Computational Models of Semantic Memory

George S. Cree
University of Toronto, Scarborough

Blair C. Armstrong
University of Toronto / Carnegie Mellon University

IN PRESS

Cambridge Handbook of Psycholinguistics

Address correspondence to:

George S. Cree
Department of Psychology
University of Toronto, Scarborough
Toronto, ON, Canada, M1C 1A4
(416) 287-7439
george.cree@utoronto.ca
http://www.utsc.utoronto.ca/~gcree

Computational models of semantic memory provide controlled environments for studying the complex, interactive nature of knowledge representation and processing. They range in scope from dealing with the components of meaning activated immediately when a word is encountered, to theories of how our complete storehouse of general world knowledge may be represented and accessed. They typically outline the format in which knowledge is stored and/or the mechanisms and time course of how that knowledge becomes available as we read, listen, or think. In this chapter we outline the contributions of three classes of computational models which have had the greatest influence on our understanding of semantic memory: semantic network theory, parallel distributed processing (PDP) networks, and word co-occurrence models.

Scientists who study semantic memory are usually concerned with word meaning, and how the meanings of multiple words can be combined to understand longer text segments. They have also been concerned with how people acquire word meanings, how they use them to draw appropriate inferences, how they can efficiently store and search vast amounts of information, why some systems of categorization seem better and more natural than others, what components of meaning become active immediately when we encounter words, and why various kinds of brain damage can lead to specific patterns of loss of word meanings. Most research has been conducted using tasks that present simple statements (e.g., “A dog is an animal.”), or combinations of words in temporal sequence (e.g., dog -> cat), and asking people to make judgments about the stimuli. To help understand how people perform these tasks, researchers have designed computer models that embody the assumptions they theorize are important about how knowledge is stored and computed.

The advantages of computer models over descriptive theories are well known (Hintzman, 1991). Modeling forces researchers to be explicit about their assumptions and the components that make up their theories. Implemented models allow researchers to test predictions, and to derive new predictions, by running simulations. They also allow researchers to create ‘artificial patients’ by removing or altering components of the model and examining the effect on behavior. Implemented models can serve as existence proofs that a theory is logically coherent, and behaves as advertised.

The goal of this chapter is to summarize the important insights that have come from each of the three main approaches to modeling semantic memory. We will outline the important contributions of each, and point the interested reader to more detailed explanations when available.

Semantic Network Theory

Semantic networks are collections of nodes that are linked together by labeled relational links. Each node typically represents a single concept, and hence these models are referred to as having localist representation schemes. The meaning of a concept is represented through a set of pointers to other nodes. A goal of this type of modeling is to determine how to link up the nodes such that the resultant knowledge structures can be used to produce realistic semantic inferences. The implementations that have best stood the test of time are those of Ross Quillian (e.g., Quillian, 1962, 1967, 1968, 1969), who was concerned with understanding both natural language and memory. Quillian's early models worked by instantiating specially coded dictionary definitions into networks of nodes and examining how inferences could be drawn from the intersections of paths emanating from target nodes.

Collins and Quillian (1969) realized that if they included assumptions about efficiency of storage, and the length of time it should take to move between nodes, it would be possible to derive predictions about how humans retrieve information. They suggested that conceptual information was stored in a hierarchy, with more general concepts (e.g., animal) at the top, and more specific concepts (e.g., canary) at the bottom (see Figure 1). Concepts were defined in two ways: as a set of features held within each concept node, and in the set of pointers to other nodes. Properties of concepts were stored at the highest node in the hierarchy for which the property held true for all concepts below (e.g., <has wings>

was stored at the bird node, but not the animal or canary node), thus implementing a form of cognitive economy.

Predictions regarding the length of time it should take participants to verify statements could be generated directly from the model. These were tested most thoroughly through use of the sentence verification task in which participants were asked to verify the truth/falsity of simple sentences (e.g., “A canary is a bird.”). This task could be simulated in the Collins and Quillian (1969) framework by starting at a node, and searching both properties in nodes, and along relational links to other nodes, until the information necessary to evaluate the statement had been found. Collins & Quillian (1969) found, as predicted, that it took longer for people to verify statements that required longer searches (e.g., traveling two nodes) than statements that required shorter searches (e.g., traveling one node).

The Collins and Quillian (1969) framework also provided a mechanism through which information could be inherited. If the system needed to learn about a new concept, then a node for that concept could be attached at the appropriate level of the hierarchy, and the concept would automatically inherit all of the appropriate information about members of that category that were stored at higher nodes.

Initial behavioral evidence appeared to support both the ideas of hierarchical organization and inheritance through cognitive economy (e.g., Collins & Quillian, 1969), but hierarchical network theory did not hold up well to further investigation. There were a series of findings that convincingly demonstrated that the strength of relation between a concept and property, or concept and concept, were more important in determining verification latency than was distance in the hierarchy (e.g., Conrad, 1972). Furthermore, the model could not explain typicality effects, such as why people are faster to verify that a robin is a bird than that a chicken is a bird (Rosch & Mervis, 1975). Additionally, it was not clear how one would decide where in the taxonomy to store concepts that belonged to more than one category (e.g., knife), or when to add a new node for items that were similar, but not identical, to those already represented. These observations, along with others, were used to argue against semantic memory as a strict taxonomic hierarchy.

Spreading activation theory (Collins & Loftus, 1975) was proposed as an alternative, and was framed as a semantic network without hierarchical organization. It was used to account for a number of behavioral phenomena that posed problems for hierarchical network theory. But this power came at a cost, as it was ultimately determined that the model was too flexible, and could be used to account for just about any pattern of data (Johnson-Laird, Herrmann, & Chaffin, 1984). This flexibility came mainly from the fact that there were no constraints on which nodes could be connected to which other nodes, and more importantly, no manner for determining the strengths of weights between nodes.

Semantic network theory was the first major computational approach designed to investigate semantic memory. It succeeded in inspiring over a decade of behavioral research, and in promoting the need for future models to be able to account for inheritance and typicality effects. The modeling framework itself is quite simple, and provides a language with which one can easily discuss predictions. Yet despite this, the framework is unsatisfying, because the concessions needed to allow it to account for the known behavioral data leave the approach too unconstrained, and hence unviable as a research framework within which to study human cognition. A small number of theories were proposed as alternative explanations of the sentence verification data (e.g., McCloskey & Glucksberg, 1979; Smith, Shoben, & Rips, 1974), but none were proposed as grand theories of semantic memory on the scale of hierarchical network theory or spreading activation theory. Despite the limitations, semantic network style models are still under development, although more as a means to implement an efficient knowledge system than for generating novel predictions. The various instantiations of ACT, for example, include a spreading activation like semantic component (Anderson, 1983). A second example, although not exactly a semantic network, is CYC, a very large database of common sense assertions, linked together by relations, designed to produce coherent inferences (Lenat & Guha, 1990). CYC is being developed with the hopes of producing a common sense reasoning component for a full artificial intelligence system (see http://www.cyc.com/).

Parallel Distributed Processing Models

PDP models, also known as connectionist networks, have grown in popularity since the early 1980's as an alternative to semantic network theory. In PDP models concepts are typically represented as distributed patterns of activity across sets of representational units. The units often represent features of concepts (e.g., <has 4 legs>), but not necessarily nameable features. Units are organized into layers, and are connected by weighted links, which are usually set through use of either an error-correcting, or Hebbian, learning algorithm.

There are several important differences between PDP and semantic network models. First, a form of cognitive economy is implemented through the fact that multiple concepts can be represented as distributed patterns across the same set of units. Second, they offer mechanisms for determining weights on connections. Third, by combining the use of distributed patterns and discovering weight strengths algorithmically, similarity amongst items emerges as an implicit result. Finally, PDP models are said to degrade gracefully, in that when they are damaged (e.g., by removing weights) knowledge about concepts is lost gradually, and not in an all-or-none manner as in localist theories.

Hinton (1981) proposed one of the first PDP alternatives to semantic network theory. He was interested in understanding the rich microstructure of semantic representations, arguing that a single relational link in a semantic network would in reality be represented by multiple units and connections in a distributed system like the brain. He designed a model in which propositional information (e.g., the moose is brown) was represented as distributed patterns of activity across sets of processing units (e.g., *role 1*: moose, *relation*: colour, *role 2*: brown). These three banks of units interacted with one another through a set of hidden *prop* (proposition) units that encoded conjunctions of roles and relations (see

Figure 2). When a pattern of activation was enacted across a layer, the system was designed to settle to a stable state (i.e., represent a correct association) by updating its activation states over time.

PDP networks that update unit states over time and settle to stable states are known as connectionist attractor networks. Hinton noted that attractor networks are appropriate for simulating semantic memory because not all points in semantic space correspond to lexical concepts (Hinton & Shallice, 1991). Semantic space can be thought of as a multidimensional space in which each dimension corresponds to a unit, and a location corresponds to a pattern of activation across those units. To borrow Hinton's example, the point in semantic space exactly halfway between rhinoceros and unicorn is not likely to correspond to anything that exists in the real world, and certainly not anything for which we have a name. It therefore makes sense to have a system that moves towards a valid semantic representation, known as an attractor, when a word is presented. Regions surrounding attractors are called attractor basins, and once the network computes a semantic state that is within one of these basins, then as long as the input to the network does not change, the network will settle into the attractor. Hinton used the properties of the attractor network to show how a distributed system could exhibit pattern completion, cognitive economy, generalization, and property inheritance.

