MODELS OF SEMANTIC MEMORY

Michael N. Jones,[1] Jon Willits,[1] & Simon Dennis[2]

[1] *Indiana University*

[2] *The Ohio State University*

Correspondence:
Michael Jones
Department of Psychological and Brain Sciences
1101 E. 10th St.
Indiana University
Bloomington, IN, 47405

Email: jonesmn@indiana.edu
Phone: (812) 856-1490
Fax: (812) 855-4691

ABSTRACT

Meaning is a fundamental component of nearly all aspects of human cognition, but formal models of semantic memory have classically lagged behind many other areas of cognition. However, computational models of semantic memory have seen a surge progress in the last two decades, advancing our knowledge of how meaning is constructed from experience, how knowledge is represented and used, and what processes are likely to be culprit in disorders characterized by semantic impairment. This chapter provides an overview of several recent clusters of models and trends in the literature, including modern connectionist and distributional models of semantic memory, and contemporary advances in grounding semantic models with perceptual information and models of compositional semantics. Several common lessons have emerged from both the connectionist and distributional literatures, and we attempt to synthesize these themes to better focus future developments in semantic modeling.

1. INTRODUCTION

Meaning is simultaneously the most obvious feature of memory—we can all compute it rapidly and automatically—and the most mysterious aspect to study. In comparison to many areas of cognition, relatively little is known about how humans compute meaning from experience. Nonetheless, a mechanistic account of semantics is an essential component of all major theories of language comprehension, reading, memory, and categorization. Semantic memory is necessary for us to construct meaning from otherwise meaningless words and utterances, to recognize objects, and to interact with the world in a knowledge-based manner.

Semantic memory typically refers to memory for word meanings, facts, concepts, and general world knowledge. For example, you know that a *panther* is a jungle cat, is more like a *tiger* than a *corgi*, and you know better than to try to pet one. The two common types of semantic information are conceptual and propositional knowledge. A *concept* is a mental representation of something, such as a *panther*, and knowledge of its similarity to other concepts. A *proposition* is a mental representation of conceptual relations that may be evaluated to have a truth value, for example, that a *panther* is a jungle cat, or has four legs and knowledge that panthers do not have gills.

In Tulving's (1972) classic modular taxonomy, declarative memory was subdivided into episodic and semantic memory, the former containing memory for autobiographical events, and the latter dedicated to generalized memory not linked to a specific event. While you may have a specific autobiographical memory of the last time you saw a *panther* at the zoo, you do not have a specific memory of when you learned that a *panther* was a jungle cat, was black, or how it is similar to a *tiger*. In this sense, semantic memory gained a reputation as the more miscellaneous and mysterious of the memory systems. While episodic memory could be studied with experimental tasks such as list learning and could be measured quantitatively by counting the

number of items correctly recognized or recalled, semantic memory researchers focused more on tasks such as similarity judgments, proposition verification, semantic priming, and free association. Unlike episodic memory, there existed no mechanistic account of how semantic memory was constructed as a function of experience. However, the field has advanced a considerable amount in the past 25 years.

A scan of the contemporary literature reveals a large number of formal models that aim to understand the mechanisms that humans use to construct semantic memory from repeated episodic experience. Modern semantic models have made truly impressive progress at elucidating how humans learn and represent semantic information, how semantic memory is recruited and used in cognitive processing, and even how complex functions like semantic composition may be accomplished by relatively simple cognitive mechanisms. Many of the current advances build from classic ideas, but only relatively recently has computational hardware advanced to a scale where we can actually simulate and evaluate these systems. Advances in semantic modeling also owe to excellent interdisciplinary collaboration, building in part on developments in computational linguistics, machine learning, and information retrieval.

The goal of this chapter is to provide an overview of recent advances in models of semantic memory. We will first provide a brief synopsis of classic models and themes in semantic memory research, but will then focus on computational developments. In addition, the focus of the chapter is on models that have a formal instantiation that may be tested quantitatively. Hence, while there are several exciting new developments in verbal conceptual theory (e.g., Louwerse's (2011) Symbol Interdependency Hypothesis), we focus exclusively on models that are explicitly expressed by computer code or mathematical expressions. In addition, the chapter assumes a sufficient understanding of the empirical literature on semantic memory.

For an overview of contemporary experimental findings, we refer the reader to a companion chapter on Semantic Memory by McRae and Jones (2013).

There are several potential ways to organize a review of the literature, and no single structure will satisfy all theorists. We opt here to follow two major clusters of cognitive models that have been prominent: distributional models and connectionist models. The division may also be broadly thought of as a division between models that specify how concepts are learned from statistical experience (distributional models), and models that specify how propositions are learned or that use conceptual representations in cognitive processes (connectionist models). Obviously, there are exceptions in both clusters that cross over, but the two literatures have had different foci. Next, we summarize some classic models of semantic memory and common theoretical debates that have extended to the contemporary models. Following the historical trends in the literature, we then discuss advances in connectionist models, followed by distributional models. Finally, we discuss hybrid approaches, new directions in models of grounded semantics and compositional semantics, and attempt to synthesize common lessons that have been learned across the literature.

2. CLASSIC MODELS AND THEMES IN SEMANTIC MEMORY RESEARCH

The three classic models of semantic memory most commonly discussed are semantic networks, feature-list models, and spatial models. These three models deserve mention here, both because they have each seen considerable attention in the literature, and because features of each have clearly evolved into modern computational models.

The semantic network has traditionally been one of the most common theoretical frameworks used to understand the structure of semantic memory. Collins and Quillian (1969) original proposed a hierarchical model of semantic memory in which concepts were nodes and propositions were labeled links (e.g., the nodes for *dog* and *animal* were connected via an ‘isa’

link). The superordinate and subordinate structure of the links produced a hierarchical tree structure (animals were divided into birds, fish, etc., and birds were further divided into robin, sparrow, etc.), and allowed the model to explain both conceptual and propositional knowledge within a single framework. Accessing knowledge required traversal of the tree to the critical branch, and the model was successful in this manner of explaining early sentence verification data from humans (e.g., the speed to verify that "a canary can sing"). A later version of the semantic network model proposed by Collins and Loftus (1975) deemphasized the hierarchical nature of the network in favor of the process of spreading activation through all network links simultaneously to account for semantic priming phenomena—in particular, the ability to produce fast negative responses. Early semantic networks can be seen as clear predecessors to several modern connectionist models, and features of them can also be seen in modern probabilistic and graphical models as well.

A competing model was the feature-comparison model of Rips, Shoben, and Smith (1973). In this model, a word's meaning is encoded as a list of binary descriptive features, which were heavily tied to the word's perceptual referent. For example, the <has_wings> feature would be turned on for a *robin*, but off for a *beagle*. Smith, Shoben, and Rips (1974) proposed two types of semantic features: Defining features that all concepts have, and characteristic features that are typical of the concept, but are not present in all cases; for example, all birds have wings, but not all birds fly. Processing in the model was accomplished by computing the feature overlap between any two concepts, and the features were allowed to vary in their contribution of importance to the concept, although how particular features came to be and how they were ranked was not fully specified. Modern versions of feature-list models use aggregate data collected from human raters in property generation tasks (e.g., McRae, de Sa, & Seidenberg, 1997).

A third type was the spatial model, which emerged from Osgood's (1952, 1971) early attempts to empirically derive semantic features using semantic differential ratings. Osgood had humans rate words on a Likert scale against a set of polar opposites (e.g., *rough-smooth, heavy-light*), and a word's meaning was then computed as a coordinate in a multidimensional semantic space. Distance between words in the space was proposed as a process for semantic comparison.[1] Featural and spatial representations have been contrasted as models of human similarity judgments (e.g., Tversky & Gati, 1982), and the same contrast applies to spatial vs. featural representations of semantic representations. We will see the feature vs. space debate emerge again with modern distributional models. Early spatial models can be seen as predecessors of modern semantic space models of distributional semantics (but co-occurrences in text corpora are used as the data on which the space is constructed rather than human ratings).

One issue with all three of these classic models is that none ever did actually learn anything. Each model relied on representations that were hand coded based on the theorist's intuition (or subjective ratings) of semantic structure, but none formally specified the cognitive mechanisms by which the representations were constructed. As Hummel and Holyoak (2003) have noted, this type of intuitive modeling may have serious consequences: "The problem of hand-coded representations is the most serious problem facing computational modeling as a scientific enterprise. All models are sensitive to their representation, so the choice of representation is among the most powerful wildcards at the modeler's disposal" (p. 247). As we will see later in the chapter, this is exactly the concern that modern distributional models address.

3. CONNECTIONIST MODELS OF SEMANTIC MEMORY

Connectionist models were among the first to specify how semantic representations might come to be learned, and how those representations might interact with other cognitive processes.

[1] One interpretation of feature comparison given by Rips et al., 1974 was also spatial distance.

Modern connectionism is a framework used to model mental and behavioral phenomena as an emergent process—one that arises out the behavior of networks of simple interconnected units (Rumelhart & McClelland, 1986). Connectionism is a very broad enterprise. Connectionist models can be used to explicitly model the interaction of different brain regions or neural processes (O'Reilly et al., 2012) or they can be used to model cognition and behavior from a "neurally-inspired" perspective, which values the way in which the models exhibit parallel processing, interactivity, and emergentism (Rumelhart & McClelland, 1986; Rogers & McClelland, 2006, 2008). Connectionist models have made a very large contribution to simulating and understanding the dynamic nature of semantic knowledge and how semantic knowledge interacts with other cognitive processes.

