress in this vein was made, however, by St. John and McClelland (McClelland et al., 1989; St. John & McClelland, 1990; St. John, 1992).

St. John and McClelland were interested in investigating verbal comprehension, not just for individual words, but for full sentences describing mini-events. Each sentence described an agent performing some action upon some recipient, often with a particular instrument—thus understanding of each event required knowledge of who the actor was, which action was taken, what item was acted upon, and what instrument was used. When one comprehends a sentence such as “The lawyer ate the spaghetti with the fork,” for instance, one knows that the lawyer is the thing doing the eating; the spaghetti, and not the fork, is what is eaten; the fork is being used by the lawyer in order to eat the spaghetti; and so on. Comprehension of such utterances requires combining the meanings

of the individual constituent words, as they come into the semantic system, into some whole or Gestalt representation of the event.

Exactly how such combination is accomplished is the subject of considerable research in psycholinguistics, extending well beyond the scope of this article. To illustrate just one of the complexities attending the question, consider the sentence "The lawyer ate the spaghetti with the sauce." Structurally, it is identical to the example sentence in the preceding paragraph; but the interpretation of the final noun ("fork" versus "sauce") is strikingly different. In the former sentence, the noun is interpreted as the instrument of the action "ate"; in the latter, the noun is interpreted as a modifier of the recipient "spaghetti." Understanding how the final noun "attaches" to the other concepts in the sentence seems to require knowledge of certain constraints deriving from the meaning of the full event—for instance, that it is impossible to use sauce to pick up and eat spaghetti noodles, or that it is unlikely that the spaghetti was served with a topping made of forks. That is, the attachment of the noun derives neither from the structural/syntactic properties of words (which should be identical in the two sentences), nor from the meanings of individual words taken in isolation (e.g. "not used as a topping for spaghetti" is not likely to be a salient property of the concept "fork").

--Figure 6 about here--

St. John and McClelland were interested both in providing a general framework for thinking about comprehension of the "whole meaning" of sentences, and in addressing attachment phenomena like that summarized above. The model they used to exemplify the framework is illustrated in Figure 6. The first bank of input units consists of localist representations of the individual words that occur in sentences. These feed

forward to a bank of hidden units which in turn feed forward to a simple recurrent layer labeled “Sentence Gestalt.” The “copying over” of patterns from the Sentence Gestalt to a context layer (labelled “Previous Sentence Gestalt”) layer allows the network to retain an internal representation of the full sequence of words preceding a current input. This context representation, rather than feeding back directly to the Sentence Gestalt, instead feeds forward into the first hidden layer, thus influencing the pattern of activation that arises there in response to a particular word input. That is, the first hidden layer forms a representation of a particular word encountered in a given sentence context; this context-sensitive word representation then feeds forward to influence the current Sentence Gestalt.

Finally, the Sentence Gestalt units feed forward to another bank of hidden units, but these also receive inputs from the layer labeled “Query.” The Query units themselves are set directly by the environment, and contain localist representations of basic questions one can pose to the network about the meaning of the sentence—questions such as “Who is the actor?”, “Who is the recipient?”, “What was the action?”, “What was the instrument?” and so on. The output layer, then, contains localist representations of the words that constitute answers to these questions: single units coding the various potential agents, recipients, actions, instruments, and so on. So the full model can be viewed as containing two parts: a “comprehension” or input system that retains representations of sequences of words in the sentence to be comprehended, and a “query” or output system that “interrogates” the model’s internal representations in order to answer questions about the meaning of the sentence. The Sentence Gestalt layer codes the representations that intermediate between these two networks.

The model's task is to take in a series of words corresponding to a meaningful sentence, and to correctly respond to queries about the sentence's meaning (ie, answer questions about "who did what to whom with what"). To find a set of weights that accomplish this task, the model is trained with backpropagation. The unit corresponding to the first word of the sentence is activated in the input, and activation flows forward through the network to the output. The network is then "queried" by activating each of the possible question-inputs in turn. With each query, the network's actual response is compared to the correct response, and the error is computed and backpropagated through all the weights in the network. Next the activation of the Sentence Gestalt layer is copied over to the Context layer; the next word in the sentence is activated in the input; input flows forward through the network to the outputs; and the model is queried again with all the various question-inputs. Effectively the model is "asked" about the full meaning of the sentence as each word comes into the input, and is trained with backpropagation to make its "best guess" as to the answers to those questions at each step in processing.

Let's set aside for a moment questions about the naturalness of this training regime, and consider the model's behavior after learning. The trained network could be presented with a full sentence (each word coming in, one at a time, in order), leading it to build up distributed pattern of activity in the "Sentence Gestalt" layer. The information coded in this representation could then be probed by activating different "Query" units—effectively asking the network to answer questions about the meaning of the sentence. The first remarkable thing was that the network could indeed successfully answer the questions. That is, although the model's internal representation of the sentence—the pattern of activity across the "Sentence Gestalt" layer—was not directly interpretable in

and of itself, it produced the correct answers to all of the probe questions, indicating that it somehow "contained" the full meaning of the test sentences.

The second remarkable thing was that this ability generalized fairly well to test-sentences the network had never before seen. For instance, when given a sentence like "the policeman ate the spaghetti with the fork," the network could correctly state that the policeman was the actor and the fork was the instrument, despite never having seen a sentence in which the policeman used a fork. The third remarkable thing was that the network's generalization behavior was sensitive to just the kinds of conceptual constraints exemplified above. When given a sentence like "The policeman ate the spaghetti with the sauce," for instance, it correctly concluded that "sauce" must be a modifier of "spaghetti," and not an instrument of "policeman" (again despite never having been trained on sentences involving policemen and spaghetti). In general, instruments associated with human beings (e.g. "fork") would tend to attach to nouns describing human beings, even when the pairings had never before been encountered; and nouns that tended not to be used as instruments did not attach to agents, even in novel sentences. The basis for this generalization should be apparent from the previous discussion of Elman's work. Human agents tend to engage in many of the same kinds of activities, using some of the same kinds of instruments; this overlap leads the SG model to represent the various different human nouns as somewhat similar to one another in the first hidden layer (and different from non-human agents), and this similarity promotes generalization to new sentence contexts.

There are other appealing aspects of the Sentence Gestalt model that will not be reviewed here. Instead it is worth focusing briefly on a seemingly artificial nature of the

training regime: the fact that the model is "queried" with all possible questions with each new word presentation, and gets faithful answers to every question during training. To what could such training possibly correspond in the real world? One answer to this question is that the training regime in St. John and McClelland's work provides a coarse proxy to the covariation of language with other aspects of experience. The verbal statements that children are trying to understand as they learn a language do not occur in isolation, but together with other sensory-motor information. When daddy says, "Look, mommy's eating her dinner with a fork!," the infant may look up to see mommy holding a fork, jamming it into the spaghetti noodles, and raising it to her mouth. The agent, action, recipient and instrument information is all contained in this event. Although children may not be explicitly querying themselves about these relationships as the SG model does, they may be doing something related—trying to anticipate who will pick up the fork, or what mommy is holding onto, or what will go into the mouth, and so on, when they hear daddy's statement and look up toward mommy. That is, correspondences between verbal statements and actual observed events may provide the statistical basis for learning to represent the meanings of full sentences.

Summary

The thread of research described in Section 3 suggested that the semantic system may serve a particular functional role: the ability to make context-appropriate inferences about the properties of objects, given their name or some other perceptual input. To accomplish this role, it is necessary for the semantic system to represent conceptual similarity relationships amongst familiar items, and to adapt these relationships as necessary according to the situation or context. Section 3 suggested that some of the

information necessary to acquire such knowledge may be present in the overlap of sensory and motor properties across different modalities and across various different situations. The work précised in the current section adds to this suggestion by showing how the semantic system can become sensitive to temporal structure, both within language and between language and other aspects of experience. Elman's (1990) work provided a simple mechanism for learning temporal structure; the work of Landauer and Dumais (1997), Burgess and Lund (1997), and others has shown how rich such structure can be, even just considering temporal structure in natural language; and the Gestalt models described by McClelland and St. John (St. John & Gernsbacher, 1995; St. John & McClelland, 1990; St. John, 1992) provide a simple framework for thinking about how coherent covariation between linguistic structure and other aspects of experience can promote the representation of meaning for full sentences and events. These developments thus begin to offer leverage on the three issues that remained unaddressed or only partially addressed at the end of the last section: the context-sensitive nature of concepts; the representation of meaning for abstract words; and the representation and processing of full events encompassing multiple items.