The model was capable of completing partial propositions. One way to understand how it accomplished this is to think of the *relation* layer as encoding contexts within which specific roles occur. A context would then interact with a pattern of activity in the *role 1* units to produce an appropriate pattern across the *role 2* units. An efficient way to represent the regularities that exist in the role-relation contexts is to determine which specific units are active together across many patterns, and to assign a hidden *prop* unit to represent those conjunctions of unit activations. Subsequently, whenever a role and relation are presented together that activate that *prop* unit, the *prop* unit would partially activate units across the *role 2* layer, which would ultimately settle to one of the concepts stored in the system, completing the pattern.

A second important property was that multiple propositions could be stored in the same set of weights. This provided a second form of cognitive economy, again different from that proposed by Collins and Quillian (1969).

Third, the model could generalize in an intelligent manner based on the similarity of representations at the role layers. Imagine, for example, that a novel concept (e.g., deer) was instantiated across the *role 1* units, and it had a representation similar to another concept stored in the network (e.g., moose). If it was activated along with an appropriate relation pattern (e.g., colour) the network was capable of completing the proposition with an appropriate *role 2* due to the fact that the pattern instantiated across the *role 1* units would sufficiently activate suitable prop units.

Finally, property inheritance, an important feature built into Collins and Quillian's (1969) semantic network theory, arises naturally in Hinton's (1981) model, assuming that appropriate representations are chosen. Collins and Quillian are said to have 'built in' property inheritance because they did not provide an account of why or how knowledge would be stored at only the highest concept node for which the feature applied. Hinton's representational scheme provided an elegant alternative. Membership in a superordinate category could be represented as the subset of role units common to each of the relevant subordinate concepts. That subset of units could then activate *prop* units fitting for the category. A new concept that shared that subset of features would automatically activate *prop* representations that were true of the category, and hence be able to provide correct inferences through pattern completion in the *role 2* layer.

Hinton's (1981) work did not, however, solve satisfactorily the problem of how to determine the strengths of the weights between units (Hinton set his weights by hand). By the mid-eighties, methods had been discovered for training sets of weights in networks that included hidden layers, sometimes referred to as deep networks. One of the most effective was backpropagation (Rumelhart, Hinton, & Williams, 1986).

Hinton (1986) showed that a system trained with backpropagation could discover implicit semantic features and represent them across a hidden layer by picking up on the regularities that existed in the inputs and the outputs of the training patterns. Specifically, he demonstrated how propositions about personal relationships, typically represented in family trees, could be coded efficiently in a distributed system. For example, a hidden unit could come to code for an aspect of a relation (represented in the *relation* units) corresponding to the idea that if the person in *role 1* is young, then the person that appears in *role 2* must also be young. This hidden unit could then take part in representing this relation across multiple propositions, not just a single instance. This was an advance over Hinton (1981) because the implicit features were not built in by the experimenter, but rather were an emergent property of the use of the learning algorithm.

David Rumelhart extended Hinton's models with a demonstration that a PDP system trained with backpropagation, and exposed to a set of concepts and features taken from an independent source, would exhibit properties consistent with human semantic memory (Rumelhart, 1990; Rumelhart & Todd, 1993). Rumelhart used the exact set of concepts, features, and relations that were depicted by Collins and Quillian (1969) in the figure that illustrated their hierarchical semantic network. Rumelhart's model was a feedforward system (i.e., not an attractor network), in which activation flowed in one direction, from inputs to outputs (see Figure 3). The model had 5 layers of units, with two input layers and one output layer. Concept name input units (e.g., canary) passed activation to representation units, and representation units passed activation to relational units, which also accepted inputs from the other input bank, known as relation units (e.g., can, is, etc.). The relational units ultimately passed activation to a set of output units that coded for features of each concept (e.g., <fly>, <sing>, etc.). Localist representations were used for both concepts and relations. Output representations were distributed patterns of features.

Rumelhart was interested primarily in the representations that would develop across the hidden units through repeated exposure to items in the training set. By using a localist representation at the input layer he had stripped out of the input all of the similarity among concepts (e.g., canary was as similar to salmon as it was to robin). Therefore, if similarities were encoded at the hidden layers during training it must be because the system was making use of similarities that existed at the output layer, and encoding them in the representation layer. This was interesting for two reasons. First, the network would be developing its own internal representations as a means of solving the mapping problem. Second, the internal representations could be examined using statistical analyses to determine which aspects of semantic knowledge were being encoded. This was an advance over Hinton's work because it used an independent, yet penetrable training set in which the implicit semantic features encoded at the hidden representation layer could be interpreted in terms of concepts, categories, and features of common objects.

Rumelhart used the model to illustrate three points, the most novel of which was that the patterns of activity developed at the representation unit hidden layer captured the featural similarities of concepts that existed at the output layer. The representation of canary, for example, was more similar to the representation of robin than it was to the representation of any of the fish, flowers, or trees. He went on to demonstrate that the activity in the representation units could be interpreted as coding for the major conceptual distinctions that existed in the training set (e.g., in one simulation representation unit 3 might be active consistently for animals but not plants). He also showed, like Hinton (1981), that the system displayed property inheritance, and that it also displayed the 'cancellation principle,' or in other words, that it could learn that there are members of a category that do not have a property that is common to most other members of the category (e.g., an ostrich is a bird but does not fly).

Hinton and Rumelhart demonstrated that PDP systems could produce patterns of behavior consistent with the human conceptual system. The models were able to generalize appropriately, demonstrate property inheritance, and develop internal representations that reflected the structure inherent in the training patterns. Although these models did not provide a full account of the behavioral phenomena associated with semantic memory, and had not been used to closely simulate any behavioral tasks, they were viewed as a promising new approach to studying semantic representation and processing.

Rogers and McClelland (2004) have significantly extended the Rumelhart framework to account for a large body of findings. An important explanatory principle in their work is that of coherent covariation of properties. This refers to the fact that some sets of features tend to co-occur together more often than others, and that these sets tend to occur more often within than across category boundaries. Many concepts, for example, <have wings>, <have beaks>, and <lay eggs>. Furthermore, those that <have wings> tend not to be <used for carpentry>. Learning in a PDP network can benefit from these regularities because a set of weight changes that benefits one concept will benefit all of the other concepts that share that same set of features. This principle is important in explaining why children first learn to differentiate concepts at the superordinate level yet prefer to name at the basic-level, why some categories are more coherent than others, how different kinds of properties can become more central to one category than another, and why causal properties appear to be more central to determining category membership than others. We will consider one example.

Theory-theorists propose that knowledge is structured around naïve, domain-specific, causal theories (e.g., Murphy & Medin, 1985). An issue discussed within this domain of study is why some concepts seem to be better and more coherent (e.g., birds), than others (e.g., large blue things found outdoors). Rogers and McClelland suggest that this can be explained, at least partially, through the fact that causal structure, as it exists in the world, leads to the coherent covariation of object properties in the environment. For example, having bird DNA tends to cause a creature to have wings, a beak, and to be able to fly. Coherent covariation of such sets of properties across objects make these objects easier to learn due to mutually beneficial weight changes across concepts, and when combined with the fact that we need to refer to these clusters of objects more often than to others (e.g., clusters of large blue things found outdoors), we can see why we will tend to find the category of birds to be a more satisfying, natural category.

Rogers and McClelland (2004) make an ambitious attempt to provide a mechanistic account of several other phenomena studied by theory-theorists. The issues are complex, and are beyond the scope of discussion in this chapter, but if one is interested in reading an excellent summary of the issues, with pointers to important directions that future work should address, then this book is highly recommended (or see McClelland & Rogers, 2003, for a brief overview).

Dynamics of Meaning Computation

Semantic computation unfolds over time. When we read or hear a word, different components of the meaning become active at different rates over the first several hundred milliseconds. Feedforward connectionist networks are not well suited for studying these computations because activation is computed across entire layers in a single sweep. Attractor networks, however, in which units update their states continuously based on both their prior states and input from other units, are well suited for this work. The networks usually have at least one set of recurrent connections in which there can be feedback activation from top-down influences (see, e.g., Figure 4). These models have been used primarily to study semantic priming, feature verification, and semantic impairments in cases of brain damage and disease.

Semantic priming refers to the finding that participants are faster to respond to a target word (e.g., eagle) if it is preceded by a semantically related word (e.g., hawk) than by an unrelated word (e.g., hook; see McNamara, 2005, for a highly readable review). Priming was originally thought to reveal the structure of semantic memory, on the assumption that if priming occurred, it was evidence that two concept nodes must be closely linked in a semantic network. PDP simulations of priming have offered alternative interpretations of why and how priming arises (Masson, 1991, 1995; Plaut, 1995; Sharkey, 1989). We will consider one example.

Plaut (1995) offered a clear demonstration of how two different kinds of priming effects could be instantiated in a single PDP system. Semantic similarity priming is due to similarity in meaning of two concepts. One way of formalizing similarity is to talk in terms of shared features: the more shared features, the stronger the priming. Associative priming, on the other hand, has traditionally been discussed as priming that occurs due to associative relations between two concepts, as indexed through association norms, in which a large number of participants are asked to respond to a word with the first word that comes to mind (e.g., ham -> sandwich). Association almost certainly reflects semantic relations, including featural similarity, and so it is difficult to nail down exactly what association is, besides a grab bag of relations. One important type of relation that we can agree is included, that is relatively distinct from featural similarity, is co-occurrence in time and/or space. For example the words ‘tennis’ and ‘elbow’ share few, if any, semantic features, but they do tend to occur together in language more often than chance. We can refer to this as word co-occurrence, and it has been demonstrated to produce reliable priming effects. Plaut (1995) discussed two different mechanisms in PDP attractor networks that may give rise to these two different kinds of priming.