Connectionist models represent knowledge in terms of weighted connections between interconnected units. A model's set of units, its connections, and how they are organized is called the model's *architecture*. Research involving connectionist models has studied a wide range of architectures, but most connectionist models share a few common features. Most models have at least one set of units designated as *input* units, as well as at least one set of units designated as *target* or *output* units. Most connectionist models also have one or more sets of intervening units between the input and output units, which are often referred to as *hidden* layers.

A connectionist model represents knowledge in terms of the strength of the weighted connections between units. Activation is fed into the input units, and that activation in turn activates (or suppresses) the units to which the input units are connected, as a function of the weighted connection strength between the units. Activation eventually propagates to the output units, with one important question of interest being, what output units will a connectionist model activate given a particular input. In this sense, the knowledge in connectionist models is typically thought of as representing the function or relationship between a set of inputs and a set of

outputs. Connectionist models should *not*, however, be confused with models that map simple stimulus-response relationships; The hidden layers between input and output layers in connectionist networks allow them to learn very complex internal representations. Models with an architecture such as the one just described, where activation flows from input units to hidden units to output units, are typically referred to as *feed-forward* networks.

A key aspect of connectionist models is that they are often used to study the learning process itself. Typically, the weights between units in a connectionist network are initialized to a random state. The network is then provided with a *training phase*, in which the model is provided with inputs (typically involving some sort of expected input from the environment), and the weights are adjusted as a function of the particular inputs the network received. Learning (adjusting the weights) is accomplished in either an *unsupervised* or a *supervised* fashion. In unsupervised learning, weights are typically adjusted according some sort of associative principle, such as Hebbian learning (Hebb, 1946; Grossberg, 1976), where weights between units are increased the more often the two units are active at the same time. In supervised learning, weights are adjusted by observing which output units the network activated given a particular input pattern, and comparing that to some goal or target output given those inputs. The weights are then adjusted so as to reduce the amount of error the network makes in terms of its activation of the "correct" and "incorrect" outputs (Kohonen, 1977; Rosenblatt, 1959; Rumelhart, Hinton, & Williams, 1986; Widrow & Hoff, 1960).

3.1 Rumelhart Networks

An illustrative example of a connectionist model of semantic memory (shown in Figure 1A) was first presented by Rumelhart & Todd (1982) and studied in detail by Rogers and McClelland (2006). This network has two sets of input units: (1) a set of units meant to represent words or concepts (e.g. *robin*, *canary*, *sunfish*, etc.), and (2) a set of units meant to represent different

types of relations (e.g. *is-a*, *can*, *has*, etc.). The network learns to associate conjunctions of those inputs (e.g. *robin+can*) with outputs representing semantic features (e.g. *fly*, *move*, *sing*, *grow*, for *robin+can*). The model accomplishes this using supervised learning, having *robin+can* activated as inputs, observing what a randomly initialized version of the model produces as an output, and then adjusting the weights so as to make the activation of the correct outputs more likely. The model is not merely learning associations between inputs and outputs—in the Rumelhart network, the inputs and outputs are mediated by two sets of hidden units, which allow the network to learn complex internal representations for each input.

A critical property of connectionist architectures using hidden layers is that the same hidden units are being used to create internal representations for all possible inputs. In the Rogers et al. example, r*obin*, *oak*, *salmon,* and *daisy* all use the same hidden units; what differentiates their internal representations is that they instantiate different distributed patterns of activation. But because the network is using overlapping distributed representations for all of the concepts, this means that during the process of learning, changing the connection weights as a result of learning about one input could potentially affect how the network represents all other items. When the network learns an internal representation (i.e. hidden unit activation state) for the input *robin+can*, and learns to associate the outputs *sing* and *fly* with that internal representation, this will mean that other inputs whose internal representations are similar to *robin* (i.e. have similar hidden unit activation states, such as *canary*) will also become more associated with *sing* and *fly*. This provides these networks with a natural mechanism for categorization, generalization, and property induction. The behavior allows researchers using connectionist models to study how these networks categorize, and to compare the predictions of the model to human behaviors.

Rogers and McClelland (2006) extensively studied the behavior of the Rumelhart networks, and found that the model provides an elegant account of a number of aspects of human concept

acquisition and representation. For example, they found that as the model acquires concepts through increasing amounts of experience, the internal representations for the concepts show progressive differentiation, learning broader distinctions first and more fine-grained distinctions later, similar to the distinctions children show (Mandler et al., 1991). In the model this happens because the network is essentially performing something akin to a principal component analysis, learning the different features in the order of the amount of variance in the input that they explain. Rogers and McClelland argued that this architecture, which combines simple learning principles with the expected structure of the environment, can be used to understand how certain features (those that have rich covariational structure) become the features that organize categories, and how conceptual structure can become reorganized over the course of concept acquisition. The general (and somewhat controversial) conclusion that Rogers and McClelland draw from their study of this model is that a number of properties of the semantic system, such as the taxonomic structure of categories (Bower et al., 1970) and role of causal knowledge in semantic reasoning (Keil, 1989), can be explained as an emergent consequence of simple learning mechanisms combined with the expected structure of the environment, and that these structural factors do not necessarily need to be explicitly built into models of semantic memory.

Feed-forward connectionist models have only been used in a limited fashion to study the actual structure of semantic memory. However, these models have been used extensively to study how semantic structure interacts with various other cognitive processes. For example, feed-forward models have been used to simulate and understand the word learning process (Gasser & Smith, 1998; Regier, 2005). These word-learning models have been used to show that many details about the representation of word meanings (like hierarchical structure), learning constraints (such as mutual exclusivity and shape bias), and empirical phenomena (such as the vocabulary spurt that children show around two years of age) emerge naturally from the structure

of environment with a simple learning algorithm, and do not need to be explicitly built into the model. Feed-forward models have also been used to model consequences of brain damage (Farah & McClelland, 1991; Rogers et al., 2004; Tyler et al., 2000), Alzheimer's disease (Chan, Salmon, & Butters, 1998), schizophrenia (Braver, Barch, & Cohen, 1998; Cohen and Servan-Schreiber, 1992; Nestor et al., 1998), and a number of other disorders that involve impairments to semantic memory (see Aakerlund & Hemmingsen, 1998, for a review). These models typically study brain disorders by lesioning the network (i.e. removing units or connections), or otherwise causing the network to behave in suboptimal ways, and then studying the consequences of this disruption. Connectionist models provide accounts of a wide range impairments and disorders, and have also been used to show that many semantic consequences of impairments and disorders, such as the selective impairment of certain categories, can be explained in terms of emergent processes deriving from the interaction of low-level features, rather than requiring explicit instantiations in the model (such as creating modular memory systems for living and nonliving things, see McRae and Cree, 2002, for a review).

3.2 Dynamic Attractor Networks

In addition to feedforward models such as the Rumelhart network, a considerable amount of semantic memory research has explored the use of *dynamical* connectionist models (Hopfield, 1982). A connectionist model becomes a dynamical model when its architecture involves some sort of bi-directionality, feedback, or recurrent connectivity. Dynamical networks allow investigations into how the activation of representations may change over time, as well as how semantic representations interact with other cognitive processes in an online fashion.

For example, Figure 2A shows McLeod, Shallice, and Plaut's (2000) dynamical network for pronouncing printed words. The network has a layer of units for encoding orthographic representations (grapheme units), a layer of units for encoding phonological representations

(phoneme units), and an intervening layer between the two that encodes the words semantic features (sememe units) - as well as additional layers of hidden units between each of these layers. Critically, the activation in this network is allowed to flow in both directions, from phonemes to sememes to graphemes, and from graphemes to sememes to phonemes. The network also has recurrent connections (the loops in figure 2A) connecting the grapheme, sememe, and phoneme layers to themselves. The combination of the bidirectional connections and recurrent connectivity allows the McLeod et al. network to establish a dynamical system where the activation at the various levels will feed back and forth eventually settling into a stable *attractor* state. The result is that these attractor networks can allow multiple constraints (e.g. the weights that establish the network's knowledge of the links between orthography and semantics, and semantics and phonology) to compete, eventually settling into a state which satisfies the most likely constraints for a given input.

As an illustration of how this works, consider an example using the McLeod et al. network, shown in Figure 2B. Here, the network is simulating the experience of a person reading words. The figure depicts a three-dimensional space, where the vertical direction (labeled "energy") represents the stability of the network's current state (versus its likelihood to switch to a new state) as activity circulates through the network. In an attractor network, only a small number of possible states are stable. These stable states are determined by the network's knowledge about the likelihood of certain orthographic, phonological, and semantic states to co-occur. And given any input, the network will eventually settle into one of these stable states. For example, if the network receives a clear case of the printed word DOG as input, and this input is not disrupted, the network will quickly settle into the corresponding DOG state in its orthographic, phonological, and semantic layers. Alternatively, if the network received a nonword like DAG as an input, it would eventually settle into a neighboring attractor state (like DOG or DIG or DAD).

Similarly, if the network receives DOG as an input, but this input is impoverished (e.g. noisy, with errors in the input signal), or disrupted (simulating masking such as might happen in a psychology experiment), this can affect the network's ability to settle into the correct attractor. In a manner corresponding well to the disruption effects that people show in behavioral experiments, an early disruption (before the network has had a chance to settle into an orthographic attractor basin) can lead the network to make a form based error (settling into the LOG basin instead). A later disruption – happening after the orthographic layer has settled into its basin but before the semantic layer has done so – can lead the network to make a semantic error, activating a code of semantic features corresponding to CAT.