5. Neuro-cognitive models

All of the models reviewed thus far are best construed as cognitive models—they offer limited insight at best as to the nature of the neural systems and processes that support semantic abilities. The final thread of research considered here encompasses neuro-cognitive models. The great majority of this work has focused on understanding impairments to semantic abilities following brain damage. Two principal questions addressed by this work are: i) How can patterns of observed semantic impairment be

explained given what we know about the cortical organization of information-processing systems in the brain, and ii) What do patterns of semantic impairment tell us about the neuroanatomical organization of the semantic system?

Until very recently, these questions have been pursued more-or-less independently of the computational issues discussed in the previous two sections. In this final section, we will consider the two most widely-studied forms of semantic impairment, and the models that have been proposed to explain them. We will see that, although these models share many properties in common, they differ in important respects that have implications for the view of semantic abilities considered in Sections 3 and 4.

5.1 Category-specific semantic impairment

The first form of semantic impairment we will consider is category-specific impairment: semantic deficits that appear to be restricted to one semantic domain while largely sparing others. By far the most commonly observed category-specific impairment involves seriously degraded knowledge of living things, with comparatively good knowledge of manmade objects (Capitani, Laiacona, Mahon, & Caramazza, 2003; Martin & Caramazza, 2003; Warrington & McCarthy, 1983). The reverse dissociation has, however, also been reported (Warrington & McCarthy, 1987; Warrington & Shallice, 1984), along with other apparently selective semantic deficits (Crutch & Warrington, 2003; Samson & Pillon, 2003), seeming to indicate that different forms of brain damage can differentially affect knowledge of different semantic domains. One straightforward interpretation of this impairment is that different parts of the brain have been "specialized" over the course of evolution for storing and retrieving semantic information

about living and nonliving things (Caramazza, 1998). In early discussions of apparent category-specific impairments, however, Warrington and Shallice (1984) suggested an alternative explanation: perhaps semantic representations of living things depend to a greater extent on knowledge of perceptual qualities, whereas semantic representations of manmade objects depend more upon knowledge of their functional characteristics. If so, then damage to regions of the brain that support knowledge of visual attributes may produce a seeming "living-things" deficit, whereas damage to regions that support knowledge of action or function may produce an apparent "manmade object" impairment.

This hypothesis had appeal for at least two reasons. First, it was consistent with what was already known about the functional organization of cortex. That is, cortical regions supporting visual perception of objects are quite removed from those that support action/object use—so the hypothesis offered a means of understanding the pattern without requiring the ad-hoc proposal of separate cortical regions for representing different kinds of concepts. Second, the hypothesis explained a few apparent exceptions to the supposed "category-specific" patterns. For instance, some patients with "living things" impairments were also seriously impaired at naming and recognizing musical instruments and minerals—artifacts that might well depend to a greater extent than usual upon knowledge of perceptual characteristics. Similarly, some patients with "manmade object" impairments also showed deficits for recognizing body-parts—arguably "living things" that are closely tied to knowledge of action and function (Warrington & Shallice, 1984).

--Figure 7 about here--

An influential computational implementation of the sensory-functional hypothesis was put forward by Farah and McClelland (1991). In addition to demonstrating that the theory was indeed tractable, simulations with the model showed that it also had some counterintuitive implications. The model, illustrated in the top panel of Figure 7, is a fully recurrent network, in which activation may flow in either direction between connected layers. For instance, visual input to the Visual layer can flow up to the Semantic layer, and then in turn to the Verbal layer; or alternatively input from the Verbal layer can flow to the Semantic layer and back to the Visual layer. Thus the units in the Semantic layer may be construed as computing mappings between visual and verbal information presented from the environment.

Representations of objects in the model take the form of distributed patterns of activity across groups of units. The units themselves can be thought of as each responding to some aspect of the entity represented by the whole pattern, though these aspects need not be nameable features or correspond in any simple way to intuitions about the featural decomposition of the concept. In the semantic layers, some units may respond to objects with some particular visual property, while others may respond to aspects of the object's functional role. In the visual layer, patterns of activity correspond to more peripheral visual representations; while patterns of activity in the verbal layer form representations of words. To present a visual stimulus to the network, the corresponding pattern of activation is clamped across Visual units; these activations feed forward to Semantic units, then on to Verbal units. The activations of Verbal units can then feed back to the Semantic units, and this dynamic flow of activation proceeds until the unit states stop changing, at which point the network is said to have settled into a

steady state or attractor. The location of such stable configurations depends upon the connection weight matrix. The role of learning in this model is to configure the weights in such a way that, when the network is presented with a particular word or picture as input, it will settle into a stable state in which the correct pattern of activity is observed across units in the visual, verbal, and semantic layers.

Farah and McClelland (1991) created representations for ten “living” and ten “nonliving” objects, by generating random patterns of -1 and +1 across all three layers of units in the model. Each unique pattern corresponded to a representation of an individual item. Representations of living and nonliving things differed only in the proportion of active semantic units in the “functional” and “perceptual” pools. These were set to match the observed ratio of perceptual to functional features of objects in dictionary definitions. Living things in the model were represented with an average of 16.1 visual and 2.1 functional units active; whereas nonliving things were represented with an average of 9.4 visual and 6.7 functional units active. All patterns had some units active in both semantic pools. The verbal and visual representations were random patterns generated in the same way for living and nonliving items. To find a configuration of weights that would allow the network to perform correctly, the model was trained with the delta rule (McClelland & Rumelhart, 1985) to associate Visual and Verbal patterns with the appropriate Semantic pattern. When the model had finished learning, it could generate the correct Semantic pattern from any Verbal or Visual input; and activation of this pattern would then correctly “fill in” the corresponding Verbal or Visual pattern.

Of interest was the model's behaviour when its semantic units were damaged. Under the sensory-functional hypothesis, units representing the functional-semantic

aspects of an item can be damaged independently of the units representing the item's perceptual-semantic properties. How did the model's performance deteriorate with increasing damage to each of these pools of units? To simulate neural trauma in the network, Farah and McClelland simply deleted some proportion of the units in either the perceptual semantic pool or the functional semantic pool. They then tested the network's ability to perform model analogues of picture naming and match-to-sample tasks. In the former, the model was presented with the picture of an object (by applying a pattern of activity to the visual units), and allowed to settle to a steady state. The resulting pattern of activity across the word units could then be read off, and compared to all the patterns in the training corpus. The model's response was considered correct if the pattern of activity across word units more similar to the correct pattern than to any other pattern. The same procedure was employed in the match-to-sample task, using a word as input and examining patterns of activity across visual units to determine the response.

Two aspects of their results are of interest. First, the model showed a clear double dissociation in its ability to name and match living and nonliving things. When visual semantic units were destroyed, the model exhibited a greater naming impairment for living relative to nonliving objects. The opposite was true when functional units were destroyed. Second, and more interesting, in neither case was the model completely unimpaired in the “spared” domain. Though the model was worse at naming living things when perceptual semantic features are destroyed, it was also impaired at naming nonliving things. Living things rely more heavily on perceptual semantic features in the model, but such features inform the representation of both living and nonliving objects to some degree. As this knowledge deteriorates in the model, it tends to affect naming

performance for both domains, albeit to differing degrees. The same graded impairments are also witnessed in the patient data—profound impairments in one domain are almost without exception accompanied by mild impairments in the relatively spared domain.

Farah and McClelland (1991) also examined the network's ability to retrieve functional and perceptual semantic information when given a picture or a word as input. Considering only the perceptual or the functional unit pools, they compared the pattern of activity in the damaged network when it had settled to the correct pattern, for each object. The network was considered to have spared knowledge of the perceptual properties of an item if the observed pattern of activity across perceptual semantic units was closest to the correct pattern; and spared knowledge of functional properties if the observed pattern across functional semantic units was closest to the correct pattern.

The simulations showed that the loss of semantic features in one modality had important consequences for the model's ability to retrieve properties in the spared modality. When perceptual semantic features were lost, the model had a tendency to generate an incorrect pattern of activity across functional semantic units, especially for living things. The reason is that the reciprocal connections among semantic features lead the network to rely on activity in perceptual semantic units to help produce the appropriate patterns across functional units. When this activation is reduced or disrupted as a result of damage, these lateral connections can interfere with the model's ability to find the correct states even in the spared units. Thus, the loss of “perceptual” semantic knowledge can precipitate a disruption of knowledge about <u>functional</u> properties, especially for categories that rely to a large extent on perceptual information in their representation. Of course, the reverse is true when functional semantic features are

damaged. So, counter-intuitively, it is not the case that patients with worse knowledge of animals than artifacts should always show preserved knowledge of functional properties under the theory—even though the theory attributes the apparent category effect to the loss of knowledge about sensory properties of objects.