Priming due to semantic overlap arises naturally in PDP networks because concepts that share features are typically represented by similar patterns of distributed activity. Computing the meaning of the prime will therefore put the network into a state in which both the prime and target are simultaneously active. Hence the system will be faster to process the meaning of the target when it is presented at the input layer after the prime, allowing for faster responding.

Plaut (1995) suggested that associative (or word co-occurrence) priming could be realized through the ability of the system to move from one attractor state to another. He manipulated the ability of the network to perform these mappings by varying the likelihood that one word followed another during training, and importantly, not resetting the activation values between training trials. This forced the network to develop weights that allowed it to move easily between attractors for words that co-occurred frequently during training, producing a priming effect when compared to non co-occurring words.

Several other models of priming have been reported, providing further insight into factors such as the role of similarity and correlational strength in predicting priming effects (McRae, de Sa, & Siedenberg, 1997), the degree of similarity required to observe priming (Cree, McRae, & McNorgan, 1999), long term semantic priming (Becker, Moscovitch, Behrmann, & Joordens, 1997), and individual and developmental differences in priming (Plaut & Booth, 2000). An important area for future work is to unpack the different relations that drive semantic and associative priming through carefully designed experiments and simulations that demonstrate the proposed mechanism(s) at work.

Feature Verification

The feature verification task involves asking participants to verify whether or not a feature is reasonably true of a concept (e.g., cat – <meows>). It can be used to reveal the time course of activation of features of a concept. McRae and colleagues have used it to examine how statistical concept-feature and feature-feature relations influence semantic computation, and have simulated the findings using a two layer attractor network framework in which word forms map to semantic features (see McRae, 2004, for a review).

Two aspects of the approach taken by McRae and colleagues deserve note. First, the semantic representations are empirically derived, being taken from a set of semantic feature production norms in which participants were asked to list the features they thought were part of a concept (McRae, Cree, Seidenberg, & McNorgan, 2005). This has the benefits of reducing degrees of freedom in modeling, allows statistics to be computed from the norms regarding the occurrence of features across concepts and categories, and it permits the use of exactly the same items in both experiments and simulations. Second, because they have been interested in how attractor networks encode feature-feature statistics, they have been less concerned with the internal representations developed across hidden layers. The modeling framework used in their simulations has therefore typically involved a set of word form units that maps to a set of fully interconnected semantic units, with perhaps a set of hidden units connected to the semantic units that can encode higher-order correlations of features across the semantic layer (sometimes called a clean-up layer). This framework allows for feature-feature relations to be directly encoded in the weights. Using this framework, McRae and colleagues have uncovered two relationships that govern speeded computation of semantic meaning in both humans and the model: correlations among features, and the distinctiveness of features.

Features are said to be correlated if they tend to occur together across concepts. Strength of correlation can be measured by computing how often feature pairs occur together across all the concepts in a set of semantic feature production norms. In a feature verification task, McRae, Cree, Westmacott, and de Sa (1999) found that features with high intercorrelational strength, a measure of how correlated a target feature is with the other features in the concept, were verified faster than control features. The accompanying attractor network simulation revealed the same effect. Analysis of the network showed that this was because the high intercorrelational strength features received strong support from other correlated features, with strong connection weights, that were also activated by the concept's word form, and thus reached higher states of activation than did control features during the early stages of processing.

Distinguishing features have been defined as features that are true of only one, or at most a few, concepts (e.g., cow - <moos>). Cree, McNorgan, and McRae (2006) demonstrated that distinguishing features are verified faster than control features by human participants, and provided a simulation that reflected this pattern of responding. Analysis of the network revealed that distinguishing features reached higher levels of activation faster than did control features because strong weights had formed during training between the word form units that represented each concept name and the distinguishing features of that concept. This makes sense, given that the goal of the network during training was to learn to settle into the correct attractor as fast as possible. Distinctive features are an excellent source of information about which attractor basin to enter, because, by definition, they occur in only a few concepts. The network can then fully settle to the correct representation by filling in the rest of the semantic representation around the distinguishing features.

The attractor network framework used by McRae and colleagues has been used more recently to explore how superordinate concepts are learned, and to simulate speeded tasks that include superordinate concepts as items. O'Connor, Cree, and McRae (2007) proposed that superordinate concepts are learned, at least partially, over many learning trials, with each trial instantiating a mapping from the word form for a superordinate concept to the semantic features of a specific category member (e.g., animal -> moose). They found that when these learning trials are interspersed with basic-level learning trials (e.g., moose -> moose) the system is able to extract the regularities with which features occur across the various category members to form superordinate attractor states. Once trained, the model produces superordinate representations that intuitively make sense (e.g., tool: <used for carpentry>, <made of metal>, <has a handle>, etc.) that capture general information about the category, and yet are not the same as any individual concept. The network has been used successfully to simulate a typicality rating task in which participants are asked to rate how good a concept is as an example of a category. The model outperformed family resemblance, the gold standard for simulating typicality ratings, on a large number of the categories on which the network was trained.

Finally, the superordinate network has been used to provide insight into some findings that are counterintuitive in terms of both semantic network theory and the standard PDP explanation of semantic priming. There have been several reports of priming studies in which high and low typicality targets produce equivalent priming effects when preceded by a superordinate word as prime (e.g., Schwanenflugel & Rey, 1986). According to the standard PDP explanation of semantic similarity priming, priming should be larger for high typicality items because they should share more features with the superordinate term. O'Connor et al. (2007) have replicated this behavioral effect, and explained why this occurs in terms of the ability of the system to move out of a superordinate vs. basic-level attractor. This serves as an example of how the dynamics of processing in an attractor network can explain patterns of human performance that do not make sense in terms of traditional semantic network modeling frameworks.

Constraints on Structure: Semantic Impairments

The patterns of semantic impairment observed in patients with brain damage and disease have served as constraints on the structure of computational models of semantic memory. Several distinct patterns of impairment have been reported. Semantic dementia, for example, tends to lead to a general impairment that first produces an inability to recall specific facts about concepts, and later influences general knowledge about concepts. Alzheimer's dementia, on the other hand, tends to produce a general semantic deficit as the disease progresses, but with a more severe impairment for living thing knowledge during the moderate to severe stages. Other complex patterns of category-specific impairments have been observed as a result of herpes simplex encephalitis, closed head injury, and stroke (see Capitani, Laiacona, Mahon, & Caramazza, 2003, for a review). Computational models offer a powerful tool for creating 'artificial patients' in which researchers can induce damage that mimics either focal damage, or diffuse brain damage, due to various diseases and injuries.

Semantic network theory was the first framework used to try to interpret the patterns of impairment (Warrington, 1975). The progression from loss of specific to general knowledge in semantic dementia can be thought of as loss of the bottom levels of a hierarchical semantic network. Category-specific semantic deficits can be interpreted as loss of branches in a hierarchical network. These explanations are ultimately unsatisfying, however, because they fail to explain why the bottom levels should be more susceptible to damage than others, or why the reported patterns of impairment do not always follow category boundaries. Thus researchers have turned to other modeling frameworks.

Tim Rogers and colleagues have used PDP networks to provide detailed accounts of what might be occurring in cases of semantic dementia (Rogers et al., 2004; Rogers & McClelland, 2004). Rogers et al. (2004) used an attractor network that mapped from verbal and visual input/output layers through a set of hidden units, labeled semantic units, that functioned much like the *prop* units in Hinton's (1981) model. When damaged, the system mimicked patient performance in confrontation naming, word and picture sorting, word to picture matching, and drawing tasks. Analyses of the network revealed that specific information was lost before general information due to the nature in which regularities across the visual and verbal patterns were stored in the semantic hidden units. Features that occurred together across numerous concepts tended to be represented in larger, more contiguous regions of the semantic space defined by those units than were distinguishing features. Therefore distinguishing features were more likely to be affected by damage, causing the system to lose the ability to discriminate among similar concepts. Another way of thinking of this is in terms of the boundaries of the attractor basins for each concept. Damage shifts the boundaries, due to loss of distinguishing dimensions, such that a pattern of activation that would have once led into the attractor basin of one concept may subsequently lead into the attractor basin of a similar concept.

Hinton and Shallice (1991) were the first to describe semantic impairments in terms of shifts in the boundaries of attractor basins in their work on acquired dyslexia. Some dyslexics, when asked to read aloud a printed word (e.g., peach), will produce a word that is related in meaning (e.g., apricot). Interestingly, many of the errors reveal a visual component, and the combination of mixed visual and semantic errors (e.g., rat – cat) is much higher than would be expected from the prevalence of independent visual or semantic errors alone. They noted that two aspects of attractor networks are important for explaining these effects. First, even if the input to the system is noisy, the system will still move towards one of its known states. Second, attractor networks are able to map similar inputs to similar regions of semantic space, something that feedforward connectionist networks do naturally, yet still settle to distant points in semantic space, as the attractor network settles into a basin of attraction. The production of 'apricot' when the system had read 'peach' was an example of the system settling into a nearby attractor due to a shift in the boundary resulting from damage.