Attractor networks have been used to study a very wide range of semantic-memory related phenomena. Rumelhart et al. (1986) used an attractor network to show how schemas (e.g., one's representations for different rooms) can emerge naturally out of the dynamics of co-occurrence of lower level objects (e.g., items in the rooms), without needing to build explicit schema representations into the model (see also Botvinick & Plaut, 2004). Like the McLeod example already described, attractor networks have been extensively used to study how semantic memory affects lexical access (Harm & Seidenberg, 2004; McLeod et al., 2000) as well as to model semantic priming (Cree, McRae, & McNorgan, 1998; McRae, et al., 1997; Plaut & Booth, 2000). Dynamical models have also been used to study the organization and development of the child lexicon (Horst, McMurray, & Samuelson, 2006; Li, Xhao, & MacWhinney, 2007), the bilingual lexicon (Li, 2009), and children's causal reasoning using semantic knowledge (McClelland & Thompson, 2007), and how lexical development differs in typical and atypical developmental circumstances (Thomas & Karmiloff-Smith, 2003).

Dynamical connectionist models have also simulated various ways that semantic knowledge impacts and interacts with sentence production and comprehension, including how

semantic constraints impact the grammaticality of sentences (Allen & Seidenberg, 1999; Dell, Chang, & Griffin, 1999; McClelland, St. John, & Taraban, 1989; Tabor & Tanenhaus, 1999; Taraban & McClelland, 1988), and how semantic knowledge assists in the learning of linguistic structure (Borovsky & Elman, 2006; Chang, Dell, & Bock, 2006; Rohde & Plaut, 2000). As with feedforward models, dynamical models have been also used to extensively study many developmental and brain disorders such as dyslexia and brain damage (Devlin et al, 1998; Hinton & Shallice, 1991; Kinder & Shanks, 2003; Lambon Ralph et al., 2001; Plaut, 1999, 2002).

4. DISTRIBUTIONAL MODELS OF SEMANTIC MEMORY

There are now a large number of computational models in the literature that may be classified as *distributional*. Other terms commonly used to refer to these models are corpus-based, semantic space, or co-occurrence models, but distributional is the most appropriate term common to all the models in that it fairly describes the environmental structure all learning mechanisms capitalize on (i.e., not all are truly spatial models, and most do not capitalize merely on direct co-occurrences). The various models differ greatly in the cognitive mechanisms they posit that humans use to construct semantic representations, ranging from Hebbian learning to probabilistic inference. But the unifying theme common to all these models is that they hypothesize a formal cognitive mechanism to learn semantics from repeated episodic experience in the linguistic environment (typically a text corpus).

The driving theory behind modern distributional models of semantic representation is certainly not a new one, and dates back at least to Wittgenstein (1953). The most famous and commonly used phrase to summarize the approach is Firth's (1957) "you shall know a word by the company it keeps," and this idea was further developed by Harris (1970) into the *distributional hypothesis* of contextual overlap. For example, *robin* and *egg* may become related because they tend to co-occur frequently with each other. In contrast, *Robin* and *sparrow* become

related because they are frequently used in similar contexts (with the same set of words), even if they rarely co-occur directly. *Ostrich* may be less related to *robin* due to a lower overlap of their contexts compared to *sparrow*, and *stapler* is likely to have very little contextual overlap with *robin*. Formal models of distributional semantics differ in their learning mechanisms, but they all have the same overall goal of formalizing the construction of semantic representations from statistical redundancies in language.

A taxonomy of distributional models is very difficult now given the large number of them and range of learning mechanisms. The models can be loosely clustered based on their notion of context (e.g., documents, words, time, etc.), or the learning mechanism they employ. We opt for the latter organization here, and just present some standard exemplars of each model type—an exhaustive description of all models is beyond the scope of this chapter (for reviews, see Bullinaria & Levy, 2007; Riordan & Jones, 2011; Turney & Pantel, 2010).

4.1 Latent Semantic Analysis

Perhaps the best-known distributional model is Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). LSA begins with a term-by-document frequency matrix of a text corpus, in which each row vector is a word's frequency distribution over documents. A document is simply a 'bag-of-words' in which transitional information is not represented. Next, a word's row vector is transformed by its log frequency in the document and its information entropy over documents ($-\sum p(x)\log_2 p(x)$; cf. Salton & McGill, 1983). Finally, the matrix is factorized using singular-value decomposition (SVD) into three component matrices, *U, Σ,* and *V*. The *U* matrix represents the orthonormal basis for a space in which each word is a point, *V* represents an analogous orthonormal document space, and *Σ* is a diagonal matrix of singular values (cf. an eigenvector) weighing dimensions in the space (see Landauer, McNamara, Dennis,

& Kintsch, 2007 for a tutorial). The original transformed term-by-document matrix, *M*, may be reconstructed as:

$$M = U\Sigma V^T, \quad (1)$$

where V^T is the transpose of *V*.

More commonly, only the top *N* singular values of Σ are retained, where *N* is usually around 300. This dimension reduction allows an approximation of the original 'episodic' matrix to be reconstructed, and has the effect of bringing out higher-order statistical relationships among words more sophisticated than mere direct co-occurrence. A word's semantic representation is then a pattern across the *N* latent semantic dimensions, and is often projected as a point in *N*-dimensional semantic space (cf. Osgood, 1952). Even though two words (e.g., *boat* and *ship*) might have had zero similarity in the original *M* matrix, indicating that they do not co-occur in the same documents, they may nonetheless be proximal in the reduced space reflecting their deeper semantic similarity (contextual similarity but not necessarily contextual overlap).

The application of SVD in LSA is quite similar to common uses of principal component analysis (a type of SVD) in questionnaire research. Given a pattern of observable scale responses to items on a personality questionnaire, for example, the theorist may apply SVD to infer a small number of latent components (e.g., extroversion, neuroticism) that are causing the larger number of observable response patterns. Similarly, LSA uses SVD to infer a small number of latent semantic components in language that explain the pattern of observable word co-occurrences across contexts. In this sense, LSA was the first model to successfully specify a function mapping semantic memory to episodic context. Landauer and Dumais (1997) were careful not to claim that humans use exactly SVD as a learning mechanism, but rather that the brain uses some dimensional reduction mechanism akin to SVD to create abstract semantic representations from experience.

The semantic representations constructed by LSA have demonstrated remarkable success at simulating a wide range of human behavioral data, including judgments of semantic similarity (Landauer & Dumais, 1997), word categorization (Laham, 2000), and discourse comprehension (Kintsch, 1998), and the model has also been applied to the automated scoring of essay quality (Landauer, Lahma, Rehder, & Schreiner, 1997). One of the most publicized feats of LSA was its ability to achieve a score on the Test of English as a Foreign Language (TOEFL) that would allow it entrance into most U.S. colleges (Landauer & Dumais). A critically important insight from the TOEFL simulation was that the model's performance peaked at the reduced 300 dimensions compared to fewer or even the full dimensionality of the Σ matrix. Even though the ability of the model (from an algebraic perspective) to reconstruct the original M matrix diminishes monotonically as dimensionality is reduced, its ability to simulate the human semantic data was *better* at the reduced dimensionalities. This finding supports the notion that semantic memory may simply be supported by a mental dimension reduction mechanism applied to episodic contexts. The dimension reduction operation brings out higher-order abstractions by glossing over variance that is idiosyncratic to specific contexts. The astute reader will note the similarity of this notion to the emergent behavior of the hidden layers of a connectionist network that also performs some dimensional reduction operation; we will return to this similarity in the discussion.

The influence of LSA on the field of semantic modeling cannot be overstated. Several criticisms of the model have emerged over the years (see Perfetti, 1998), including the lack of incremental learning, neglect of word-order information, issues about what exact cognitive mechanisms would perform SVD, and concerns over its core assumption that meaning can be represented as a point in space. However, LSA clearly paved the way for a rapid sequence of advances in semantic models in the years since its publication.

4.2 Moving Window Models

An alternative approach to learning distributional semantics is to slide an *N*-word window across a text corpus, and to apply some lexical association function to the co-occurrence counts within the window at each step. While LSA represents a word's episodic context as a document, moving-window models operationalize a word's context in terms of the other words that it is commonly seen with in temporal contexts. Compared to LSA's batch learning mechanism, this allows moving-window models to gradually develop semantic structure from simple co-occurrence counting (cf. Hebbian learning) as a text corpus is experienced in a continuous fashion. In addition, several of these models inversely weight co-occurrence by how many words intervene between a target word and its associate, allowing them to capitalize on word-order information.

The prototypical exemplar of a moving-window model is the Hyperspace Analogue to Language model (HAL; Lund & Burgess, 1996). In HAL, a co-occurrence window (typically, the 10 words preceding and succeeding the target word) is slid across a text corpus, and a global word-by-word co-occurrence matrix is updated at each one-word increment of the window. HAL uses a ramped window in which co-occurrence magnitudes are weighted inversely proportional to distance from the target word. A word's semantic representation in the model is simply a concatenation of its row and column vectors from the global co-occurrence matrix. The row and column vectors reflect the weighted frequency with which each word preceded and succeeded, respectively, the target word in the corpus. Obviously, the word vectors in HAL are both high dimensional and very sparse. Hence, it is common to only use the column vectors with the highest variance (typically about 10% of all words are then retained as 'context' words; Lund & Burgess). Considering its simplicity, HAL has been very successful at accounting for human behavior in semantic tasks, including semantic priming (Lund & Burgess, 1996), and asymmetric

semantic similarity as well as higher-order tasks such as problem solving (Burgess & Lund, 2000).

In HAL, words are most similar if they have appeared in similar positions relative to other words (paradigmatic similarity; e.g., *bee-wasp*). In fact, Burgess and Lund (2000) have suggested that the structure learned by HAL is very similar to what an SRN (Elman, 1990) would learn if it could scale up to such a large linguistic dataset. In contrast, it is known that LSA gives stronger weight to syntagmatic relations (e.g., *bee-honey*) than does HAL, since LSA ignores word order, and both types of similarity are important factors in human semantic representation (Jones, Kintsch, & Mewhort, 2006).