5.2 The convergence model

The second well-studied form of semantic impairment is the progressive and profound degeneration of semantic knowledge observed in the syndrome known as semantic dementia (SD). There are three remarkable facts about SD that constrain theories about the neural basis of semantic abilities. First, the semantic impairment appears to encompass knowledge of all kinds of concepts, tested in all modalities of reception and expression. In contrast to the "category-specific" cases described above, for instance, patients with SD show equally poor knowledge about living and nonliving things (Garrard, Lambon Ralph, & Hodges, 2002). They are profoundly anomic (Hodges, Graham, & Patterson, 1995; Lambon Ralph, Graham, Ellis, & Hodges, 1998; Rogers, Ivanoiu, Patterson, & Hodges, 2006), but their impairments are not restricted to language: they show serious deficits recognizing line drawings of common objects (Rogers, Lambon Ralph, Hodges, & Patterson, 2003), drawing pictures of objects after a brief delay (Bozeat et al., 2003), colouring black-and-white line drawings of common objects (Rogers, Patterson, Hodges, & Graham, 2003), assessing the usual function of every-day objects (Bozeat, Lambon Ralph, Patterson, & Hodges, 2002), matching a sound (such as a telephone ring) to a picture of the item that makes the sound (Adlam, Rogers, Salmond, Patterson, & Hodges, in press; Bozeat, Lambon Ralph, Patterson, Garrard, & Hodges,

2000)—effectively any task that requires them to make an inference about an object's properties (regardless of whether the item is depicted or denoted by a word).

Second, other aspects of cognitive functioning are remarkably spared in the disorder. Patients with SD are generally well-oriented in space and time; show comparatively normal episodic and recognition memory; have speech that is grammatical and, apart from word-finding problems, fluent; have normal or near-normal perception; show no attentional dysfunction; and perform well on tests of reasoning and problem-solving (Patterson & Hodges, 2000).

Third, the neuropathology that produces SD is not widespread in the brain, but is relatively circumscribed. The condition follows from the temporal-lobe variant of fronto-temporal dementia, a disease that produces a slowly-progressing deterioration of cortical gray matter in the anterior temporal lobes of the brain. Although the pathology is often more pronounced in the left hemisphere, it is virtually always bilateral, and in some cases can be worse in the right hemisphere.

On the basis of these observations above, Rogers et al. (2004) proposed a theory about the neural basis of semantic memory, illustrated in the bottom panel of Figure 7. Like the approaches discussed in Sections 3 and 4, the theory proposes that semantic memory serves a key function: to promote inferences about the properties of objects and events that are not directly perceived in the environment. For instance, when encountering a line drawing of a banana, representation of the depicted object's shape may depend predominantly upon perceptual and not semantic processes; but the semantic system then promotes retrieval of the item's name, its characteristic color, its taste, the

actions required to peel it, and so on. In this sense, the "meaning" of the image inheres in the coactivation of various associated sensory, motor, and linguistic representations.

Different kinds of sensory, motor, and linguistic information are known to be coded in widely distributed and functionally specialized cortical regions—with some regions specialized, for instance, for colour perception, others for motion perception, others for representation of orthographic or phonological words forms, and so on (Chao, Haxby, & Martin, 1999; Martin & Chao, 2001). On the basis of the neuroanatomical observations from SD, Rogers et al. (2004) suggested that these widely-distributed sensory, motor, and linguistic representations communicate with one another via the anterior temporal-lobe regions affected in SD. That is, the anterior temporal lobes act as a kind of "hub" or "convergence zone" (Damasio & Damasio, 1994) that promotes the interactive activation of linguistic, perceptual and motor representations. When the hub deteriorates as a consequence of disease, this degrades the ability to map between such surface forms.

Rogers et al. (2004) used a simplified implementation of the theory to illustrate some desirable consequences of this proposal. The model's architecture (the black ovals in the second panel of Figure 7) was similar to that of the Farah-McClelland model: it included a layer to code visual shape representations, a layer to code verbal inputs/outputs, and an intermediating hidden layer (labeled "Semantics" in the Figure). Units in the Visual layer were understood to represent visual properties of objects that could be directly perceived; whereas units in the Verbal layer were understood to represent individual words. Visual and Verbal units could get direct input from the environment, and both layers send connections to and received connections from the

intermediating Semantic units. Thus the model could be presented with a visual input (corresponding to a pattern of activity across Visual units), a single name or word (corresponding to activation of a single Verbal unit), or a phrase describing an object's properties (corresponding to a pattern of activation across Verbal units).

This Convergence model contrasted with the Farah-McClelland model in three important ways. First, the patterns of activity that constituted the Visual and Verbal representations were not random vectors, but instead captured aspects of similarity apparent in line drawings of common objects, and in the verbal statements we tend to make about such objects. That is, items with many visual properties in common were represented with overlapping patterns in the Visual layer; whereas items to which similar spoken predicates apply were represented with similar patterns in the Verbal layer. Second, no "semantic" representations were assigned. Instead, the model was simply trained (using a backpropagation algorithm suited to recurrent networks) to complete mappings between individual names, visual representations, and verbal descriptions of various objects. The patterns of activation that arose across Semantic units in the trained models thus constituted learned internal representations, just as in the models described in Sections 3-4. Third, the Convergence model proposed no functional specialization of the intermediating semantic units.

Rogers et al. (2004) simulated the neuropathology of SD by removing an increasing proportion of the weights projecting into or out from the Semantic layer. The simulation experiments were able to replicate several interesting aspects of impairment in SD, and made a variety of new predictions about the consequences of temporal-lobe damage for semantic memory (Lambon Ralph, Lowe, & Rogers, in press; Rogers et al.,

2004). Rather than reviewing all of these results, we will instead focus on one aspect of the simulations that provides a clue as to why the cortical semantic network might employ a convergent architecture.

The key observation concerns the fact that the learned internal representations in the model end up capturing the semantic similarity relations existing amongst the items in the training corpus, for essentially the same reasons discussed earlier with respect to the Rumelhart model. More interestingly, the authors showed that these acquired similarity relations differed from those apparent in the overlap of the model's Visual and Verbal patterns considered independently. Specifically, from overlap in visual features, the category of fruits was largely intermingled with manmade objects; whereas, from overlap in verbal features, the same items were represented as quite distinct from both manmade objects and from animals. The internal representations formed across Semantic units in the Convergence model captured a blend of these similarity relations: Fruits were represented as i) similar to one another, ii) distinct from both manmade objects and animals, but iii) considerably more similar to the former than the latter. This counter-intuitive finding (that fruits may be represented as more similar to manmade objects than to animals) predicted that patients with SD should be more likely to confuse fruits with artifacts than with animals, a prediction that was confirmed in a subsequent sorting experiment (Rogers et al., 2004).

In other words, the simulation showed that the intermediating representations that arise from learning in a convergent architecture can capture similarity structure that is not directly apparent in any individual surface representation. This observation is important precisely because surface representations—the sensory, motor, and linguistic

representations from which “meanings” are thought to arise—often do not seem to faithfully capture semantic/conceptual similarities. Lightbulbs and pears may have similar shapes; fire engine and strawberries have similar colours; potato-mashers and plungers engage similar praxis; and so on. The Convergence model suggests that, although conceptual similarity structure may not be directly captured by any of these surface representations, it may be apparent in the pattern of overlap across the different kinds of representation. Thus the explanation as to why, computationally, the cortex should employ a convergent architecture is as follows: To acquire representations that capture conceptual similarity relations (and thus promote appropriate generalization of stored information to newly encountered items), the semantic system must be sensitive to overlap across widely-distributed surface representations; and such sensitivity depends in turn upon there being, somewhere in the cortical semantic network, a region where all these different kinds of information converge.

Summary

The two semantic syndromes described above seem to point to different conclusions about the neuroanatomical organization of the semantic system. Studies of patients with apparent category-specific impairment seem to suggest that there exists a certain degree of functional specialization within the semantic system, and theorists vary considerably in their opinions as to the degree and of nature of such functional specialization. On the other hand, studies of patients SD seem to suggest that there exists in the anterior temporal cortex a relatively circumscribed region that is critical to semantic processing for all variety of concepts and all modes of reception and expression. The Farah-McClelland model may be viewed as an effort to find a middle way between a

complete balkanization of the semantic system and a fully homogeneous system. Related efforts have been put forward by Plaut (2002), Humphreys and Forde (Humphreys & Forde, 2001), Tyler and colleagues (Tyler, Moss, Durrant-Peatfield, & Levy, 2000), Devlin and colleagues (Devlin, Gonnerman, Andersen, & Seidenberg, 1998), Lambon Ralph, Lowe and Rogers (in press), and many others. Although there is as yet no clear consensus as to the resolution of these issues, it is apparent that computational models are providing important tools for an increasing number of researchers interested in the neural basis of semantic abilities.