Counter to intuition, Hinton and Shallice (1991) learned that it didn't matter which set of weights they damaged, be they early in the system or late, the network still produced visual, semantic, and mixed errors. This was not what they had expected to observe, and is not what would be predicted from traditional models used in cognitive neuropsychology, which would predict that damage to weights between the input and semantic layer would be the most likely to produce visual errors. They discovered that this was because damage anywhere in the system would lead to shifts in the boundaries of the attractors, and so based on the set of concepts that were represented close in the semantic space, damage anywhere could lead to errors that would be classified as visual, semantic, or mixed, depending on which incorrect attractor the system now mapped into. Plaut and Shallice (1993) provided an extended investigation of these issues.

Attractor networks have also been applied to interpreting category-specific semantic deficits. Four factors have been found to be important in explaining the observed patterns of deficits. The first is that different modalities of information may be differentially important for discriminating among objects from various categories (Warrington & Shallice, 1984, Warrington & McCarthy, 1987).

Functional information is important for discriminating among tools, for example, and hence damage to regions that code functional information could differentially disrupt the ability to discriminate tools. The second is that living and nonliving things differ in the ratio of perceptual to functional features (Farah & McClelland, 1991). Farah and McClelland instantiated these principles in a PDP network and demonstrated how they can give rise to category-specific semantic deficits in a system in which living and nonliving thing concepts are both represented over a set of semantic processing units with specialization by modality. Hence they demonstrated that a category-level behavioral deficit could emerge from a system in which knowledge was stored by modality of knowledge, not category.

The most counterintuitive simulation reported by Farah and McClelland (1991) was one in which damage to perceptual properties gave rise to an inability to activate functional information for living things. Analysis of the model revealed that this occurred because activation of the functional features was reliant on the mass of activation they got from perceptual properties, and thus when the perceptual properties were impaired, there was not enough activation in the system to push the functional features above threshold. This finding ran counter to the (still) generally held prediction that if damage to perceptual features was the cause of living thing deficits, then patients with living things deficits should still be able to answer questions about the functional properties of living things.

The second two factors were highlighted by Devlin, Gonnerman, Anderson, and Seidenberg (1998) in a model they used to explain how severity of impairment can interact with the structure of the semantic system to produce different patterns of impairment as damage progresses. Devlin et al. noted that concepts from different domains vary in terms of the average degree of correlation among features, with living things features tending to be more correlated. They also differ in terms of proportion of distinctive features, with nonliving things tending to have more informative features. As a result, when mild damage impairs the ability to activate features, knowledge about nonliving things will be most likely to be lost, because there are relatively more distinctive features in the representations of nonliving things, and because the correlations among shared features of living things will mutually reinforce one another, and preserve their ability to become activated. As the damage progresses, more of the correlated features will become impaired, and they will eventually lose the ability, en masse, to reinforce one another. Because correlated features tend to be the ones that help discriminate living things, there will be a shift from a trend for nonliving things deficits to clear deficits for living things. Subsequent patient data has confirmed the living things deficit in moderate to severe AD, but not the predicted initial nonliving things impairment in mild AD. It remains to be seen whether models that use more realistic semantic representations would produce the same predictions for mild AD.

Recent theories and simulations have focused on the role of topographic representation schemes in explaining patterns of deficits (Jankowicz, Becker, & Howell, 2003; Miikkulainen, 1997; Vinson, Vigliocco, Cappa, & Siri, 2003). Plaut (2002), for example, has brought attention to the idea that the brain's natural tendency to prefer short over long neural connections can lead the system to develop graded topographic, modality specific knowledge representations when mapping between visual and tactile inputs to verbal and action-based outputs. Simmons and Barsalou (2003) have presented a potential modeling framework based on Damasio's (1989) idea of topographically organized convergence zones that appears to explain many of the trends in impairment observed in cases of category-specific semantic deficits, but their theory remains to be implemented. In contrast, Tyler and colleagues have argued for a single semantic system that has internal structure based on patterns of shared and distinctive properties across concepts and categories, and have implemented a number of their claims in PDP networks (Durrant-Peatfield, Tyler, Moss, & Levy, 1997; Greer et al., 2001).

Despite success in explaining aspects of patient performance, no single model has yet been able to explain all of the patterns of impairment observed in the various cases of brain damage and disease that have been reported. PDP models have clearly provided explanations of how these deficits might arise that go far beyond what was possible to explain with traditional semantic network theories. All of these explanations appeal to the regularities with which specific features pattern across concepts and categories. It is therefore imperative that researchers derive large scale, realistic representations for each input and output modality if they hope to be able to simulate the complete spectrum of deficits that have been reported. We turn next to state of the art methods that are being used to derive such representations.

Deriving Representations and Exploring Structure: Word Co-occurrence Models

A major problem in modeling semantic memory has been how best to derive realistic semantic representations. Solutions have involved using hand crafted features (Hinton, 1981), randomly generated features (e.g., Plaut, 1995), and semantic feature production norms (McRae et al., 2005). There are positives and negatives with each approach, relating to how many concepts are to be included, how realistic the representations need to be in structure both at the level of an individual concept, and in terms of how concepts cluster into different categories, the word classes for which representations are required (e.g., noun, verb, adjective, etc), and the resources required/available to derive the representations (see Harm, 1998, for extended discussion). As a result researchers have been searching for methods that will produce representations that are all encompassing, richly structured, realistic, and can be generated fairly effortlessly via simple learning rules applied to rich data sources.

Two solutions to the problem of generating large scale representations have emerged. The first is to mine relational information from a large lexical database such as WordNet (Fellbaum, 1998; see Harm, 1998, for a computational implementation of feature mining). The second, which has generated much more research interest, has been to mine large text corpora for information about how words tend to co-occur around one another. The resultant models have become known as word co-occurrence models. The two most prevalent are Latent Semantic Analysis (LSA; Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990; Landauer & Dumais, 1997), and the Hyperspace Analogue to Language (HAL; Lund & Burgess, 1996), but there are several new comers, including Correlated Occurrence

Analogue to Lexical Semantics (COALS; Rohde, Gonnerman, & Plaut, 2007), Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007), and the Topic model (Griffiths, Steyvers, & Tenenbaum, 2007), that seek to improve on early methods.

LSA learns word meaning from a text corpus by examining the patterns of word co-occurrence across discrete passages of text (e.g., segmented documents). The first stage of LSA consists of constructing a global co-occurrence matrix in which each row represents a word, each column represents a document, and each cell represents the frequency with which a given word occurs in a given document. This procedure is illustrated in Table 1.

The matrix produced during the first stage of LSA is usually subjected to a normalization procedure designed to control for word frequency (see Landauer & Dumais, 1997, for details) and is compressed using singular value decomposition (SVD), a process which serves to emphasize the contexts in which words do and do not occur, and thus a specific word's meaning. Conceptually, this process can be likened to a factor analysis, in which a large number of items are reduced to a small number of latent factors. Consequently, words which may not have been strongly related in the text itself could share a common latent factor, and thus be related to one another. It should be noted however that the resulting dimensions of each word vector, although related to the original vectors, no longer directly correspond to any of the words from the corpus on which the model was trained.

Given the fact that LSA focuses on a word's co-occurrence within a general context (e.g., a document), LSA is particularly well suited to capture the general contextual associations and similarities which exist between words (e.g., bread and butter). However, LSA overlooks the importance of positional information in deriving word meaning (e.g., butter is spread on bread, bread is not spread on butter), and thus fails to capture the positional similarities that exist between words.

LSA has been used to model a wide variety of tasks, including: improving a child's ability to summarize via LSA-based feedback (Kintsch, Steinhart, Stahl, & LSA Research Group, 2000), assessing the importance of word order in the meaning of a document (Landauer, Laham, Rehder, & Schreiner, 1997), metaphor comprehension (Kintsch, 2000; Kintsch & Bowles, 2002), and modeling performance on the Test of English as a Foreign Language (TOEFL). Additional information and publications related to LSA can be found at http://lsa.colorado.edu/.

Representation vectors in the HAL model are built from information about the proximal co-occurrence of words within a large body of text. In HAL, the initial co-occurrence matrix is defined such that there is a row and a column for each word in the corpus. A moving window, usually 10 words in size, is passed over the entire text corpus. As the window passes across the text, information is computed regarding the distance of other words within the moving window, relative to the target word at the center of the window. A context word immediately beside the target word receives a value of 10, whereas if there are three words between the context word and the target word, the context word receives a 6. Each row denotes the co-occurrence of a context word before the target word, and each column denotes the co-occurrence of a context word after the target word. This procedure is illustrated in Table 2.

Once the entire corpus has been processed, meaning vectors are extracted by concatenating the row and column for a given word. The first half of this vector then represents the frequency with which context words ‘x’ tend to occur preceding ‘y’ in the corpus, whereas the row vectors represent the frequencies with which the context words ‘y’ tend to follow ‘x’ in the corpus. Thus, the vector representations of each word in the HAL model preserve some positional information. To reduce the size of the vectors, which at this stage of processing contain two dimensions for each word encountered in the corpus, the variance of each word-vector is computed, and only the word dimensions corresponding to the word vectors with the greatest variance are conserved. Trimming words based on the overall variance in their meaning vectors makes sense because the words with the largest variance contribute the most to the calculation of the overall similarity between vectors.

As with LSA, the similarity of the vectors generated by the HAL model can be compared to one another by calculating the distance between vectors in the semantic space. Given that these word vectors represent the positional context within which each word occurs, HAL is well suited to capturing the positional similarities between words. For example, pen and pencil will tend to have similar word vectors because these words tend to occur in the same position surrounded by the same set of context words. However, in capturing positional information, the HAL model is largely insensitive to the types of contextual information to which LSA is sensitive (e.g., the strong association between bread and butter, despite the fact that bread and butter represent very different objects).