Several recent modifications to HAL have produced models with state-of-the-art performance at simulating human data. One concern in the original model was that chance frequencies can produce spurious similarities in the global matrix: A higher frequency word has a greater chance of randomly occurring with any other word and, hence, high-frequency words end up being more semantically similar to a target independent of semantic similarity. Recent versions of HAL, such as Hidex (Shaoul & Westbury, 2006) factor out chance occurrence by weighting co-occurrence by inverse frequency of the target word, which is similar to LSA's application of log-entropy weighting, but after learning the matrix. A second modification to HAL was proposed by Rohde, Gonnerman, and Plaut (2005) in their COALS model (Correlated Occurrence Analogue to Lexical Semantics). In COALS, there is no preceding/succeeding distinction within the moving window, and the model uses a co-occurrence association function based on Pearson's correlation to factor out the confounding of chance co-occurrence due to frequency. Hence, the similarity between two words is their normalized covariational pattern over all context words. In addition, COALS performs SVD on this matrix. Although these are

quite straightforward modifications to HAL, COALS heavily outperforms its predecessor on human tasks such as semantic categorization (Riordan & Jones, 2011).

A similar moving window model was used by McDonald and Lowe (1998) to simulate semantic priming. In their model, there is no predecessor/successor distinction, but all words are simply represented by their co-occurrence in the moving window with a small number of predefined "context words." While many applications of HAL tabulate the entire matrix and then discard the 90% of column vectors with the least amount of variance, McDonald and Lowe's context word approach specifies the context words (columns) a priori, and tabulates row vectors for each target word but only in relation to the predefined context words. This context word approach, where as few as 100 context words are used as the columns, has also been successfully used by Mitchell et al. (2008) to predict fMRI brain activity associated with humans making semantic judgments about nouns. Slightly more modern versions of these context-word models use log likelihood or log odds rather than raw co-occurrence frequency as matrix elements (Lowe & McDonald, 2000), and some even apply SVD to the word-by-word matrix (e.g., Budi, Royer, & Pirolli, 2007) to bring out latent word relationships.

Moving window models such as HAL have surprised the field with the array of "deep" semantic tasks they can explain with relatively simple learning algorithms based on counting repetitions. They also tie large-scale models of statistical semantics with other learning models such as compound cuing (McKoon & Ratcliff, 1992) and cross-situational word learning (Smith & Yu, 2008).

4.3 Random Vector Models

A entirely different take on contextual representation is seen in models that use random representations for words that gradually develop semantic structure through repeated episodes of the word in a text corpus. The mechanisms used by these models are theoretically tied to

mathematical models of associative memory. For this reason, random vector models tend to capitalize on both contextual co-occurrence as LSA does, and also associative position relative to other words as models like HAL and COALS do, representing both in a composite vector space.

In the Bound Encoding of the Aggregate Language Environment model (BEAGLE; Jones & Mewhort, 2007), semantic representations are gradually acquired as text is experienced in sentence chunks. The model is based heavily on mechanisms from Murdock's (1982) theory of item and associative memory. The first time a word is encountered, it is assigned a random initial vector known as its environmental vector, e_i. This vector is the same each time the word is experienced in the text corpus, and is assumed to represent the relatively stable physical characteristics of perceiving the word (e.g., its visual form or sound). The random vector assumption is obviously an oversimplification, assuming that all words are equally similar to one another in their environmental form (e.g., *dog* is as similar to *dug* as it is to *carburetor*), but see Cox, Kachergis, Recchia, and Jones (2010) for a version of the model that builds in preexisting orthographic structure.

In BEAGLE, each time a word is experienced in the corpus, its memory vector, m_i, is updated as the sum of the random environmental vectors for the other words that occurred in context with it, ignoring high-frequency function words. Hence, in the short phrase "A dog bit the mailman," the memory representation for *dog* is updated as $m_{dog} = e_{bit} + e_{mailman}$. In the same sentence, $m_{bit} = e_{dog} + e_{mailman}$ and $m_{mailman} = e_{dog} + e_{bit}$ are encoded. Even though the environmental vectors are random, the memory vectors for each word in the phrase have some of the same random environmental structure summed into their memory representations. Hence, m_{dog}, m_{bit}, and $m_{mailman}$ all move closer to one another in memory space each time they directly co-occur in contexts. In addition, latent similarity naturally emerges in the memory matrix—even if *dog* and *pitbull* never directly co-occur with each other, they will become

similar in memory space if they tend to occur with the same words (i.e., similar contexts). This allows higher-order abstraction, achieved in LSA by SVD, to emerge in BEAGLE naturally from simple Hebbian summation. Rather than reducing dimensionality after constructing a matrix, BEAGLE sets dimensionality a priori, and the semantic information is distributed across dimensions evenly. If fewer or more dimensions are selected (provided a critical mass is used), the information is simply distributed over fewer or more dimensions. Multiple runs of a model on the same corpus may produce very different vectors (unlike LSA or HAL), but the overall similarity structure of the memory matrix on multiple runs will be remarkably similar. In this sense, BEAGLE has considerable similarity to unsupervised connectionist models.

The use of random environmental representations allows BEAGLE to learn information as would LSA, but in a continuous fashion and without the need for SVD. But the most interesting aspect of the model is that the random representations allow the model to encode word order information in parallel by applying an operation from signal processing known as convolution to bind together vectors for words in sequence. Convolution-based memory models have been very successful as models of both vision and paired-associate memory, and BEAGLE extends this mechanism to encode n-gram chunk information in the word's representation. The model uses circular convolution, which binds together two vectors, with dimensionality *n*, into a third vector of the same dimensionality:

$$\text{for } i = 0 \text{ to } n-1: \qquad z_i = \sum_{j=0}^{n-1} x_{j \bmod_n} * y_{(i-j) \bmod_n} \tag{2}$$

BEAGLE applies this operation recursively to create an order vector representing all the environmental vectors that occur in sequences around the target word, and this order vector is also summed into the word's memory vector. Hence, the memory vector becomes a pattern of elements that reflects the word's history of co-occurrence with, and position relative to, other

words in sentences. Words that appear in similar contexts and similar syntactic roles within sentences will become progressively more similar. Jones, et al. (2006) have demonstrated how this integration of context and order information in a single vector representation allows the model to better account for patterns in semantic priming data.

An additional benefit of having order information encoded in a word's memory vector is that the convolution mechanism used to encode sequence information may be inverted to decode sequential expectancies for a word from its learned history. This decoding operates in a similar fashion to how Murdock (1982) retrieves an associated target given a cue in paired-associate learning. The model can make inferences about likely transitions preceding or following a word and can build up expectancies for which words should be upcoming in sentence processing tasks using the same associative mechanism it uses for learning (see Jones & Mewhort, 2007). Although it only learns lexical semantic structure, BEAGLE naturally displays complex rule-like syntactic behavior as an emergent property of its lexicon. Further, it draws a theoretical bridge between models of lexical semantics and associative memory suggesting that they may be based on the same cognitive mechanisms.

A similar approach to BEAGLE, known as *random indexing*, has been taken by Kanerva and colleagues (Kanerva, 2009; Kanerva, Kristoferson, & Holst, 2000). Random indexing uses similar principles to BEAGLE's summation of random environmental vectors, but is based on Kanerva's (1988) theory of sparse distributed memory. The initial vector for a word in random indexing is a sparse binary representation: A very high dimensional vector in which most elements are zeros with a small number of random elements switched to ones (a.k.a., a "spatter code"). A word's memory representation is then a sum of initial vectors for the other words that it has appeared in contexts with. Words that are semantically similar will tend to be additive on

the random elements that they share nonzero values on, which leads to a similarity structure remarkably similar to LSA, but without the need for SVD.

Sahlgren, Holst, & Kanerva (2008) have extended random indexing to encode order information as does BEAGLE in their Random Permutation Model (RPM). RPM encodes contextual information the same way as standard random indexing. Rather than convolution, it uses a permutation function to encode the order of words around a target word. The permutation function may be applied recursively to encode multiple words at multiple positions around the target word, and this order vector is also added to the word's memory representation. Like BEAGLE, a word's memory vector is a distributed pattern that contains information about its co-occurrence with and position relative to other words. However, in RPM this representation is a sparse hyperdimensional vector, which contains less noise than does BEAGLE's dense Gaussian vectors. In comparative simulations, RPM has been shown to outperform BEAGLE on simple associative tasks (Recchia, Jones, Sahlgren, & Kanerva, 2010).

Howard and colleagues (e.g., Howard, Shakar, & Jagadisan, 2011) have taken a different approach to learning semantic representations, binding local item representations to a gradually changing representation of context by modifying the Temporal Context Model (TCM; Howard & Kahana, 2002) to learn semantic information from a text corpus. TCM uses static vectors representing word form, similar to RPM's initial vectors or BEAGLE's environmental vectors. However, the model binds words to temporal context, a representation that changes gradually with time, similar to oscillator-based systems. In this sense, the model is heavily inspired by hippocampal function. Encountering a word reinstates its previous temporal contexts when encoding its current state in the corpus. Hence, while LSA, HAL, and BEAGLE all treat context as a categorical measure (documents, windows, and sentences, respectively, are completely different contexts), TCM treats context as a continuous measure that is gradually changing over

time. In addition, while all the aforementioned models are essentially batch learners or ignore previous semantic learning when encoding a word, a word's learned history in TCM contributes to its future representation. This is a unique and important feature of TCM compared to other models.

Howard et al. (2011) trained a predictive version of TCM (pTCM) on a text corpus to compare to established semantic models. pTCM continuously attempts to predict upcoming words based on reinstated temporal context. In this sense, the model has many features in common with both BEAGLE and SRNs (Elman, 1990), allowing it to represent both context and order information within the same composite representation. Howard et al. demonstrate impressive performance from pTCM on linguistic association tasks. In addition, the application of TCM in general to semantic representation makes a formal link to mechanisms of episodic memory (which at its core, TCM is) as well as findings in cognitive neuroscience (see Polyn & Kahana, 2008).