An important direction for future efforts will be to relate these neuro-cognitive models back to the computational issues motivating the more abstract models reviewed in earlier sections of this chapter. That is, rather than asking "What architecture best explains the pattern of sparing and impairment observed from different forms of brain damage," we may begin to ask, "What architectures yield the computational properties that, from more abstract semantic theories, we believe the semantic system must possess?"

Conclusion and open issues

This overview indicates that, for many of the challenging puzzles currently facing research in human semantic memory, the beginnings of answers exist in the literature. Important questions about category coherence and feature weighting may be addressed by the fact that certain network architectures promote sensitivity to high-order covariance structure amongst stimulus properties across different modalities of reception and expression. Context-sensitivity may also reflect sensitivity to higher order correlational structure, in that any particular situation or context constrains which of an item's

properties are “important” or relevant, and which similarity relationships are best used to govern generalization and induction. One way of understanding such influences is to proposed that the distributed semantic representations that govern performance in the moment are shaped, not only by the particular item in question, but also by a representation of the current context, as is the case in the Rumelhart model (see also the chapter on cognitive control by De Pisapia, Repovs and Braver). Finally, the semantic system’s ability to comprehend full events, as well as its knowledge of “abstract” properties—properties that are not plausibly instantiated directly in sensory and motor systems—may derive, at least in part, from its sensitivity to temporal structure.

Important directions for future work involve drawing these various threads together, in three different respects. First, the existing work is dispersed across a variety of models employing quite different architectures, differing degrees of abstraction, and different assumptions about the nature learning and of the information available to the semantic system. It is not clear how the different pieces fit together into a single framework—a model in which coherent covariation amongst perceptual, motor, and linguistic properties, and sensitivity to temporal structure, and representation of task context, all contribute together to semantic representation and processing. Clearly the development of such a model is beyond the current state of the art, but important next steps will involve addressing at least some components of this uber-system.

Second, this chapter has focused predominantly on parallel-distributed-processing approaches to semantic memory—not because there are no other computational approaches, but because these other approaches typically focus on a slightly different set of issues. For instance, semantic memory is clearly important for human induction and

inference; but induction and inference also constitutes a domain of study in its own right, in which Bayesian approaches are probably most influential. Similarly, studies of categorization, though clearly overlapping with issues addressed here, also constitute a separate domain of study, in which mathematical approaches (including prototype and instance-trace models) are the norm. As previously mentioned, these overlapping domains of study, and the methods they adopt, are reviewed in other chapters of the Handbook. An important direction for future research in semantic cognition and in these other domains will be to understand whether the theoretical approaches adopted there differ fundamentally from those described in the current chapter, or whether they constitute different formal descriptions of the same underlying processes.

Finally, there is clearly much to be done in relating computational theories of semantic abilities to information processing in the brain. Although most theories about the neural basis of semantic cognition support the notion that semantic memory arises from the association of perceptual, motor, and linguistic representations that are widely distributed in the brain, there remain many open questions about the structure and properties of the cortical semantic network. For instance, how can sensory-motor learning lead to knowledge of conceptual similarity relations? How are abstract properties represented in the brain, if the semantic system is built upon sensory and motor properties? How does the brain achieve the flexibility and context-sensitivity observed in the semantic system? What cortical mechanisms support conceptual development, and to what extent are these driven by experience versus maturation? The simulations reviewed in the current chapter provide intriguing clues about the answers to these questions; the

next decade of research will need to integrate these computational ideas with the emerging picture from neuroscience.

References

Adlam, A.-L., Rogers, T. T., Salmond, C. H., Patterson, K., & Hodges, J. R. (in press). Semantic dementia and fluent primary progressive aphasia: Two sides of the same coin? *Brain*.

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--178.

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

Bertenthal, B. (1993). Infants' perception of biomechanical motions: {Instrinsic} image and knowledge-based constraints. In C. Grandrud (Ed.), *Visual perception and cognition in infancy* (pp. 175--214). Hillsdale, NJ: Erlbaum.

Billman, D., & Knutson, J. (1996). Unsupervised concept learning and value systematicity: {A} complex whole aids learning the parts. *Journal of Experimental Psychology: {Learning}, Memory, and Cognition, 22*, 458--475.

Bodner, G. E., & Masson, M. E. (2003). Beyond spreading activation: An influence of relatedness proportion on masked semantic priming. *Psychonomic Bulletin and Review, 10*(3), 645-652.

Bozeat, S., Lambon Ralph, M. A., Graham, K. S., Patterson, K., Wilkin, H., Rowland, J., et al. (2003). A duck with four legs: Investigating the structure of conceptual knowledge using picture drawing in semantic dementia. *Cognitive Neuropsychology, 20*(1), 27-47.

Bozeat, S., Lambon Ralph, M. A., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Nonverbal semantic impairment in semantic dementia. *Neuropsychologia, 38*, 1207-1215.

Bozeat, S., Lambon Ralph, M. A., Patterson, K., & Hodges, J. R. (2002). When objects lose their meaning: what happens to their use? *Cognitive, Affective, and Behavioral Neuroscience, 2*(3), 236-251.

Bruner, J. S., Goodnow, J. J., & Austin, G. A. (1956). *A study of thinking*.New York: Wiley.

Burgess, C., & Lund, K. (1997). Modelling Parsing Constraints with High-Dimensional Context Space. *LCP, 12*(2), 177--210.

Capitani, E., Laiacona, M., Mahon, B. Z., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology*, *20*(3-6), 213-261.

Caramazza, A. (1998). The interpretation of semantic category-specific deficits: What do they reveal about the organization of conceptual knowledge in the brain? *Neurocase, 4*, 265--272.

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

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience, 2*(10), 913-919.

Collins, A. M., & Loftus, E. F. (1975). A Spreading-Activation Theory of Semantic Processing. *PR, 82*, 407--428.

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

Crestani, F. (1997). Application of spreading activation techniques in information retrieval. *Atrtificial Intelligence Review, 11*(6), 453-482.

Crutch, S. J., & Warrington, E. K. (2003). The selective impairment of fruit and vegetable knowledge: A multiple processing channels account of fine-grain category-specificity. *Cognitive Neuropsychology, 20*(3-6), 355-372.

Csibra, G., Gergely, G., Biro, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: The perception of `pure reason' in infancy. *Cognition, 72*(3), 237--267.

Damasio, A. R., & Damasio, H. (1994). Cortical systems underlying knowledge retrieval. In C. Koch (Ed.), *Large-scale Neuronal Theories of the Brain* (pp. 61-74). Cambridge, MA: MIT Press.

Dell, G. S. (1986). A Spreading-Activation Theory of Retrieval in Sentence Production. *Psychological Review, 93*(3), 283--321.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: {A} computational account. *Journal of Cognitive Neuroscience, 10*(1), 77--94.

Elman, J. L. (1990). Finding Structure in Time. *Cognitive Science, 14*, 179--211.

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--357.

Garrard, P., Lambon Ralph, M. A., & Hodges, J. R. (2002). Semantic dementia: A category-specific paradox. In E. M. Forde & G. W. Humphreys (Eds.), *Category specificity in brain and mind* (pp. 149-179). Hove, UK: Psychology Press.

Gelman, R. (1990). First principles organize attention to and learning about relevant data: Number and the animate/inanimate distinction as examples. *Cognitive Science, 14*, 79--106.

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

Glenberg, A., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory and Language, 43*, 379-401.

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--1222.

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

Hodges, J. R., Graham, N., & Patterson, K. (1995). Charting the progression in semantic dementia: Implications for the organisation of semantic memory. *Memory, 3*, 463--495.

Howard, M. W., & Kahana, M. J. (2002). When does semantic similarity help episodic retrieval? *Journal of Memory and Language, 46*, 85-98.

Humphreys, G. W., & Forde, E. M. (2001). Hierarchies, similarity, and interactivity in object-recognition: On the multiplicity of `category-specific' deficits in neuropsychological populations. *Behavioral and Brain Sciences, 24*(3), 453-509.