The HAL model has been used to successfully account for a variety of phenomena, including: resolving semantic ambiguity (Burgess, 2001), modeling verb morphology (Burgess & Lund, 1997a), representing abstract words and concrete words (Audet & Burgess, 1999), emotional connotations (Burgess & Lund, 1997b), and how word meaning is learned (Li, Burgess, & Lund, 2000). Additional information related to HAL can be found at: http://hal.ucr.edu/.

COALS is one of the more recent co-occurrence models, and represents an attempt at improving HAL both via theoretically motivated improvements to the algorithmic procedures used to derive a word's representation, and in the case of the COALS-SVD model, via the use of SVD to reduce the dimensionality of the meaning vectors, and emphasize latent information in word representations as in LSA.

The process of creating the co-occurrence matrix for COALS is much the same as that for HAL, with a few notable exceptions. COALS disregards whether a context word occurs before or after a key word, and uses a ramped 4-word window instead of 10. Additionally, the initial sparse matrix that COALS generates is reduced by discarding all but the columns representing the words with the highest word frequency.

Rohde et al. (2007) argue that the most interesting data contained in the norms is not the raw co-occurrence patterns of words (i.e., do words x and y tend to co-occur in general), but rather the conditional co-occurrence patterns (i.e., does x tend to occur around y more or less often than it does in general). They determine the conditional co-occurrence by computing Pearson's correlation coefficients between the different word-pairs in the matrix, and replace the individual cell values in the co-occurrence matrix with these correlation values. To simplify the matrix, the negative correlations are set to zero, as they likely do not have a large impact on determining a word's meaning (e.g., it is easier to guess a word from a small number of positively correlated words, such as 'wings' and 'beak', than a small number of negatively correlated words, such as 'cement' and 'chair'; although see Griffiths et al., 2007 for a discussion on the importance of small negative associations). As a final step, the correlation coefficients are square rooted to magnify the differences between the different correlations. Optionally, the final COALS matrix can be compressed using a special form of SVD, as in LSA, to produce low dimensionality binary-valued vectors which may be better suited for use in connectionist models.

The COALS model, and especially the COALS-SVD variant, has been shown to be approximately equal or superior to other co-occurrence models (e.g., HAL, LSA) and other WordNet-based models on a variety of word-pair similarity ratings (e.g., Miller Charles Ratings, Miller & Charles, 1991; Finkelstein et al. ratings, Finkelstein et al., 2002), and on multiple-choice vocabulary tests (e.g., the TOEFL as first used by Landauer & Dumais, 1997; the Reader's Digest Word Power quizzes as used by Jarmasz & Szpakowicz, 2003). Additionally, Rohde et al. (2007) have provided a thorough analysis of how different model parameters (e.g., number of dimensions conserved after trimming based on frequency, size of text corpus, size of ramped window) influence results. This has provided researchers with a valuable tool in understanding how the different model parameters interact, and how to optimize the model's efficacy in a specific situation. Additional information related to COALS can be found at: http://dlt4.mit.edu/~dr/COALS/.

BEAGLE is a fourth approach to deriving meaning from text, and takes inspiration from Murdock’s (1982) work on associative memory theory. The main benefits of BEAGLE are that it incorporates a mechanism through which both context and transitional information can be encoded within the same vector representation, it learns continuously, and its representations develop and can be accessed throughout learning.

Words are represented as holographic vectors. Each unique word in the corpus is assigned an environmental vector of normally distributed random values that stands for the word form. Each word is also assigned a memory vector, which is updated each time that word appears in text to reflect information about the context in which the word appears. As the model processes a sentence (this model’s unit of processing), a context vector is computed which reflects the sum of the environmental vectors for the other words in the sentence. This context vector is then added to the memory vector for the word and represents the associative component of the word’s representation. Transitional information is encoded by adding to the memory vector via non-commutative circular convolution (Plate, 1995). Conceptually, this process corresponds to creating additional ‘position’ vectors which reflect the order with which individual words, pairs of words, and groups of words are associated (i.e., co-occur) with the word of interest.

Word vectors thus come to represent both context and positional information. Similar vectors are produced for words with similar meanings that may not have occurred in the same contexts. For example, kerosene and paraffin may not co-occur frequently in sentences, but could still come to have similar memory vectors. These memory vectors differ from those produced by the other word co-occurrence models because contextual or positional information (or both) can be extracted from a single memory representation. BEAGLE has been shown to perform well on a variety of cognitive tasks including: semantic priming, associative priming, mediated priming (Jones, Kintsch, & Mewhort, 2006), and modeling the developmental trajectories of language acquisition (Riordan & Jones, 2007).

Additionally, BEAGLE has been shown to be largely insensitive to the dimensionality of a word's memory vectors, so long as they are sufficiently large to represent all of the learned information. This is an important distinction between this model and the other three models which have been discussed (particularly LSA), as the dimensionality of the word representations have been shown to significantly modulate the model's performance (Landauer & Dumais, 1997; Rohde et al., 2007).

The Topic model (Griffiths et al., 2007) is a fifth approach to extracting meaning from text, and is based on earlier models (Blei, Ng, & Jordan, 2003) developed to generate documents based on the underlying topics under discussion. Recent work has expanded the application of the Topic model to invert this inference, such that this model can now be used to infer the general topic(s) of a document dynamically based on the contributions of each word's meanings to different topics.

Many variants of the Topic model have been proposed, but the most interesting cases for our purposes are the recent variants which are able to represent both the positional and associative information in text (Griffiths, Steyvers, Blei, & Tenenbaum, 2005), and which are able to dynamically adjust the number of topics required to adequately capture the structure of meanings in words (Griffiths & Steyvers, 2004; Griffiths et al., 2007). To accomplish these goals, the topic model uses various Bayesian statistical procedures (e.g., Marcov Chain Monte Carlo algorithms, Latent Dirichlet Allocation) to generate distributed representations of latent topics and words as the probabilities that particular words will occur in particular contexts.

The Topic model has several advantages relative to previous word co-occurrence models, primarily because it does not represent words as locations in a semantic space (BEAGLE's holographic representations may also possess these advantages, but to date this has not been explicitly tested). One of the most important of these differences is the ability to represent the different meanings of semantically ambiguous words (i.e., homonymous and polysemous words), which tap different semantic representations depending on the context in which they are encountered (e.g., 'bank' can refer to either a financial institution or the border of a river, depending on the context). This is an important step forward relative to previous models, which conflated the meanings of semantically ambiguous words to produce representations in which the different meanings were averaged into a single representation. For example, in the previously discussed accounts, words such as ‘bank’ are represented as an average of the points in semantic space representing each different interpretation of the word. Thus, the location in semantic space in which these words are represented does not capture the fact that these words are strongly associated with two different clusters of knowledge. In contrast, the Topic model can represent each meaning of a word as occurrences in two distinct latent topics, thus conserving a representation of each particular interpretation.

A second major advantage of the Topic model is that it does not suffer from the disadvantages of purely spatial representations of memory, as outlined by Tversky (1977). Tversky originally criticized spatial accounts of memory because they failed to fully capture human similarity judgments. To borrow Tversky’s examples, semantic space accounts fail to capture the fact that similarity can be asymmetric (e.g., North Korea is judged to be more like China than China is to North Korea), and that similarity does not always obey the triangle equality, in which meaning ‘x’ can resemble meaning ‘y’, meaning ‘y’ can resemble meaning ‘z’, but meaning ‘x’ does not resemble meaning ‘z’ (e.g., Jamaica is similar to Cuba, Cuba is similar to Soviet Russia, but Jamaica is not similar to Soviet Russia). To account for these trends, Tversky argued that specific features must be defined so that the total number of shared and non-shared features can be contrasted to one another to assess similarity. The Topic model provides solutions to both of these issues because each topic within which a word occurs with a high probability can be interpreted as a feature, and these features can be used as the basis for similarity judgments which are not subject to the problems of pure semantic space accounts.

The Topic model has been successfully employed in many tasks including dynamically selecting a correct meaning of an ambiguous word (Griffiths et al., 2007), word association, document classification

(Blei et al., 2003), including identifying scientific topics which are rising in popularity based on abstracts from the Proceedings of the National Academy of Sciences (Griffiths & Steyvers, 2004), document modeling, and collaborative filtering (Blei et al., 2003). Tools for implementing the topic model are available at http://psiexp.ss.uci.edu/research/programs_data/toolbox.htm.

Although the recent development of COALS, BEAGLE, and Topic has precluded a direct comparison of these models to date, it is interesting to note that an important component of these models' success is the fact that they attempt to integrate the contextual and positional information which is captured by LSA and HAL respectively. Given the substantially different approaches to representing association employed by these techniques, a direct comparison of these models will undoubtedly provide useful insights into the finer details of semantic representations and the mechanisms used to derive them. Furthermore, the combination of the components responsible for each model's particular successes within a single framework promises to be an interesting research avenue in future years.

There are two important limitations to current versions of word co-occurrence models that should be noted. First, representations are not sufficiently grounded in perception and action (Glenberg & Robertson, 2000). Word meanings are generated from the co-occurrence of a word with other words. This is certainly one important mechanism through which humans learn word meaning, but humans are also exposed to other sources. For example, they learn co-occurrences between words and sounds, tastes, smells, physical objects in the environment, and actions that can be performed on those objects. This is why some speak of language as being grounded in the environment – something current models are missing. Second, the vectors created by co-occurrence models can be compared to one another to assess their similarity, but there is more to meaning than the simple fact that two words *are* related – we also know *how* they are related. For example, the vectors for dog and bark may be more similar to one another than dog and meow, but there is no obvious way of determining why they are similar. The

Topic model begins to assess this issue, but has yet to map the particular representations of topics onto a knowledge source other than word/context co-occurrence, nor has it provided a deep understanding of the representational structure and processes underlying semantic memory. These issues are ripe for future research, and the solutions will be major contributions to the study of semantic memory.