4.4 Probabilistic Topic Models

Considerable attention in the cognitive modeling literature has recently been placed on Bayesian models of cognition (see Austerweil, et al., this volume), and mechanisms of Bayesian inference have been successfully extended to semantic memory as well. Probabilistic *topic models* (Blei, Ng, & Jordan, 2003; Griffiths, Steyvers, & Tenenbaum, 2007) operate in a similar fashion to LSA, performing statistical inference to reduce the dimensionality of a term-by-document matrix. However, the theoretical mechanisms behind the inference and representation in topic models differ markedly from LSA and other spatial models.

An assumption of a topic model is that documents are generated by mixtures of latent "topics," where a topic is a probability distribution over words. While LSA makes a similar assumption that latent semantic components can be inferred from observable co-occurrences

across documents, topic models go a step further, specifying a fully generative model for documents (a procedure by which documents may be generated). The assumption is that when constructing documents, humans are sampling a distribution over universal latent topics. For example, one might construct a document about a recent beetle infestation by mixing topics about insects, forests, the ecosystem, etc., with varying weights. To generate each word within this document, one samples a topic according to the document's mixture weights, and then samples words from that topic's probability distribution over words.

To train the model, Bayesian inference is used to reverse the generative process: Assuming that topic mixing is what generates documents, the task of the model is to invert the process and statistically infer the set of topics that were responsible for generating a given set of documents. The formal instantiation of a topic model can be technically intimidating to the novice modeler—based on Latent Dirichlet Allocation algorithms, Markov Chain Monte Carlo algorithms etc. (see Griffiths et al., 2007; Griffiths, Steyvers, Blei, & Tenenbaum, 2005). But it is important to note that the theoretical ideas underlying the model are actually quite simple and elegant and are based on the same ideas posited for how children infer unseen causes for observable events (Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

Consider the analogy of a dermatologist: Given that disease X is present, symptoms A, B, and C are expected to manifest. The task of a dermatologist is one of causal inference however—given a set of co-occurring symptoms she must infer the unseen disease or diseases that produced to the observed data. Over many instances of the same co-occurring symptoms, she can infer the likelihood that they are the result of a common cause. The topic model works in an analogous way, but on a much larger scale of inference and with mixtures of causal variables. Given that certain words tend to co-occur in contexts and this pattern is consistent over many contexts, the model infers the likely latent "topics" that are responsible for generating the co-occurrence

patterns, where each document is a probabilistic mixture of these topics. Each topic is a probability distribution over words, and a word's meaning can be captured by the probability that it was generated by each topic (just as each disease would be a probability distribution over symptoms, and a symptom is a probability distribution over possible diseases that generated it).

Figure 3, reproduced from Steyvers and Griffiths (2007), illustrates this process. Assuming that document co-occurrences are being generated by the process on the left, the topic model attempts to statistically infer (on the right) the most likely topics and mixtures that would have generated the observed data. It is important to note that like LSA, topic models tend to assume a simple bag-of-words representation of a document, neglecting word-order information (but see Andrews & Vigliocco, 2010; Griffiths, et al., 2005). Similar to LSA, each document in the original co-occurrence matrix may be reconstructed by determining the document's distribution over N topics (reflecting its gist, g), using this distribution to select a topic for each word w_i, and then generating a word from the distribution of words conditioned on the topic:

$$P(w_i|g) = \sum_{z_i=1}^{N} P(w_i|z_i)P(z_i|g), \quad (3)$$

where w is the distribution of words over topics, and z is the distribution of topics over words. In practice, topic models construct a prior on the degree of mixing of topics in a document, and then estimate the probability distributions of topics over words and documents over topics using Gibbs sampling (Griffiths & Steyvers, 2004).

The probabilistic inference machinery behind topic models results in at least three major differences in topic models when compared to other distributional models. Firstly, as mentioned above, topic models are generative. Secondly, it is often suggested that the topics themselves have a meaningful interpretation, such as finance, medicine, theft, etc., whereas the components of LSA are difficult to interpret, and the components of models like BEAGLE are purposely not

interpretable in isolation from the others. It is important to note, however, that since the number of topics (and the value of the priors) is set a priori by the theorist, there is often a considerable amount of hand-fitting and intuition that can go into constructing topics that are meaningful (similar to 'labeling' factors in a factor analysis). Thirdly, words in a topic model are represented as probability distributions rather than as points in semantic space—this is a key distinction between topic models and the above spatial models. It allows topic models to naturally display asymmetric associations, which are commonly seen in free association data but require additional assumptions to explain with spatial models (Griffiths, et al., 2007; but see Jones, Gruenenfelder, & Recchia, 2011). Representing a word's meaning with a probability distribution also naturally allows for polysemy in the representation compared to vector representation models that collapse multiple meanings to a single point. For these reasons, topic models have been shown to produce better fits to free association data than LSA, and they are able to account for disambiguation, word-prediction, and discourse effects that are problematic for LSA (Griffiths et al., 2007).

4.5 Retrieval-Based Semantics

Kwantes (2005) proposed an alternative approach to modeling semantic memory from distributional structure. Although not named in his publication, Kwantes' model is commonly referred to as the *constructed semantics model* (CSM), a name that is paradoxical given that the model posits that there is no such thing as semantic memory. Rather, semantic behavior exhibited by the model is an emergent artifact of retrieval from episodic memory. While all other models put the semantic abstraction mechanism at encoding (e.g., SVD, Bayesian inference, vector summation), CSM actually encodes the episodic matrix and performs abstraction as needed when a word is encountered.

CSM is based heavily on Hintzman's (1986) Minerva 2 model which was used as an existence proof that a variety of behavioral effects that had been used to argue for two distinct memory stores (episodic and semantic) could naturally be produced by a model that only had memory for episodes. So-called "prototype effects" were simply an artifact of averaging at retrieval in the model, not necessarily evidence of a semantic store. CSM extends Minerva 2 almost exactly to a text corpus. In CSM, the memory matrix is the term-by-document matrix (i.e., it assumes perfect memory of episodes). When a word is encountered in the environment, its semantic representation is constructed as an average of the episodic memories of all other words in memory, weighted by their contextual similarity to the target. The result is a vector that has higher-order semantic similarities accumulated from the lexicon. This semantic vector is similar in structure to the memory vector learned in BEAGLE by context averaging, but the averaging is done on the fly, it is not encoded or stored.

Although retrieval-based models have received less attention in the literature than models like LSA, they represent a very important link to other instance-based models, especially exemplar models of recognition memory and categorization (e.g., Nosofsky, 1986). The primary reason limiting their uptake in model applications is likely due to the heavy computational expense required to actually simulate their process (Stone, Dennis, & Kwantes, 2011).

5. GROUNDING SEMANTIC MODELS

Semantic models, particularly distributional models, have been criticized as psychologically implausible because they learn from only linguistic information and do not contain information about sensorimotor perception contrary to the grounded cognition movement (for a review, see de Vega, Glenberg, & Graesser, 2008). Hence, representations in distributional models are not a replacement for feature norms. Feature-based representations contain a great deal of sensorimotor properties of words that cannot be learned from purely linguistic input, and

both types of information are core to human semantic representation (Louwerse, 2008). Riordan and Jones (2011) recently compared a variety of feature-based and distributional models on semantic clustering tasks. Their results demonstrated that whereas there is information about word meaning redundantly coded in both feature norms and linguistic data, each has its own unique variance and the two information sources serve as complimentary cues to meaning.

Research using recurrent networks trained on child-directed speech corpora has found that pretraining a network with features related to children's sensorimotor experience produced significantly better word learning when subsequently trained on linguistic data (Howell, Jankowicz, & Becker, 2005). Durda, Buchanan, and Caron (2009) trained a feedforward network to associate LSA-type semantic vectors with their corresponding activation of features from McRae et al.'s (2005) norms. Given the semantic representation for *dog,* the model attempts to activate correct output features such as <has fur> and inhibit incorrect output features such as <made of metal>. After training, the network was able to infer the correct pattern of perceptual features for words that were not used in training because of their linguistic similarity to words that were learned.

Several recent probabilistic topic models have also explored parallel learning of linguistic and featural information (Andrews, Vigliocco, & Vinson, 2009; Baroni, Murphy, Barba, & Poesio, 2010; Steyvers, 2009). Given a word-by-document representation of a text corpus and a word-by-feature representation of feature production norms, the models learn a word's meaning by simultaneously considering inference across documents and features. This enables learning from joint distributional information: If the model learns from the feature norms that *sparrows* have beaks, and from linguistic experience that *sparrows* and *mockingbirds* are distributionally similar, it will infer that *mockingbirds* also have beaks, despite having no feature vector for *mockingbird*. Integration of linguistic and sensorimotor information allows the models to better

fit human semantic data than a model trained with only one source (Andrews et al., 2009). This information integration is not unique to Bayesian models, but can also be accomplished within random vector models (Jones & Recchia, 2010; Vigliocco, Vinson, Lewis, & Garrett, 2004) and retrieval-based models (Johns & Jones, 2012).

6. COMPOSITIONAL SEMANTICS

The models we have considered thus far are designed to extract the meaning of individual terms. However, the sentence "John loves Mary" is not just the sum of the words it contains. Rather "John" is bound to the role LOVER and "Mary" is bound to the role LOVEE. The study of how sentence structure determines these bindings is called compositional semantics. Recent work has begun to explore mechanisms for compositional semantics by applying learning mechanisms to the already learned lexicon of a distributional model (Mitchell & Lapata, 2010).