Johnson, S. (2000). The recognition of mentalistic agents in infancy. *Trends in Cognitive Science, 4*, 22--28.

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

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

Keil, F. (1989). *Concepts, Kinds, and Cognitive Development*.Cambridge, MA: MIT Press.

Kruschke, J. K. (1992). {ALCOVE:} {An} Exemplar-Based Connectionist Model of Category Learning. *Psychological Review, 99*(1), 22--44.

Lambon Ralph, M. A., Graham, K., Ellis, E., & Hodges, J. R. (1998). Naming in semantic dementia--What matters? *Neuropsychologia, 36*, 125-142.

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (in press). The neural basis of category-specific semantic deficits for living things: Evidence from semantic dementia, HSVE and a neural network model. *Brain*.

Landau, B., Smith, L., & Jones, S. (1988). The importance of shape in early lexical learning. *Cognitive Development, 3*, 299-321.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction, and representation of knowledge. *PR, 104*(2), 211--240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to Latent Semantic Analysis. *Discourse Processes, 25*, 259-284.

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

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

Martin, A., & Caramazza, A. (2003). Neuropsychological and neuroimaging persepctives on conceptual knowledge: An introduction. *Cognitive Neuropsychology, 20*(3-6), 195-212.

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

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General, 114*, 159--188.

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

Medin, D. L., & Shaffer, M. M. (1978). Context Theory of Classification Learning. *PR, 85*, 207--238.

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

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

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

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

Nosofsky, R. M. (1986). Attention, similiarity and the identification-categorization relationship. *JEPLMC, 115*(1), 39--57.

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

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

Plaut, D. C. (2002). Graded modality-specific specialisation in semantics: A computational account of optic aphasia. *Cognitive Neuropsychology, 19*(7), 603-639.

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.

Rogers, T. T., Ivanoiu, A., Patterson, K., & Hodges, J. (2006). Semantic memory in Alzheimer's disease and the fronto-temporal dementias: A longitudinal study of 236 patients. *Neuropsychology, 20*(3), 319-335.

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

Rogers, T. T., Lambon Ralph, M. A., Hodges, J. R., & Patterson, K. (2003). Object recognition under semantic impairment: The effects of conceptual regularities on perceptual decisions. *Language and Cognitive Processes, 18*(5/6), 625-662.

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

Rogers, T. T., Patterson, K., Hodges, J. R., & Graham, K. (2003). *Colour knowledge in semantic dementia: It's not all black and white.* Paper presented at the Cognitive Neuroscience Society Annual Meeting, New York, New York.

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

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology, 7*, 573--605.

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

Rosch, E., Simpson, C., & Miller, R. S. (1976). Structural bases of typicality effects. *Journal of Experimental Psychology: Human Perception and Performance, 2*, 491--502.

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

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

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. In D. E. Meyer & S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in*

Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience (pp. 3--30). Cambridge, MA: MIT Press.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical Learning by 8-Month-Olds. *Science, 274*(5294), 1926--1928.

Samson, D. S., & Pillon, A. (2003). A case of impaired knowledge for fruits and vegetables. *Cognitive Neuropsychology, 20*(3-6), 373-400.

Smith, J. D. (2002). Exemplar theory's predicted typicality gradient can be tested and disconfirmed. *Psychological Science, 13*, 437-442.

St. John, M. A., & Gernsbacher, M. A. (1995). Syntactic comprehension: Practice makes perfect and frequency makes fleet. In *The Seventeenth Annual Conference of the Cognitive Science Society*.Hillsdale, NJ: Cognitive Science Society

Erlbaum.

St. John, M. A., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46*, 217-257.

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

Steyvers, M., Griffiths, T. L., & Dennis, S. (2006). Probabilistic inference in human semantic memory. *Trends in Cognitive Science, 10*(7), 327-334.

Tulving, E. (1972). Episodic and semantic memory. In E. Tulving & W. Donaldson (Eds.), *Organization of memory* (pp. 381-403). New York: Academic Press.

Tyler, L., Moss, H. E., Durrant-Peatfield, M. R., & Levy, J. P. (2000). Conceptual structure and the structure of concepts: A distributed account of category-specific deficits. *Brain and Language, 75*(2), 195--231.

Warrington, E. K., & McCarthy, R. (1983). Category-Specific Access Dysphasia. *Brain, 106*, 859--878.

Warrington, E. K., & McCarthy, R. (1987). Categories of Knowledge: {Further} Fractionation and an Attempted Integration. *Brain, 110*, 1273--1296.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain, 107*, 829-854.

Wilson, R. A., & Keil, F. C. (2000). The shadows and shallows of explanation. In F. C. K. a. R. A. Wilson (Ed.), *Explanation and Cognition* (pp. 87--114). Boston, MA: MIT Press.

Wittgenstein, L. (1953). *Philosophical investigations*.Oxford: Blackwell.

Yeh, W., & Barsalou, L. (2006). The situated nature of concepts. *American journal of Psychology, 119*(3), 349-384.

Zaki, S. F., & Nosofsky, R. (2004). False prototype enhancement effects in dot pattern categorization. *Memory and Cognition, 32*(3), 390-398.

Figure captions

Figure 1. The architecture of Hinton's (1981) seminal model of semantic memory.

Figure 2. Rumelhart's (1990; Rumelhart and Todd, 1993) model, subsequently used as the basis for Rogers and McClelland's (2004) theory of semantic memory.

Figure 3. Multidimensional scaling of internal representations for 8 items at 10 equally-spaced intervals during training of the Rumelhart model. The labelled end-points indicate the similarities amongst the representations at the end of learning, whereas the lines trace the trajectory of these representations throughout learning.

Figure 4. Bottom: Mean Euclidean distance between plant and animal, bird and fish, and robin and canary internal representations throughout training of the Rumelhart model. Middle: Average magnitude of the error signal propagating back to representation units from properties that reliably discriminate plants from animals, birds from fish, or the canary and robin, when the network is presented with the canary as input at different points during learning. Top: Activation of different output properties when the network is queried about the canary. The properties include one shared by animals (can move), one shared by birds (can fly), and one unique to the canary (can sing).

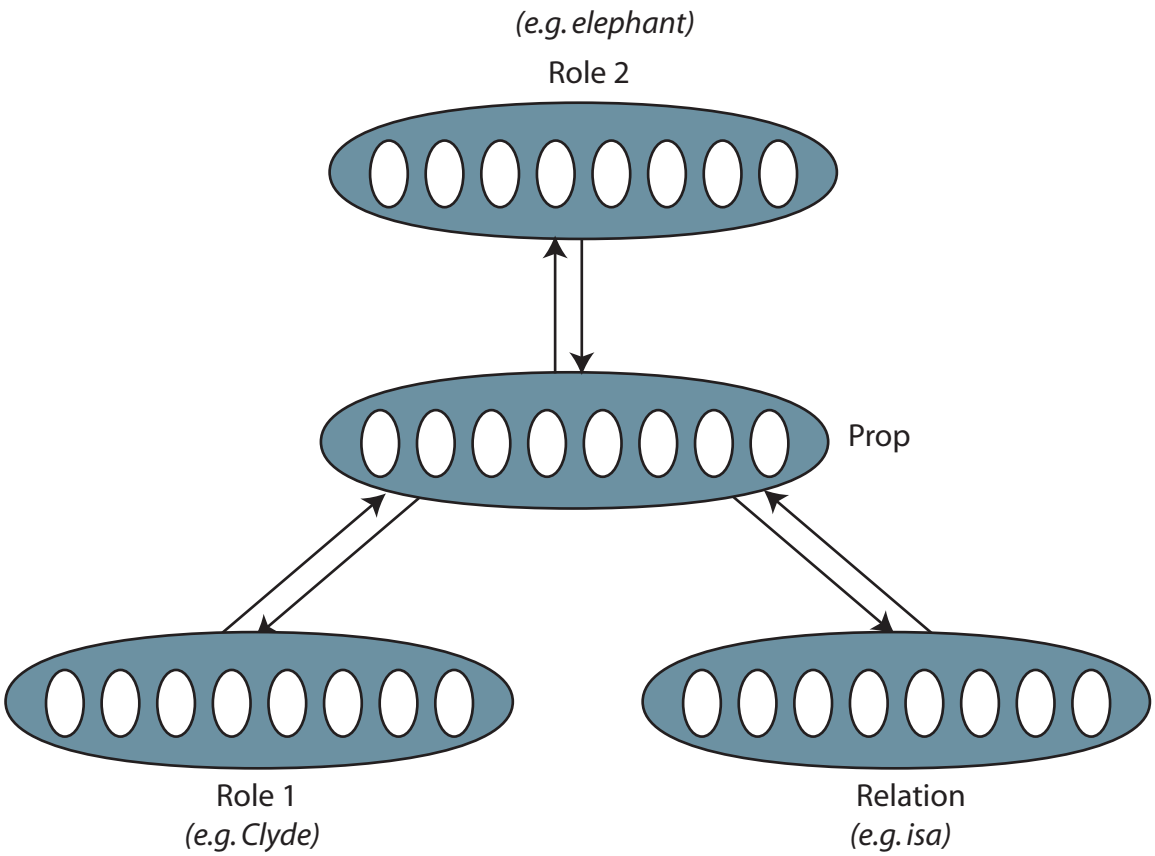
Figure 5. The architecture of a simple recurrent network (SRN; Elman, 1990).