Conclusions

We have reviewed the three major approaches to modeling semantic memory. Emphasis has been placed on highlighting findings that run counter to semantic network theory, the oldest, and probably still most widely accepted framework outside of the field for thinking about semantic memory. These novel findings include new ways of thinking about cognitive economy and property inheritance, explanations of differences in the speeded computation of components of meaning during word reading, insight into the often puzzling patterns of deficits observed in cases of damage, and new methods for thinking about and deriving word meaning. We are far from a complete understanding of semantic memory, but hopefully it is clear how computational models have contributed to our current understanding, and will play an important role in future research.

It is interesting to consider why semantic network theory remains the dominant model used to think about semantic memory despite the fact that its flaws have been well publicized. Several reasons are likely. First, many believe that the new alternatives produce the same old predictions. This is clearly not true. We have reviewed several cases where the predictions that would be derived from semantic network theory are at odds with the accounts provided by PDP models. Second, the alternative theories are more difficult to understand due to the inclusion of formulae, and the perceived impenetrability of the representations derived from the simulations. This may be partially true, but it is clearly not a good reason to avoid the approaches, and new techniques are being developed for visualizing the knowledge stored in PDP networks. Finally, there was a clear shift in the early 1980’s away from the study of ‘semantic memory’ and towards the new and emerging field of ‘concepts and categorization’ (see Rips &

Medin, 2005, for a review). Researchers in the new field articulated the flaws with the old ways of thinking about knowledge representation, and provided successful new alternatives in their place. However, computational models of semantic memory continued to be developed within the field of PDP modeling, and as we hope we have outlined clearly, many exciting advances have been made. Indeed, as some have argued, the two fields are now ready for re-integration, with PDP modeling perhaps providing a rigorous framework within which to incorporate the descriptive proposals of theory-theory (Rehder & Murphy, 2003; Rogers & McClelland, 2004), a process which may be facilitated and enhanced by the use of rich knowledge representations from word co-occurrence models. Regardless of the labels used to describe the work, it is the meaning of the work that is important, and there are clearly still many issues that need to be explored. Computational modeling will play an important role in providing a descriptive language for that work.

References

Anderson, J. A. (1983). *The Architecture of Cognition*. Cambridge, MA: Harvard University Press.

Audet, C., & Burgess, C. (1999). Using a high-dimensional memory model to evaluate the properties of abstract and concrete words. *Proceedings of the Cognitive Science Society*, pp. 37-42. Hillsdale, N.J.: Lawrence Erlbaum Associates.

Becker, S., Moscovitch, M., Behrmann, M., & Joordens, S. (1997). Long-term semantic priming: A computational account and empirical evidence. *Journal of Experimental Psychology: Learning, Memory and Cognition, 23*, 1059-1082.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3, 993-1022.

Burgess, C. (2001). Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling. In Gorfein, D.S. (Ed.), *On the Consequences of Meaning Selection: Perspectives on Resolving Lexical Ambiguity*. APA Press.

Burgess, C., & Lund, K. (1997). Representing abstract words and emotional connotation in high-dimensional memory space. *Proceedings of the Cognitive Science Society,* pp. 61-66. Hillsdale, NJ: Lawrence Erlbaum Associates

Burgess, C., & Lund, K. (1997). Modeling parsing constraints with high-dimensional context space. *Language and Cognitive Processes, 12*, 177-210.

Capitani, E., Laiacona, M., Mahon, B., & Caramazza, A. (2003). What are the facts of semantic category-specific deficits? A critical review of the clinical evidence. *Cognitive Neuropsychology, 20(3/4/5/6),* 213-261.

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

Collins, A. M., & Loftus, E. F. (1975). A spreading-activation theory of semantic processing. *Psychological Review, 82(6)*, 407-428.

Conrad, C. (1972). Cognitive economy in semantic memory. *Journal of Experimental Psychology, 92(2)*, 149-154.

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

Cree, G. S., McNorgan, C., & McRae, K. (2006). Distinctive features hold a privileged status in the computation of word meaning: Implications for theories of semantic memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 32(4)*, 643-658.

Damasio, A. R. (1989). The brain binds entities and events by multiregional activation from convergence zones. *Neural Computation, 1*, 123-132.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing By Latent Semantic Analysis. *Journal of the American Society for Information Science, 41*, 391-407.

Devlin, J. T., Gonnerman, L. M., Anderson, 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.

Durrant-Peatfield, M. R., Tyler, L. K., Moss, H. E., & Levy, J. P. (1997). The distinctiveness of form and function in category structure: A connectionist model. In M. G. Shafto & P. Langley (Eds.), *Proceedings of the 19th Annual Conference of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

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.

Fellbaum, C. (1998). *Wordnet: An electronic lexical database.* Cambridge, MA: MIT Press.

Finkelstein, L., Gabrilovish, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The concept revisited. *ACM Transactions on Information Systems, 20(1)*, 116-131.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol grounding and meaning: A comparison of high-dimensional and embodied theories of meaning. *Journal of Memory & Language, 43(3)*, 379-401.

Greer, M. J., van Casteren, M., McLellan, S. A., Moss, H. E., Rodd, J., Rogers, T. T., & Tyler, L. K. (2001). The emergence of semantic categories from distributed featural representations. In J. D. Moore & K. Stenning [Eds.], *Proceedings of the 23rd Annual Conference of the Cognitive Science Society*, pp. 358-363. London, UK: Lawrence Erlbaum Associates.

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. *Advances in Neural Information Processing Systems 15.*

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101,* 5228-5235.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2005). Integrating topics and syntax. *Advances in Neural Information Processing Systems 17.*

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114,* 211-244.

Harm, M.W. (1998). Division of Labor in a Computational Model of Visual Word Recognition (Doctoral dissertation, University of Southern California, 1998).

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

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

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence, 46*, 47-75.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review, 98(1),* 74-95.

Hintzman, D. L. (1991). Why are formal models useful in psychology? In W. E. Hockley and S. Lewandowsky, (Eds.), *Relating Theory and Data: Essays on Human Memory in Honor of Bennet B. Murdock*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Jankowicz, D., Becker, S. & Howell, S. R. (2003). Modelling Category-Specific Deficits using Topographic, Corpus-Derived Representations. *Poster presented at the 44th annual meeting of the Psychonomic Society, Vancouver, B.C., November 6-9 2003.*

Jarmasz, M., & Szpakowicz, S. (2003). Roget's thesaurus and semantic similarity. In *Proceedings of the Conference on Recent Advances in Natural Language Processing*, pp. 212-219. Borovets, Bulgaria.

Johnson-Laird, P. N., Herrmann, D. J., & Chaffin, R. (1984). Only connections: A critique of semantic networks. *Psychological Bulletin, 96*, 292-315.

Jones, M. N., Kintsch, W., & Mewhort, D. J. K. (2006). High-dimensional semantic space accounts of priming. Journal of Memory and Language, *55*, 534-552.

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review, 114(1)*, 1-37.

Kintsch, W. (2000). Metaphor comprehension: A computational theory. *Psychonomic Bulletin and Review, 7*, 257-266.

Kintsch, W., & Bowles, A. (2002) Metaphor comprehension: What makes a metaphor difficult to understand? *Metaphor and Symbol, 17*, 249-262.

Kintsch, E., Steinhart, D., Stahl, G. & LSA research group (2000). Developing Summarization Skills through the use of LSA-based feedback, Interactive Learning Environments, *8*, 87-109.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104(2)*, 211-240.

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

Landauer. T. K., Laham, D., Rehder, B., & Schreiner, M. E. (1997). How well can passage meaning be derived without using word order? A comparison of latent Semantic Analysis and humans. In M. G. Shafto & P. Langley, (Eds.), *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, pp. 214-417. Mahwah, NJ: Erlbaum.

Lenat, D. & Guha, R. V. (1990). *Building Large Knowledge-Based Systems: Representation and Inference in the CYC Project*. Reading, Mass: Addison-Wesley Publishing Co.

Li, Burgess, & Lund (2000). The acquisition of word meaning through global lexical co-occurrences. In *Proceedings of the 30th Child Language Research Forum*, pp. 167-178. Stanford, CA: CSLI, 2000.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces form lexical co-occurrence. *Behavioral Research Methods, Instruments, & Computers, 28(2),* 203-208.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4*, 310-322.

McCloskey, M., & Glucksberg, S. (1979). Decision processes in verifying category membership statements: Implications for models of semantic memory. *Cognitive Psychology, 11*, 1-37.

McNamara, T. P. (2005). *Semantic priming*. New York, NY: Psychology Press.

McRae, K., Cree, G. S., Westmacott, R., & de Sa, V. R. (1999). Further evidence for feature correlations in semantic memory. *Canadian Journal of Experimental Psychology, 53*, 360-373.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavioral Research Methods, Instruments, and Computers, 37*, 547-559.

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

McRae, K. (2004). Semantic memory: Some insights from feature-based connectionist attractor networks. In B. H. Ross (Ed.), *The Psychology of Learning and Motivation: Advances in Research and Theory*, Vol. 45 (pp. 41-86). San Diego, CA: Academic Press.