Dennis (2004, 2005) argued that extracting propositional structure from sentences revolves around the distinction between syntagmatic and paradigmatic associations. Syntagmatic associations occur between words that appear together in utterances (e.g. run fast). Paradigmatic associations occur between words that appear in similar contexts, but not necessarily in the same utterances (e.g. deep and shallow). The *syntamatic paradigmatic model* proposes that syntagmatic associations are used to determine which words could have filled a particular slot within a sentence. The set of these words form role vectors which are then bound to fillers by paradigmatic associations to form a propositional representation of the sentence. The syntagmatic paradigmatic mechanism has been shown to be capable of accounting for a wide range of sentence processing phenomena. Furthermore, it is capable of taking advantage of regularities in the overlap of role patterns to create implicit inferences that Dennis (2005) claimed are responsible for the robustness of human commonsense reasoning.

7. COMMON LESSONS AND FUTURE DIRECTIONS

Models of semantic memory have seen impressive developments over the past two decades that have greatly advanced our understanding of how humans create, represent, and use meaning from experience. These developments are thanks in part to advances in other areas, such as machine learning, and to better large-scale norms of semantic data on which to fit and compare the models. In general, distributional models have been successfully used to better explore the statistical structure of the environment and to understand the mechanisms that may be used to construct semantic representations. Connectionist models are an excellent compliment, elucidating our understanding of semantic processing, and how semantic structure interacts with other cognitive systems and tasks. An obvious and important requirement for the future is to start to bring these insights together, and several hybrid models are now emerging in the literature.

Several important themes have emerged that are common to both the connectionist and distributional literatures. The first is the clear is the importance of data reduction. Whatever specific mechanism humans are using to construct conceptual and propositional knowledge from experience, it is likely that this mechanism learns by focusing on important statistical factors that are constant across contexts, and by throwing away factors that are idiosyncratic to specific contexts. In a sense, capacity constraints on human encoding, storage, and retrieval may give rise to our incredible ability to construct and use meaning.

A second common theme is the importance of data scale in semantic modeling. In both connectionist and distributional models, the issue of data scale vs. mechanistic complexity has been brought to the forefront of discussion in the literature. A consistent emerging theme is that simpler models tend to give the best explanation of human data, both in terms of parsimony and quantitative fit to the data, when they are trained on linguistic data that is on a realistic scale to what humans experience. For example, although simple context-word moving window models

are considerably simpler than LSA and do not perform well at small data scales, they are capable of scaling up to learn from human-scale amounts of linguistic data (a scale not necessarily possible to learn with LSA), and consistently outperform more complex models such as LSA with large data (e.g., Louwerse, 2011; Recchia & Jones, 2009). This leads to potential concern that earlier theoretical advancements with models trained on so-called 'toy datasets' (artificial language corpora constructed to test the model's structural learning) may have been overly complex. To fit human behavioral data with a corpus that is far smaller and without the deep complexities inherent in real language, the model must necessarily be building complexity into the architecture and mechanism whereas humans may be using a considerably simpler mechanism, offloading considerable statistical complexity already present in their linguistic environment.

A third common theme is that complex semantic structures and behaviors may be an emergent property of the lexicon. Emergence is a key property of connectionist models, and we have seen that complex representations of schemata, hierarchical categories, and syntactic processing may be emergent consequences of many connectionist models (e.g., Rogers & McClelland, 2007). But emergence is also a natural consequence of distributional models. In several cases, the same mechanisms used to learn semantic representations may be applied to the learned representations to simulate complex behaviors, such as BEAGLE's ability to model sentence comprehension as an emergent property of order information distributed across the lexicon (Jones & Mewhort, 2007). Topic models also possess a natural mechanism for producing asymmetric similarity and polysemous processing through conditional inference.

Learning to organize the mental lexicon is one of the most important cognitive functions across development, laying the fundamental structure for future semantic learning and communicative behavior. Semantic modeling has a very promising future, with potential to

further our understanding of basic cognitive mechanisms that give rise to complex meaning structures, and how these mental representations are used in a wide range of higher-order cognitive tasks.

GLOSSARY

Compositional Semantics: The process by which a complex expression (e.g., a phrase or sentence) is constructed from the meanings of its constituent concepts.

Concept: A mental representation generalized from particular instances, and knowledge of its similarity to other concepts.

Connectionist Model: A model that represents knowledge as weighted network of interconnected units. Behavioral phenomena are an emergent process of the full network.

Context: In semantic models, context is typically considered the 'window' within which two words may be considered to co-occur, and it is one of the major theoretical differences between distributional models. Context may be considered to be discrete units, such as sentences or documents, or it may be more continuous, such as in moving-window or temporal context models.

Distributional Model: A general approach to concept learning and representation from statistical redundancies in the environment.

Dynamic Network: A connectionist network whose architecture involves bi-directionality, feedback, or recurrent connectivity.

Feature Comparison Model: A classic model of semantic memory that represents concepts as vectors of binary features representing the presence or absence of features. For example, the *has_wings* element would be turned on for robin, but off for golden retriever.

Paradigmatic and Syntagmatic Relations: Paradigmatic similarity between two words emphasizes their synonymy or substitutability (*bee-wasp*), whereas syntagmatic similarity emphasizes associative or event relations (e.g., *bee-honey*, *wasp-sting*).

Proposition: A mental representation of conceptual relations that may be evaluated to have a truth-value. For example, knowledge that birds have wings.

Random Vector Model: A semantic model that begins with some sort of randomly generated vector to initially represent a concept. Over linguistic experience, an aggregating function gradually produces similar vector patterns among words that are semantically related. They allow for study of the time course of semantic acquisition.

Semantic Memory: Memory for word meanings, facts, concepts, and general world knowledge. Typically not tied to a specific event.

Semantic Network: A classic graphical model of semantic memory that represents concepts as nodes and semantic relations as labeled edges between the nodes. Often, the hypothetical process of *spreading activation* is used to simulate behavioral data such as semantic priming from a semantic network.

Singular-Value Decomposition: A statistical method of factorizing an *m x n* matrix, **M**, into an *m x m* unitary matrix, **U**, an *m x n* diagonal matrix, **Σ**, with diagonal entries that are the singular values, and an *n x n* unitary matrix, **V**. The original matrix may be recomposed as $\mathbf{M} = \mathbf{U\Sigma V^T}$, where $\mathbf{V^T}$ is the transpose of **V**.

Spatial Model: A model that represents word meaning as a point in a multidimensional space, and that typically applies a geometric function to express conceptual similarity.

Supervised and Unsupervised Learning: Supervised learning typically trains the model on a set of labeled exemplars (i.e., the true output of each training exemplar is known), whereas in unsupervised learning the model must discover structure in the data without the benefit of known labels.

Topic Model: A generative probabilistic model that uses Bayesian inference to abstract the mental "topics" used to compose a set of documents.

ACKNOWLEDGEMENTS

This work was supported by National Science Foundation grant BCS-1056744 and National Institutes of Health grant R01MH096906 to MNJ, and Defence Research and Development Canada grant W7711-067985 to SD. JW was supported by postdoctoral training grant NIH T32 DC000012.

REFERENCES

Aakerlund, L., Hemmingsen, R. (1998). Neural networks as models of psychopathology. *Biological Psychiatry, 43*, 471-82.

Allen, J., & Seidenberg, M. S. (1999). The emergence of grammaticality in connectionist networks. In B. MacWhinney. (Ed.), *Emergentist Approaches to Language: proceedings of the 28th Carnegie symposium on cognition* (pp. 115-151). Hillsdale, NJ: Lawrence Earlbaum Associates.

Andrews, M., & Vigliocco, G. (2010). The hidden-Markov topic model: A probabilistic model of semantic representation. *Topics in Cognitive Science, 2,* 101-113.

Andrews, M., Vigliocco, G. & Vinson, D. P. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review, 116*, 463-498.

Baroni, M., Murphy, B., Barbu, E., & Poesio, M. (2010). Strudel: A corpus-based semantic model based on properties and types. *Cognitive Science 34,* 222-254.

Botvinick, M., & Plaut, D. (2004). Doing without schema hierarchies: A recurrent connectionist approach to normal and impaired sequential action. *Psychological Review, 111,* 395-429

Braver, T. S., Barch, D. M., Cohen, J. D. (1999). Cognition and control in schizophrenia: a computational model of dopamine and prefrontal function. *Biological Psychiatry,* 46, 312-28.

Bower, G. H. (1970). Organizational factors in memory. *Cognitive Psychology, 1,* 18-46.

Budiu, R., Royer, C., Pirolli, P. L. (2007). Modeling information scent: a comparison of LSA, PMI and GLSA similarity measures on common tests and corpora. *Proceedings of Recherche d'Information Assistée par Ordinateur* (RIAO). Pittsburgh, PA, 8.

Bullinaria, J., A., & Levy, J., P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods, 39*, 510-526.

Burgess, C. & Lund, K. (2000) The dynamics of meaning in memory. In E. Dietrich & A. B. Markman (Eds.), *Cognitive dynamics: Conceptual and representational change in humans and machines* (pp. 117–156). Mahwah, NJ: Erlbaum.

Chang, F., Dell, G. S., & Bock, K. (2006). Becoming syntactic. *Psychological Review,* 113, 2, 234-272.

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

Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning & Verbal Behavior, 8*(2), 240-247.

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

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

De Vega, M., Glenberg, A. M., & Graesser, A. C. (2008). *Symbols and embodiment: debates on meaning and cognition*. Oxford, UK: Oxford University Press.

Dell, G. S., Chang, F., & Griffin, Z. M. (1999). Connectionist models of language production: Lexical access and grammatical encoding. *Cognitive Science, 23(*4), 517-542.