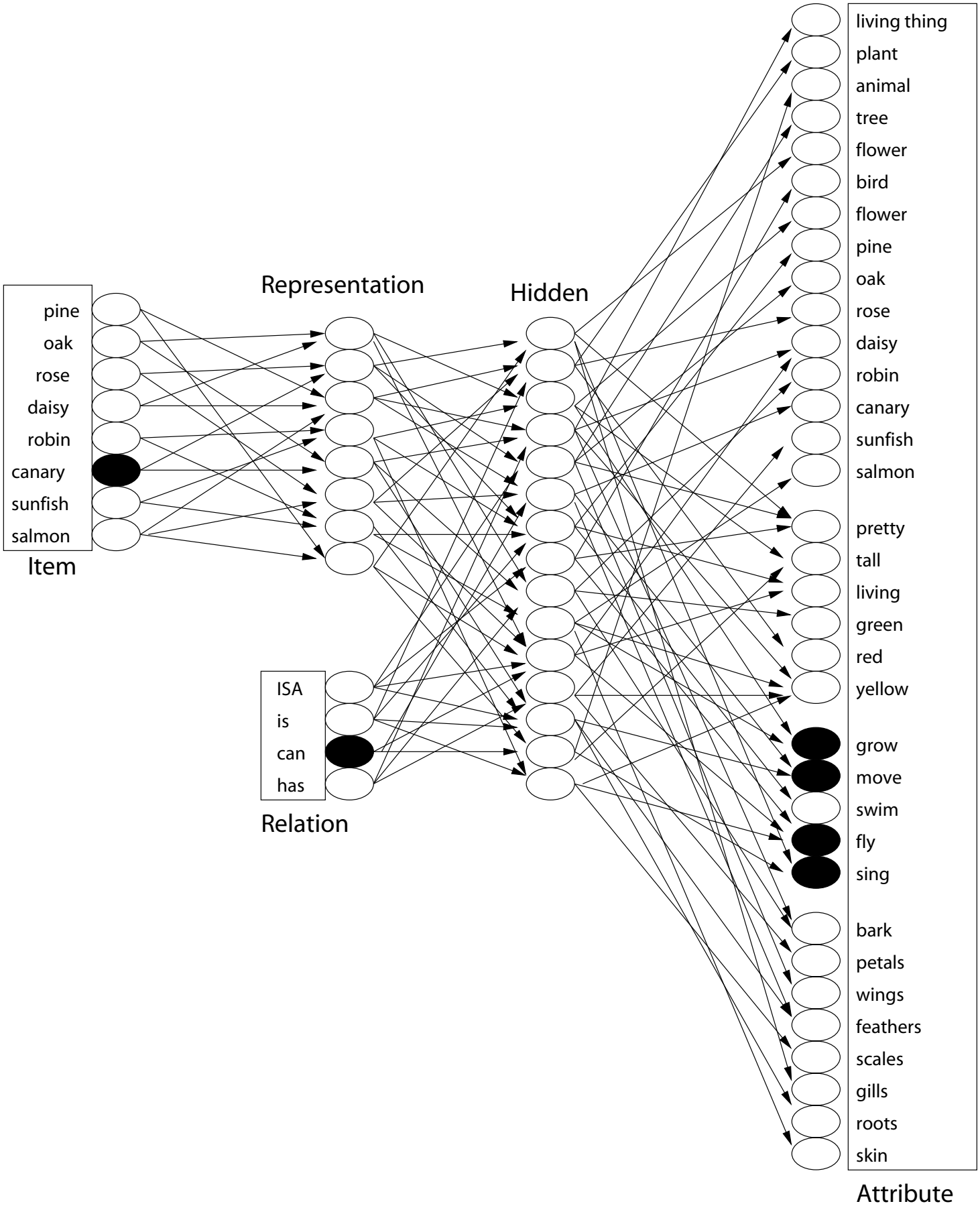
Figure 6. The architecture of the Sentence Gestalt Model (McClelland, St. John and Taraban, 1989).

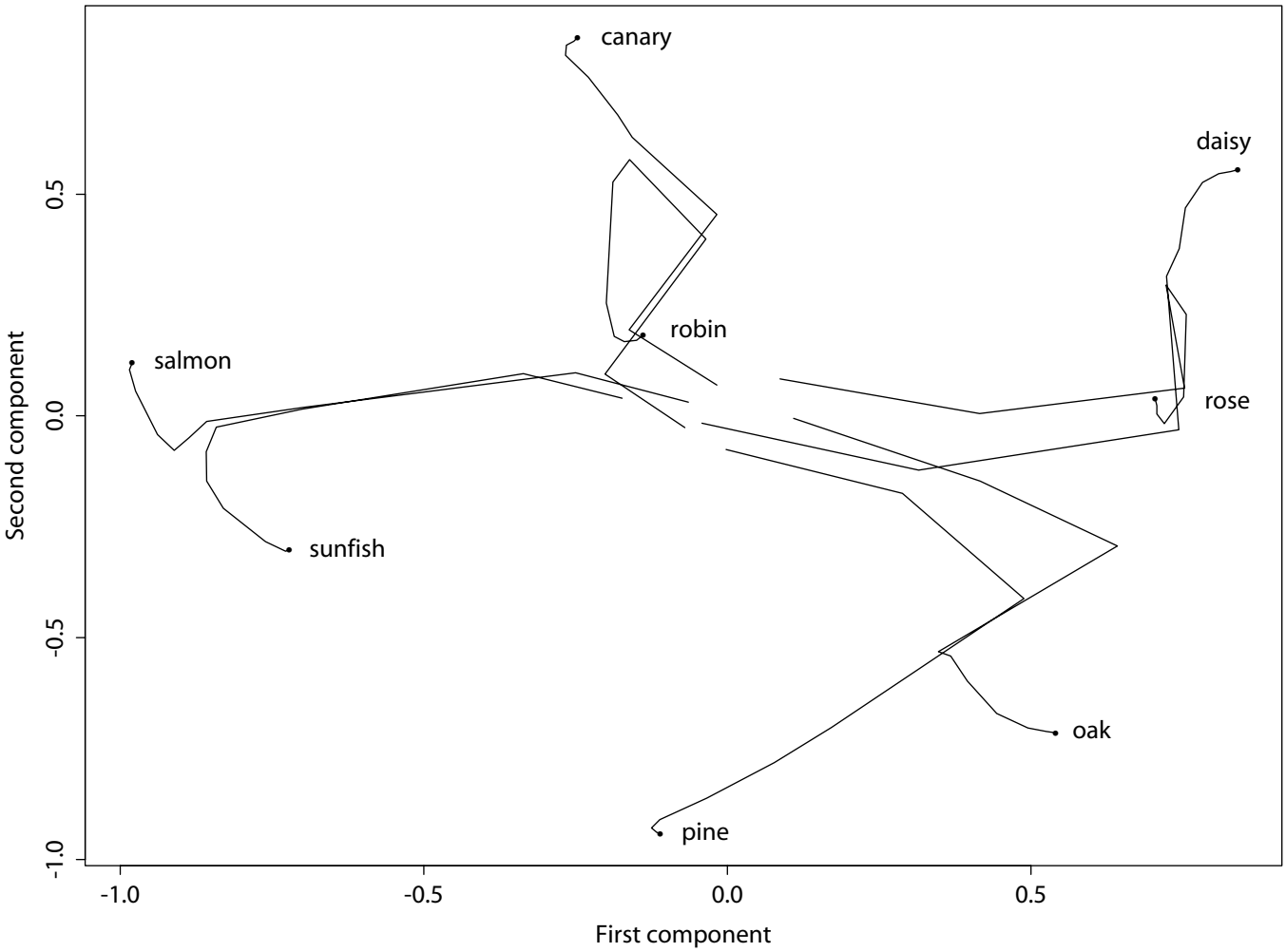
Figure 7. Panel A. The Farah-McClelland model. Pabel B: The Convergence theory of semantic memory. Unit pools shown in black were implemented in the models described by Rogers et al. (2004) and Lambon Ralph, Lowe and Rogers (in press).