Masson, M. E. J. (1991). A distributed memory model of context effects in word identification. In D. Besner and G. W. Humphreys (Eds.), *Basic Processes in Reading: Visual word recognition*, 233-263. Hillsdale, NJ: Erlbaum.

Masson, M. E. J., (1995). A distributed memory model of semantic priming. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 21(1)*, 3-23.

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience, 4(4),* 310-322.

Miikkulainen, R. (1997). Dyslexic and category-specific aphasic impairments in a self-organizing feature map model of the lexicon. *Brain and Language, 59*, 334-366.

Miller, G. A., & Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes, 6(1)*, 1-28.

Murdock, B. B. (1982). A theory for the storage and retrieval of item and associative information. *Psychological Review, 89*, 609-626.

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

O’Connor, C., Cree, G. S., & McRae, K. (2007). Conceptual hierarchies arise from the dynamics of learning and processing: Insights from a flat attractor network. In prep.

Plaut, D. C. (1995). Semantic and associative priming in a distributed attractor network. In *Proceedings of the Seventeenth Annual Conference of the Cognitive Science Society*, pp. 37-42, Pittsburgh, PA. Hillsdale, NJ: Lawrence Erlbaum Associates.

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

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: Empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review, 107*, 786-823.

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

Plate, T. A. (1995). Holographic reduced representations. *IEEE Transactions on Neural Networks, 6*, 623-641.

Quillian, M. R. (1962). A revised design for an understanding machine, *Mechanical Translation, 7*, 17–29.

Quillian, M. R. (1968). Semantic Memory. In M. Minsky (Ed.), *Semantic Information Processing*, pp. 227-270, Cambridge, MA: MIT Press.

Quillian, M. R. (1967). Word Concepts: A theory and simulation of some basic semantic capabilities. *Behavioral Science, 12(5),* 410-430.

Quillian, M. R. (1969). The teachable language comprehender. *Communications of the Association for Computing Machinery, 12*, 459-475.

Rehder, B., & Murphy, G. L. (2003). A knowledge-resonance (KRES) model of category learning. *Psychonomic Bulletin & Review, 10(4),* 759-784.

Riordan, B., & Jones, M. N. (2007). Comparing semantic space models using child-directed speech. In D. S. MacNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Cognitive Science Society,* 599-604.

Rips, L. J. & Medin, D.L. (2005). Concepts, Categories, and Semantic Memory. In K. Holyoak & R. Morrison (Eds.), *Cambridge Handbook of Thinking and Reasoning*, pp. 37-72. Cambridge, U.K.: Cambridge University Press.

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A parallel distributed processing approach*. MIT Press.

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

Rogers, T. T., & Plaut, D. (2002). Connectionist perspectives on category specific deficits. In E. Forde and G. Humphreys, (Eds.), *Category Specificity in Mind and Brain*, pp. 251-289. Hove, East Sussex, U.K.: Psychology Press.

Rohde, D. L., Gonnerman, L. M., & Plaut, D. C. (2007). An improved method for deriving word meaning from lexical co-occurrence. Cognitive Science, Submitted.

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

Rumelhart, D. E. (1990). Brain style computation: Learning and generalization. In S. F. Zornetzer, J. L. Davis, and C. Lau [Eds.], *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. Rumelhart, J. L. McClelland, and the PDP Research Group, (Eds.), *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 and S. Kornblum (Eds.), *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, pp. 3-30. Cambridge, MA: MIT Press.

Schwanenflugel, P. J., & Rey, M. (1986). Interlingual semantic facilitation: Evidence for a common representational system in the bilingual lexicon. *Journal of Memory and Language, 25*, 605-618.

Sharkey, N. E. (1989). The lexical distance model and word priming. *Journal of Memory and Language, 31*, 543-572.

Simmons, W. K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology, 20(3/4/5/6),* 451-486.

Smith, E. E., Shoben, E. J., & Rips, L. J. (1974). Structure and process in semantic memory: A featural model for semantic decisions. *Psychological Review, 81*, 214-241.

Tversky, A. (1977). Features of similarity. *Psychological Review, 84*, 327-352.

Vinson, D.P., Vigliocco, G., Cappa, S. & Siri, S. (2003). The breakdown of semantic knowledge along semantic field boundaries: Insights from an empirically-driven statistical model of meaning representation. *Brain & Language, 86*, 347-365.

Warrington, E. K. (1975). The selective impairment of semantic memory. *Quarterly Journal of Experimental Psychology, 27*, 635-658.

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

Warrington, E. K., & McCarthy, R. (1987). Categories of knowledge: Further fractionation and an attempted integration. *Brain, 110*, 1273-1296.

Author Notes

The authors would like to acknowledge the insightful comments and discussion on previous versions of this manuscript provided by D. C. Plaut, D. J. K. Mewhort, M. Jones, K. Simmons , S. Gotts, and A. Martin. The creation of this manuscript was supported in part by a Natural Sciences and Engineering Research Council Discovery Grant to G. S. C. and a Canada Graduate Scholarship to B. C. A.

Table 1: LSA representation for eight sentences.

A1: Divert more power from the warp engines.
A2: Can the engines sustain warp five given their current state?
A3: Captain, the impulse engines are buckling under the stress.
A4: Drop out of warp and switch to impulse engines.

B1: Captain to transporter room, five to beam up.
B2: I tried to beam them down, but the transporter beam is destabilizing.
B3: The transporter is ready to beam you down Captain.
B4: Lock on the transporter and beam them up.

	A1	A2	A3	A4	B1	B2	B3	B4
the	1	1	2					1
warp	1	1		1		1		
engines	1	1	1	1				
captain			1		1		1	
to				1	2	1		
transporter					1	1	1	1
beam					1	2	1	1
up					1	1		1

Each column represents a sentence, and each row represents how frequently each word occurred in that sentence. In this example, words which did not occur at least three times across all sentences have been removed from the table.

Table 2: HAL representation matrix for the sentence "Captain to transporter room, five to beam up."

	captain	to	transporter	room	five	beam	up
captain							
to	6	2	3	4	5		
transporter	4	5					
room	3	4	5				
five	2	3	4	5			
beam		1	2	3	4		
up			1	2	3	5	

Each row denotes the co-occurrence of context words (column labels) BEFORE the key word (the row label). Cells with a co-occurrence value of zero were left blank. Each column denotes the co-occurrence of context words (row labels) AFTER the key word (the column label). In this example, a window of 5 words was used to record co-occurrence for illustrative purposes.

Figure Captions.

Figure 1. Architecture of the hierarchical network theory as proposed by Collins and Quillian (1969).

Figure 2. Architecture of Hinton's (1981) PDP model designed to encode propositional information. Note that the number of units in each layer is meant to be illustrative, and does not reflect the number used in Hinton's simulations.

Figure 3. Architecture of Rumelhart and Todd's (1991) PDP network designed to encode the information present in Collins and Quillian's (1969) hierarchical network theory. In this illustration, 'canary' and 'can' have been activated at the input layer, and the network was produced the correct output pattern across the attribute layer.

Figure 4. Architecture of Hinton and Shallice's (1991) attractor network. Note the connections leading back from the clean-up units to the sememe units. Also, note that the self-connections at the sememe layer represent the fact that small sets of sememe units were interconnected, each set representing values on one attribute dimension, and that not all sememe units were fully interconnected.

Figure 1.

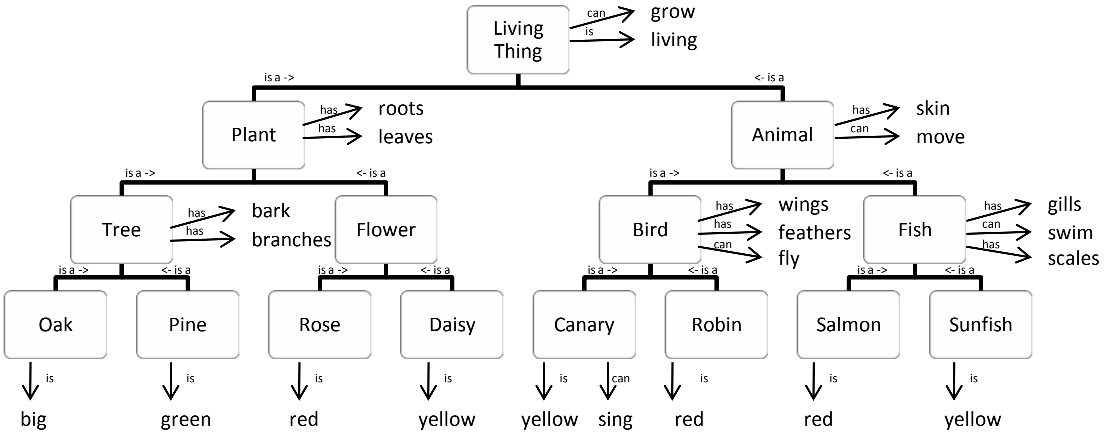

Living Thing
can
is
grow
living
is a ->
<- is a
Plant
has
has
roots
leaves
Animal
has
can
skin
move
is a ->
<- is a
is a ->
<- is a
Tree
has
has
bark
branches
Flower
Bird
has
has
can
wings
feathers
fly
Fish
has
can
has
gills
swim
scales
is a ->
<- is a
is a ->
<- is a
is a ->
<- is a
is a ->
<- is a
Oak
Pine
Rose
Daisy
Canary
Robin
Salmon
Sunfish
is
is
is
is
is
can
is
is
is
big
green
red
yellow
yellow
sing
red
red
yellow

Figure 2.