Dell, G. S., Juliano, C., & Govindjee, A. (1993). Structure and content in language production: A theory of frame constraints in phonological speech errors. *Cognitive Science, 17,* 149–195.

Dennis, S. (2004). An unsupervised method for the extraction of propositional information from text. *Proceedings of the National Academy of Sciences of the United States of America, 101*(Suppl 1), 5206-5213.

Dennis, S. (2005). A memory-based theory of verbal cognition. *Cognitive Science, 29*, 145-193.

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*, 77-94.

Durda, K., Buchanan, L., & Caron, R. (2009). Grounding co-occurrence: Identifying features in a lexical co-occurrence model of semantic memory. *Behavior Research Methods, 41*, 1210-1223.

Elman, J. (1990). Finding structure in time. *Cognitive Science, 14,* 179-211.

Farah, M. J., & 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.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. In Philological Society (Great Britain) (Ed.), *Studies in linguistic analysis*. Oxford: Blackwell.

Griffiths, T. L., & Steyvers, M. (2003). 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,* 537-544.

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

Grossberg, S. (1976). Adaptive pattern classification and universal recoding, I: Parallel development and coding of neural feature detectors. *Biological Cybernetics*, *23*, 121-134.

Harm, M. W., & Seidenberg, M. S. (2004). Computing the Meanings of Words in Reading: Cooperative Division of Labor Between Visual and Phonological Processes. *Psychological Review, 111*, 662-720.

Harris, Z. (1970). Distributional structure. In *Papers in Structural and Transformational Linguistics* (pp. 775-794). Dordrecht, Holland: D. Reidel Publishing Company.

Hebb, D. (1946). The Organization of Learning.

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

Horst, J.S., McMurray, B., and Samuelson, L.K. (2006) Online processing is essential for learning: Understanding fast mapping and word learning in a dynamic connectionist architecture. In R. Sun (Ed) *Proceedings of the 28th meeting of the Cognitive Science Society* (pp. 339-334).

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, *79*, 2554–2558.

Howard, M. W., & Kahana, M. J. (2002). A distributed representation of temporal context. *Journal of Mathematical Psychology, 46,* 269-299.

Howard, M. W., Shankar, K. H., & Jagadisan, U. K. (2010). Constructing semantic representations from a gradually changing representation of temporal context. *Topics in Cognitive Science, 3,* 48-73.

Howell, S., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language, 53*, 258-276.

Hummel, J. E., & Holyoak, K. J. (2003). A symbolic-connectionist theory of relational inference and generalization. *Psychological Review, 110*, 220-264.

Johns, B. T., & Jones, M. N. (2012). Perceptual inference from global lexical similarity. *Topics in Cognitive Science, 4:*1, 103-120.

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*, 1-37.

Kanerva, P. (2009). Hyperdimensional computing: An introduction to computing in distributed representations with high-dimensional random vectors. *Cognitive Computation, 1*, 139-159.

Keil, F. C. (1989). *Concepts, kinds, and cognitive development*. Cambridge, MA: MIT Press.

Kinder, A., & Shanks, D. R. (2003). Neuropsychologicla dissociations between priming and recognition: A single-system connectionist account. *Psychological Review, 110*, 728-744.

Kintsch, W. (1998). Comprehension: A paradigm for cognition. New York: Cambridge University Press.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, *43*, 59-69.

Kwantes, P. J. (2005). Using context to build semantics. *Psy Bull & Review, 12,* 703-710.

Lambon Ralph, M. A., McClelland, J. L., Patterson, K., Galton, C. J., & Hodges, J. R. (2001). The relationship between object naming and semantic impairment: Neuropsychological evidence and a computational model. *Cognitive Neuroscience*. *13*, 341-356.

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

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. 412-417). Mahwah, NJ: Erlbaum.

Lowe, W. & McDonald, S. (2000). The direct route: Mediated priming in semantic space. *Proceedings of the 22th Annual Conference of the Cognitive Science Society*.

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science, 3*, 273-302.

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

Mandler, J. M., Bauer, P. J. & McDonough, L. (1991) Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology*, *23*, 263-298.

McClelland, J. L. & Thompson, R. M. (2007). Using domain-general principles to explain children's causal reasoning abilities. *Developmental Science, 10*(3), 333-356.

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

McDonald, S. & Lowe, W. (1998). Modelling functional priming and the associative boost. *Proceedings of the 20th Annual Conference of the Cognitive Science Society* (pp. 667-680).

McKoon, G., & Ratcliff, R. (1992). Spreading activation versus compound cue accounts of priming: Mediated priming revisited. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 1155-1172.

McLeod, P., Shallice, T., and Plaut, D.C. (2000). Visual and semantic influences in word recognition: Converging evidence from acquired dyslexic patients, normal subjects, and a computational model. *Cognition, 74,* 91-114.

McRae, K., & Cree, G. (2002). Factors underlying category-specific semantic deficits. In E. M. E. Forde, & G. Humphreys (Eds.), *Category-specificity in mind and brain*. East Sussex, UK: Psychology Press.

McRae, K., Cree, G., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods, 37*, 547-59.

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

McRae, K., & Jones, M. N. (in press). Semantic memory. In D. Reisberg (Ed.) *The Oxford Handbook of Cognitive Psychology*.

Mitchell, T. M., Shinkanerva, S. V., Carlson, A., Chang, K., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320,* 1191-1195.

Nestor, P. G., Akdag, S. J., O'Donnell, B. F., Niznikiewicz, M., Law, S., Shenton, M. E., & McCarley, R. W. (1998). Word recall in schizophrenia: a connectionist model. *American Journal of Psychiatry, 155*(12), 1685-1690.

O'Reilly, R. C., Munakata, Y., Frank, M. J., Hazy, T. E., and Contributors (2012). *Computational Cognitive Neuroscience*. Wiki Book, 1st Edition. URL: http://ccnbook.colorado.edu.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Review, 49,* 197-237.

Osgood, C. E. (1971). Explorations in semantic space: A personal diary. *Journal of Social Issues, 27,* 5-62.

Perfetti, C. (1998). The limits of co-occurrence: Tools and theories in language research. *Discourse Processes, 25*, 363-377.

Plaut, D. C. (1999). A connectionist approach to word reading and acquired dyslexia: Extension to sequential processing. *Cognitive Science, 23*, 543-568.

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

Plaut, D. C., and 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.

Recchia, G. L., & Jones, M. N. (2009). More data trumps smarter algorithms: Training computational models of semantics on very large corpora. *Behavior Research Methods, 41*, 657-663.

Regier, T. (2005). The emergence of words: Attentional learning of form and meaning. *Cognitive Science, 29*, 819-865.

Riordan, B., & Jones, M. N. (2011). Redundancy in linguistic and perceptual experience: Comparing distributional and feature-based models of semantic representation. *Topics in Cognitive Science, 3*, 1-43.

Rips, L. J., Shoben, E. J., & Smith, E. E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning & Verbal Behavior, 12,* 1-20.

Rogers, T. T. ,& McClelland, J. L. (2006). *Semantic Cognition*. Cambridge, MA: MIT Press.

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

Rohde, D. L. T., and Plaut, D. C. (1999). Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition, 72,* 67-109.

Rohde, D. L. T., Gonnerman, L., & Plaut, D. C**.** (2009). An improved model of semantic similarity based on lexical co-occurrence. *Cognitive Science*.

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

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

Rumelhart, D. E., McClelland, J. L., & the PDP research group. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition*. *Volume I*. Cambridge, MA: MIT Press.

Rumelhart, D. E. & Todd, P. M. (1993) Learning and connectionist representations. In D. E. Meyer & S. Kornblum's (Eds.) *Attention and performance XIV: Synergies in experimental psychology, artificial intelligence, and cognitive neuroscience*, ed. D. E. Meyer & S. Kornblum, pp. 3–30. MIT Press.

Sahlgren, M., Holst, A., & Kanerva, P. (2008). Permutations as a means to encode order in word space. *Proceedings of the 30th Conference of the Cognitive Science Society*, p. 1300-1305.

Salton, G., & McGill, M. (1983). *Introduction to modern information retrieval*. McGraw-Hill.

Shaoul, C., & Westbury, C. (2006). Word frequency effects in highdimensional co-occurrence models: A new approach. *Behavior Research Methods, 38*, 190–195.

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.

Smith, L., & Yu, C. (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition, 106,* 1558-1568.

Steyvers, M. (2009). Combining feature norms and text data with topic models. *Acta Psychologica, 133*, 234-243.

Steyvers, M., & Griffiths, T. (2008). Probabilistic topic models. In T. Landauer, D. McNamara, S. Dennis, & W. Kintsch (Eds.) *Handbook of latent semantic analysis*. Mahwah, NJ: Erlbaum.

Stone, B., Dennis, S. & Kwantes, P. J. (2011). Comparing methods for single paragraph similarity analysis. *Topics in Cognitive Science. 3* (1). 92-122.

Tabor, W. & Tanenhaus, M.K. (1999). Dynamical theories of sentence processing. *Cognitive Science, 23*, 491-515.

Taraban, R. & McClelland, J. L. (1988). Constituent attachment and thematic role assignment in sentence processing: Influences of content-based expectations. *Journal of Memory & Language, 27*, 597-632.

Thomas, M. S. C., & Karmiloff-Smith, A. (2003). Modeling language acquisition in atypical phenotypes. *Psychological Review, 110*(4), 647-682.

Turney, P. D. and Pantel, P. (2010). From frequency to maning: Vector space models of semantics. *Journal of Artificial Inteligence Research*, 37, 141-188.