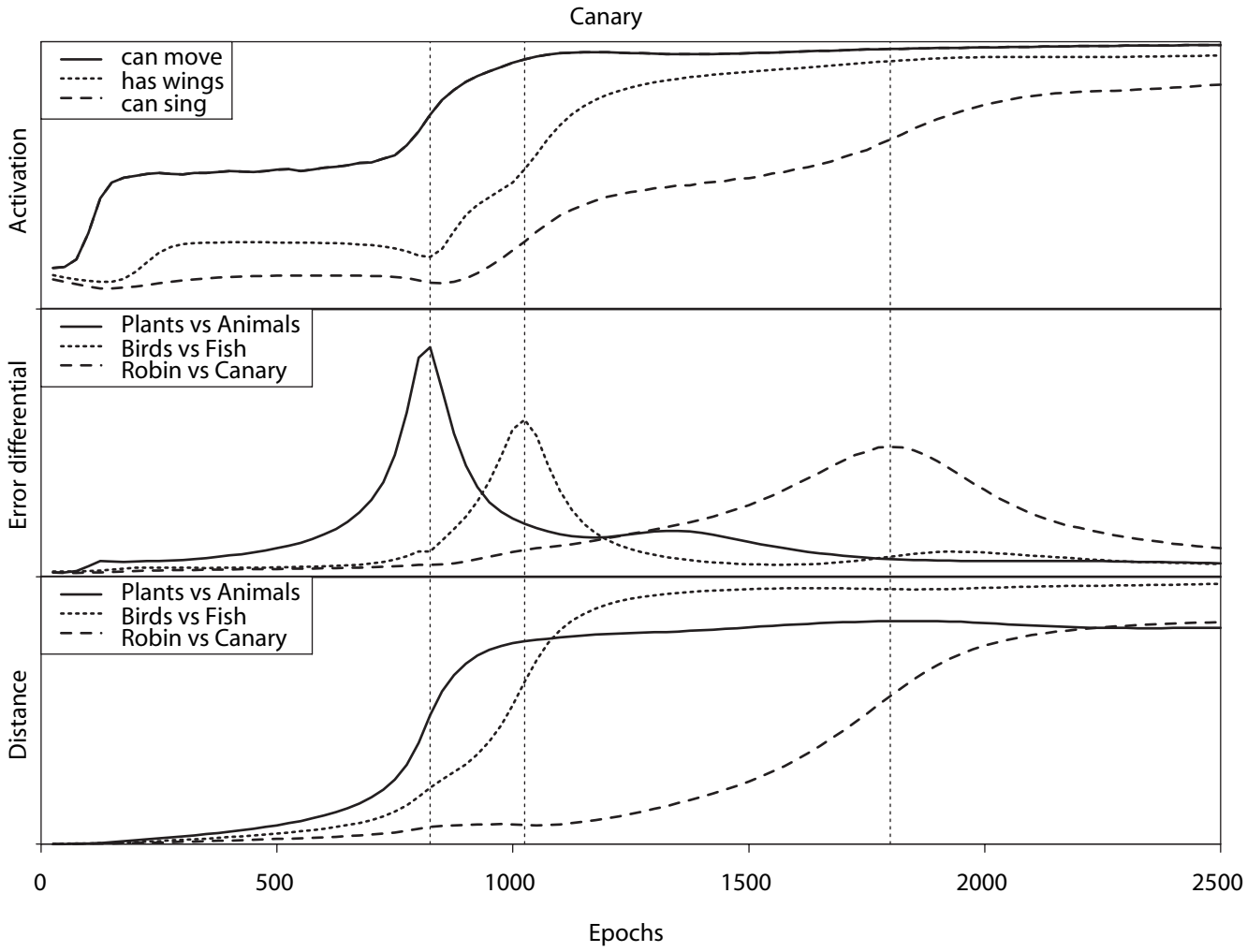
Notes

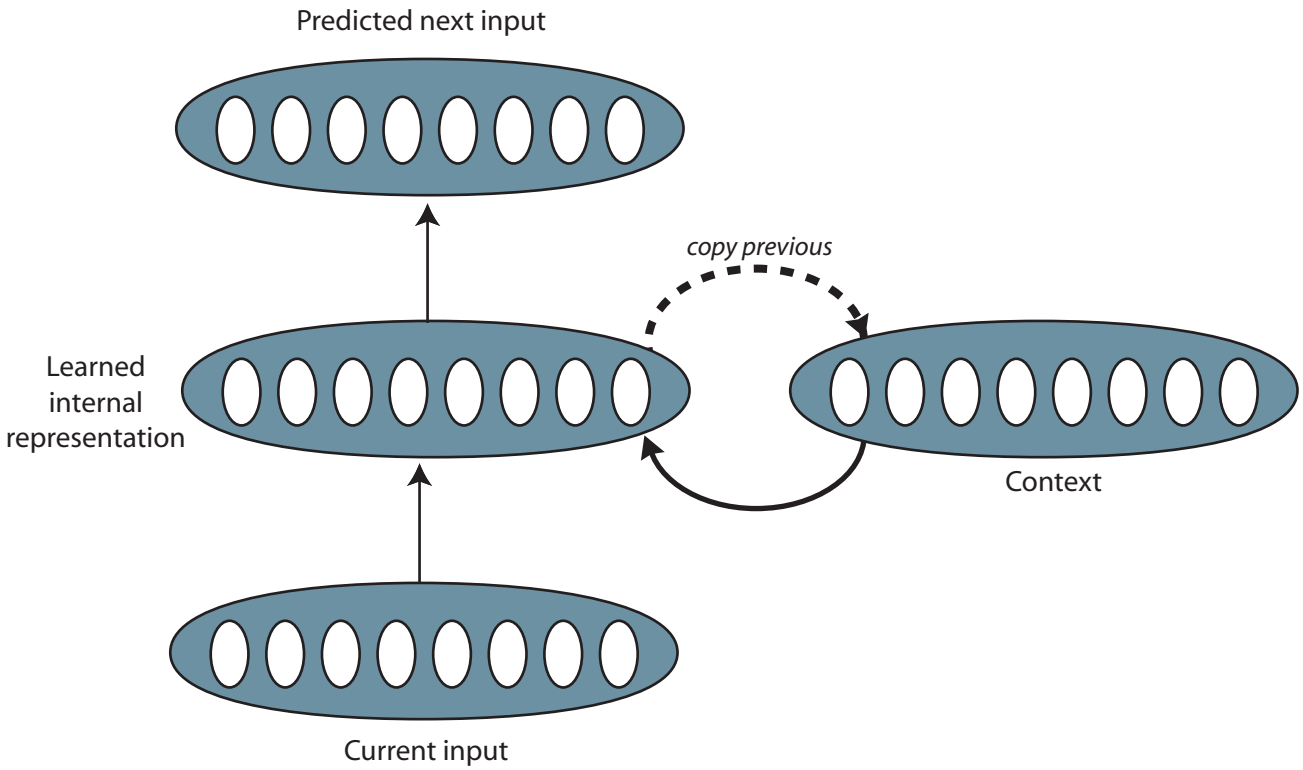
[1] In a more realistic model, the representation of a novel item would be achieved by recurrent connections projecting back from the attribute units toward the representation units; such a model is discussed in the final section of this chapter. To simulate this recurrent process in the feed-forward model shown in Figure 2, Rumelhart used a technique called backpropagation-to-activation: beginning with a neutral pattern of activation across Representation units, activation was propagated forward to the outputs. Error was computed on just the "bird" output unit, and the derivative of this error was calculated, without changing any weights, in a backward pass. The activations of the Representation units were then adjusted to reduce the error on the "bird" unit. That is, the model adapted its internal representations by changing activations on the Representation units until it found a pattern that strongly activated the "bird" output unit. This pattern thus constitutes a representation of the novel item given just the information that it is a bird.