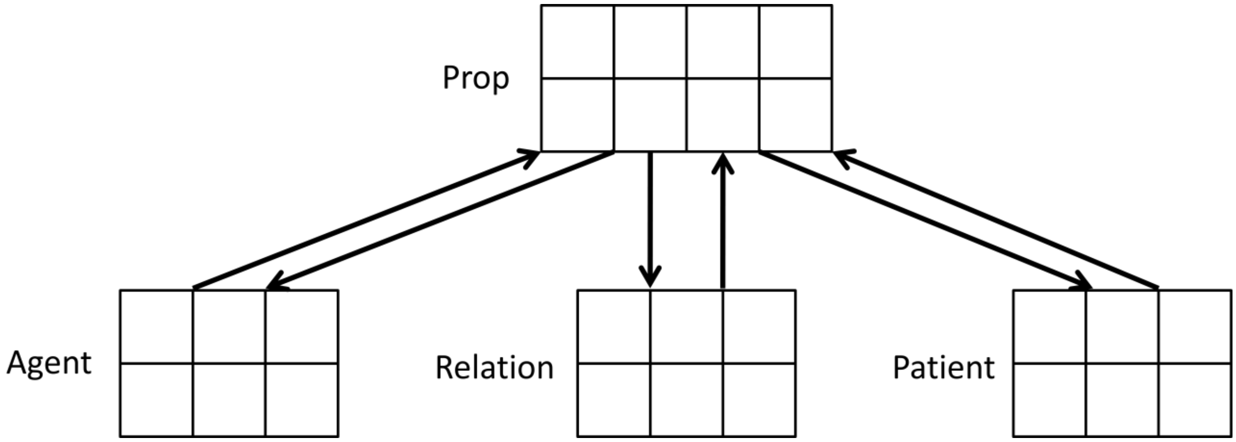


Figure 3.

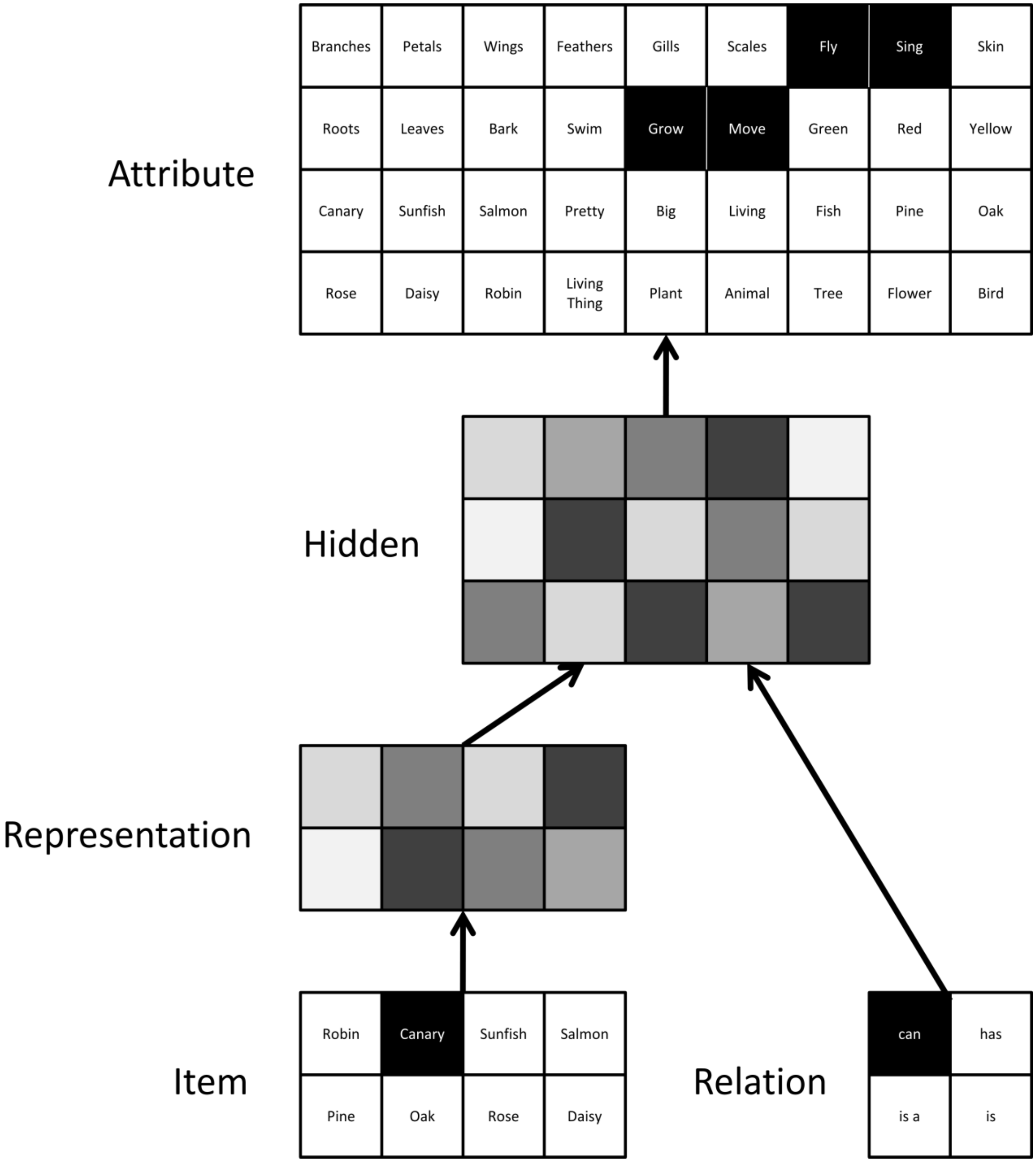
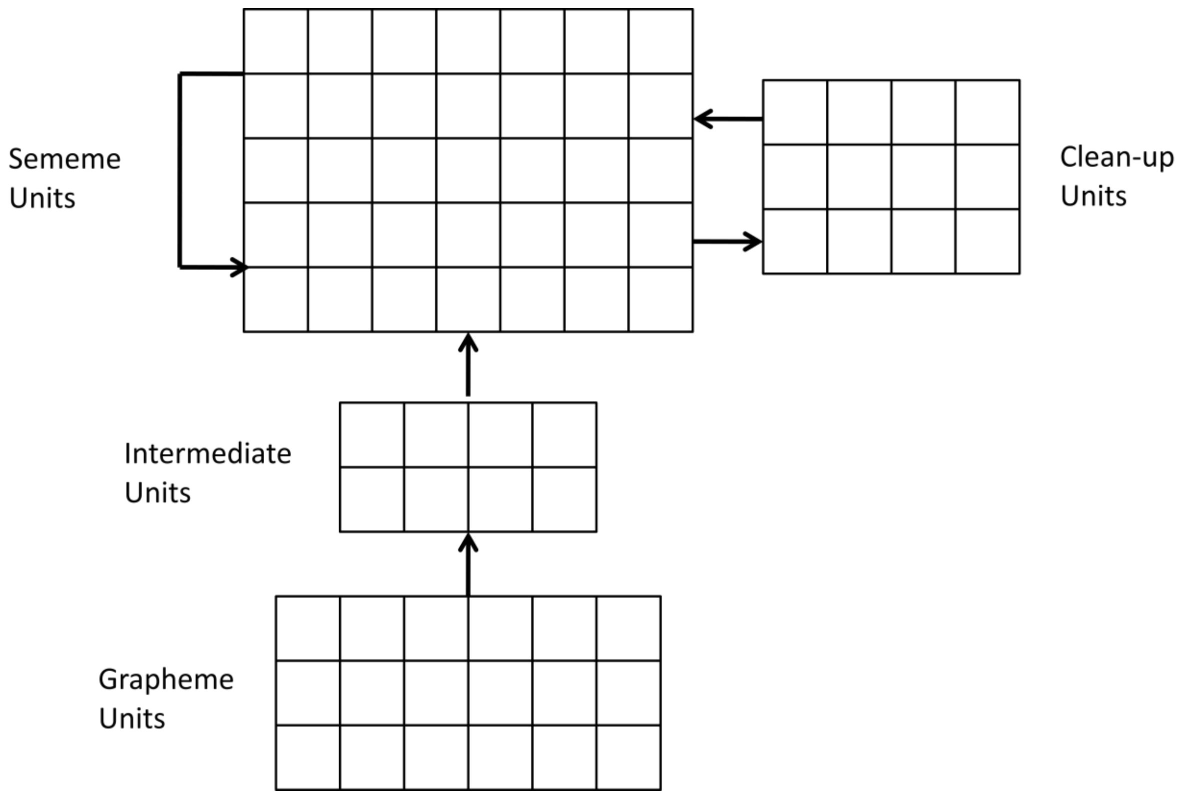

Attribute
Branches
Petals
Wings
Feathers
Gills
Scales
Fly
Sing
Skin
Roots
Leaves
Bark
Swim
Grow
Move
Green
Red
Yellow
Canary
Sunfish
Salmon
Pretty
Big
Living
Fish
Pine
Oak
Rose
Daisy
Robin
Living Thing
Plant
Animal
Tree
Flower
Bird
Hidden
Representation
Item
Robin
Canary
Sunfish
Salmon
Pine
Oak
Rose
Daisy
Relation
can
has
is a
is

Figure 4.


Sememe
Units
Clean-up
Units
Intermediate
Units
Grapheme
Units