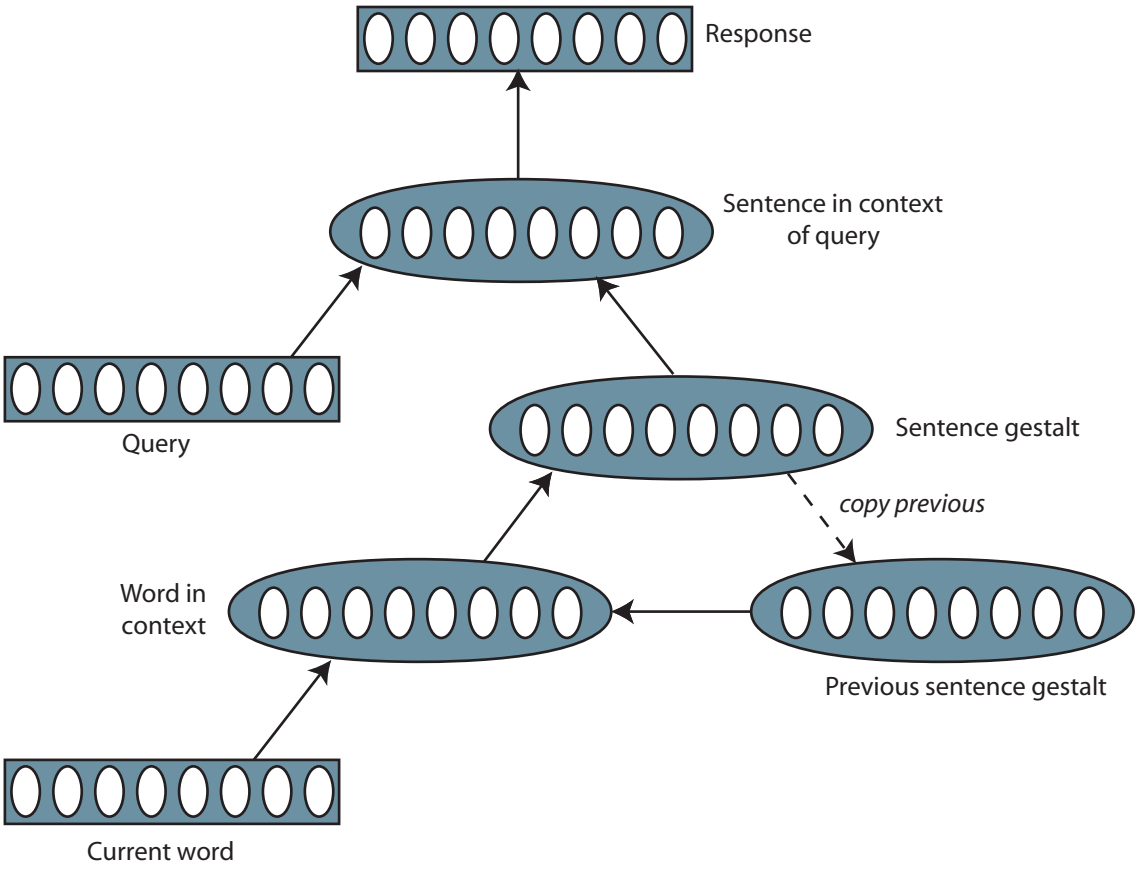

(e.g. elephant)
Role 2
Prop
Role 1
(e.g. Clyde)
Relation
(e.g. isa)

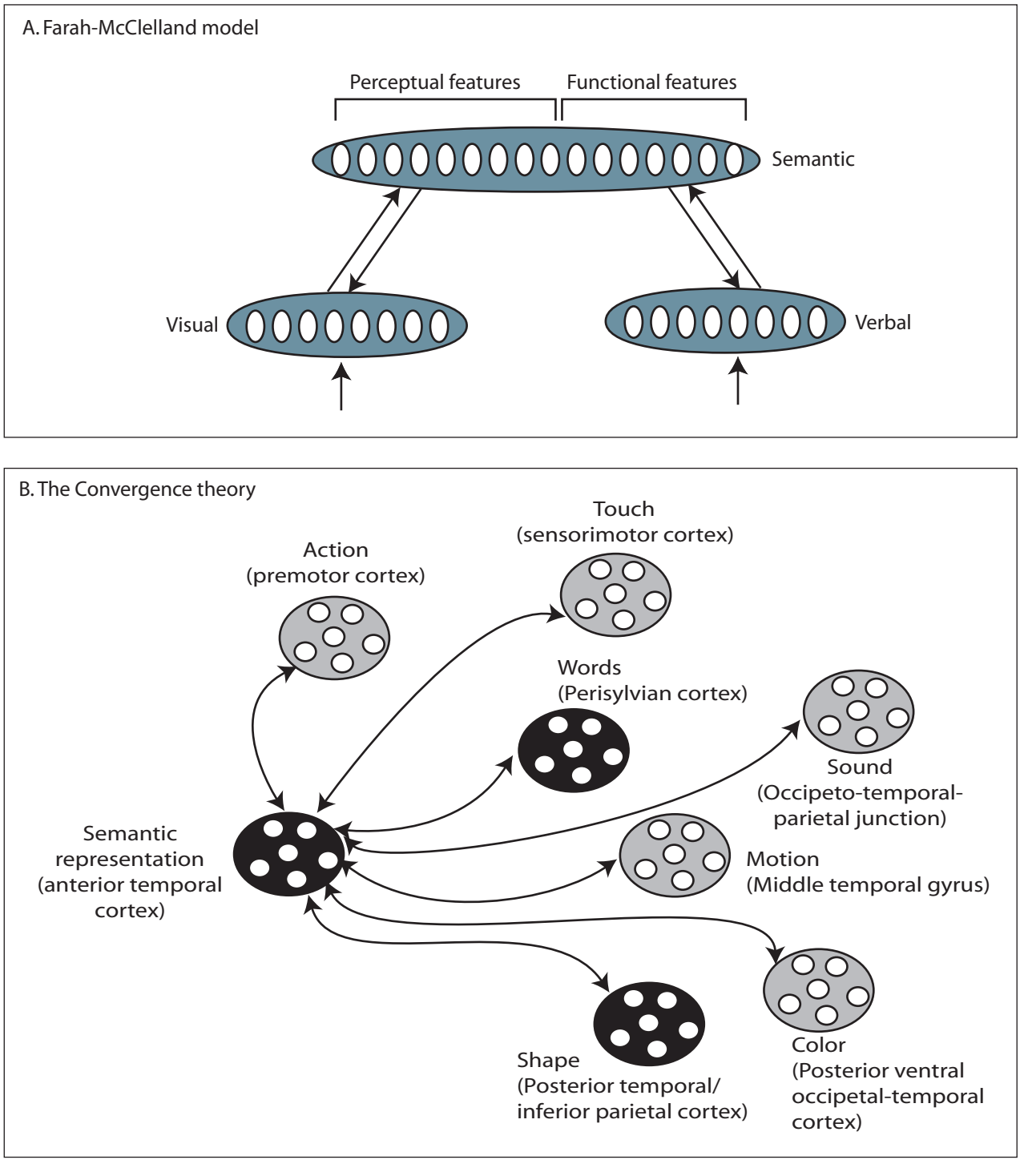

Representation
Hidden
pine
oak
rose
daisy
robin
canary
sunfish
salmon
Item
ISA
is
can
has
Relation
living thing
plant
animal
tree
flower
bird
flower
pine
oak
rose
daisy
robin
canary
sunfish
salmon
pretty
tall
living
green
red
yellow
grow
move
swim
fly
sing
bark
petals
wings
feathers
scales
gills
roots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


canary
daisy
robin
salmon
rose
sunfish
oak
pine
Second component
First component
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5


Canary
can move
has wings
can sing
Activation
Plants vs Animals
Birds vs Fish
Robin vs Canary
Error differential
Plants vs Animals
Birds vs Fish
Robin vs Canary
Distance
0
500
1000
1500
2000
2500
Epochs


Predicted next input
copy previous
Learned
internal
representation
Context
Current input


Response
Sentence in context
of query
Query
Sentence gestalt
copy previous
Word in
context
Previous sentence gestalt
Current word


A. Farah-McClelland model
Perceptual features
Functional features
Semantic
Visual
Verbal
B. The Convergence theory
Touch
(sensorimotor cortex)
Action
(premotor cortex)
Words
(Perisylvian cortex)
Sound
(Occipeto-temporal-
parietal junction)
Semantic
representation
(anterior temporal
cortex)
Motion
(Middle temporal gyrus)
Shape
(Posterior temporal/
inferior parietal cortex)
Color
(Posterior ventral
occipetal-temporal
cortex)