Tversky, A. & Gati, I. (1982). Studies of similarity. In E. Rosch & B. Lloyd Eds.), *On the nature and principle of formation of categories* (pp. 79-98). Hillsday, N.J: Erlbaum Press.

Tyler, L. K., Durrant-Peatfield, M. R., Levy, J. P., Voice, J. K., & Moss, H. E. (1996). Distinctiveness and correlations in the structure of categories: Behavioral data and a connectionist model. *Brain and Language, 55*, 89-91.

Vigliocco, G., Vinson, D. P., Lewis, W., & Garrett, M. (2004). Representing the meanings of object and action words: The featural and unitary semantic space hypothesis. *Cognitive Psychology, 48*(4), 422-488.

Widrow, B., & Hoff, M. E., Jr. (1960). Adaptive switching circuits. In 1960 IRE WESCON Convention Record, Part 4 (pp. 96-104). New York: IRE.

Wittgenstein, Ludwig (1953). Philosophical Investigations. Blackwell Publishing.

Table 1. Highly cited semantic models and the specific sub-processes that comprise the models.

	Representational Structure		**Representational Transformation**		
Model	Unit Type	Unit Span	Normalization	Abstraction	**Comparison Process**
HAL	Ordered Word-by-Word Co-occurrence Matrix	Distance Weighted 10-word window	Conditional Probabilities (Matrix Row Sum)	None	City Block Distance Similarity
COALS	Ordered Word-by-Word Co-occurrence Matrix	10-word window	Correlational Normalization	Principle Components Analysis	Correlational Similarity
LSA	Unordered Word-by-Document Count Matrix	Pre-Defined Document	Log Entropy	Singular Value Decomposition	Cosine Similarity
Topic Models	Unordered Word-by-Document Count Matrix	Pre-Defined Document	None	Latent Dirichlet Allocation	Inner Product Similarity
BEAGLE	Ordered Word-by-Word Matrix	Sentence Window	None	Random Vector Accumulation	Cosine Similarity

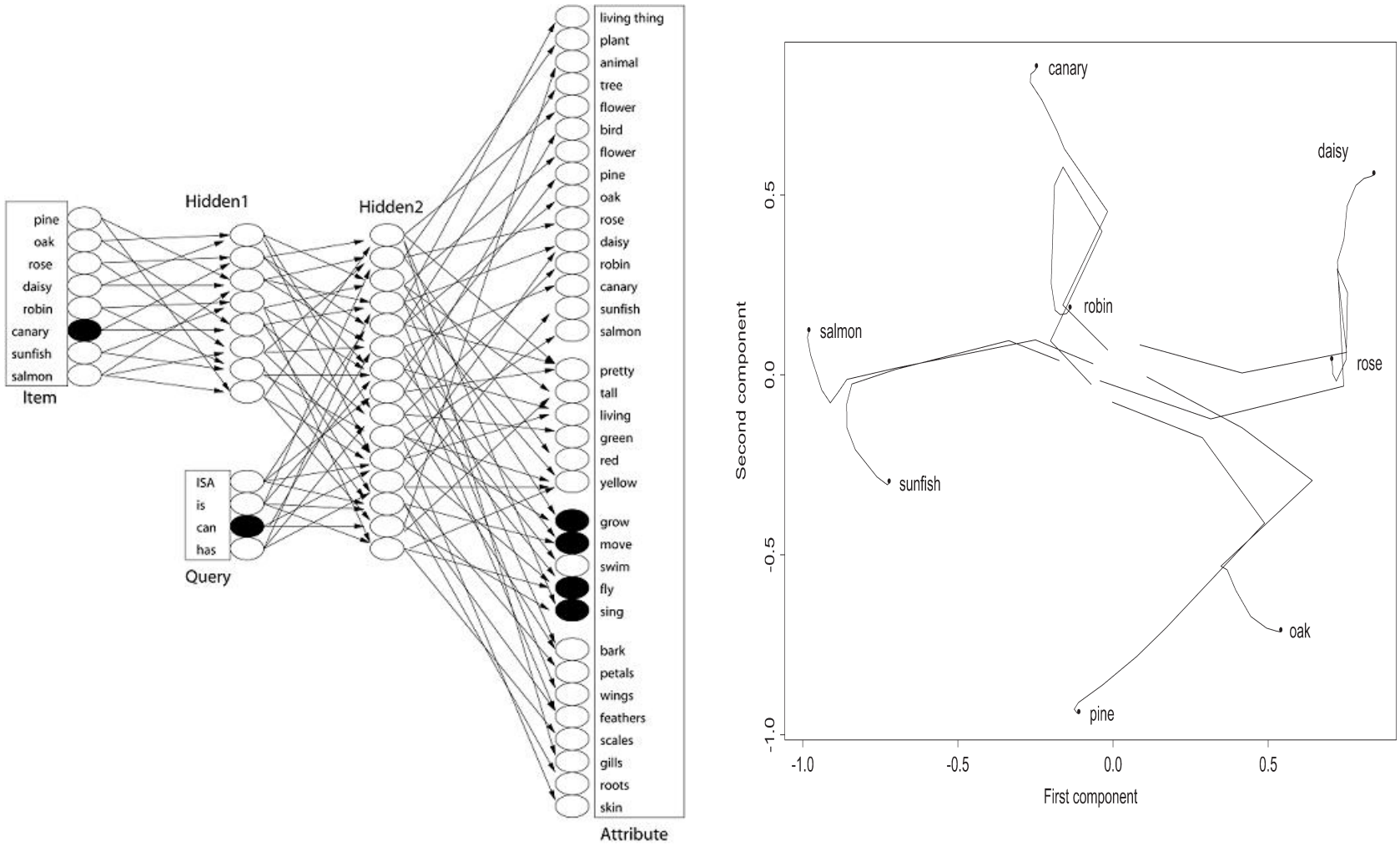

Figure 1A (Left). The network architecture used by Rogers and McClelland (2006), showing the output activation of appropriate semantic features given an input concept and an input relation. **Figure 1B (Right).** A graph of the network's learning trajectory, obtained by performing a multidimensional scaling on the network's hidden unit activations for each input. As the network obtains more experience with each concept, it progressively learns to make finer and finer distinctions between categories.

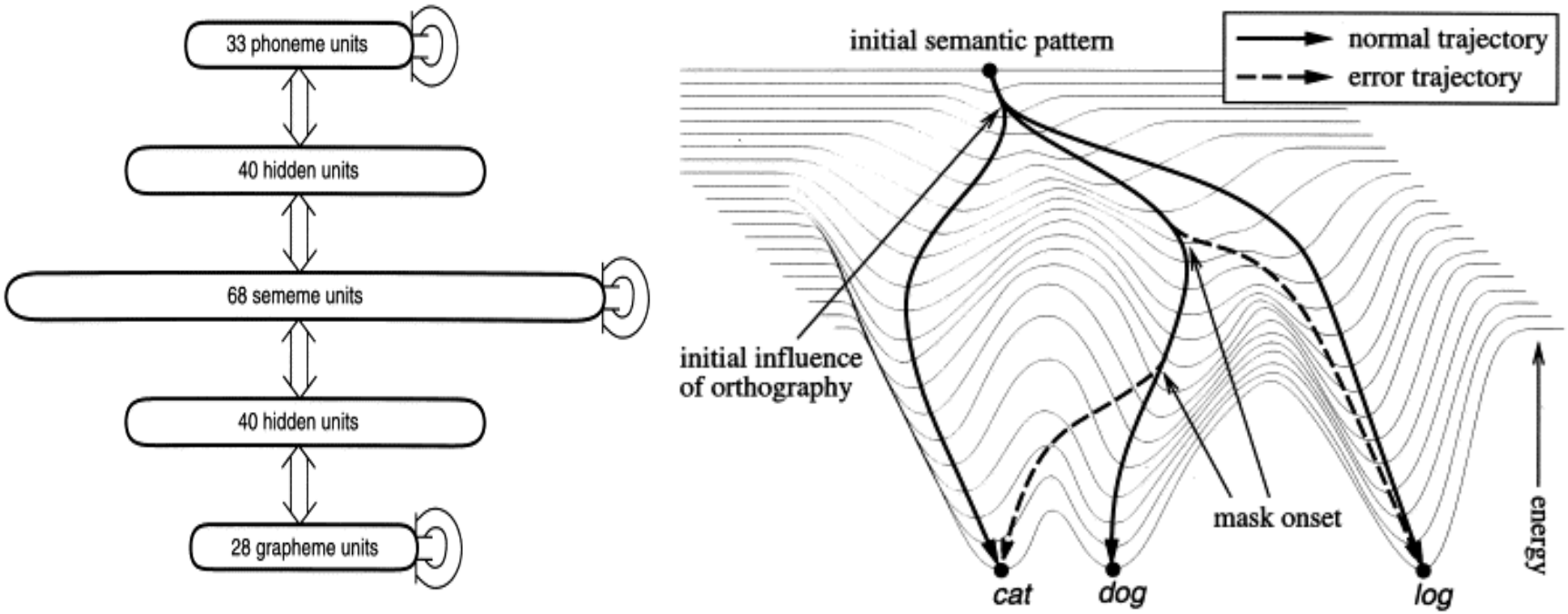

Figure 2a (Left). A prototypical example of a semantic attractor network, from McLeod, Shallice, & Plaut (2000).
Figure 2B (Right). An example of the network's behavior, simulating the experience of a person reading the word "dog". The weights of the network have created a number of attractor spaces, determined by words' orthographic, phonological, and semantic similarity. Disrupting the input (such as presenting the participant with a stimulus mask) at different stages has different effects. Early masking leads to a higher likelihood of falling into the wrong perceptual attractor (LOG instead of DOG). Later masking leads to a higher likelihood of falling into the wrong semantic attractor (CAT instead of DOG).

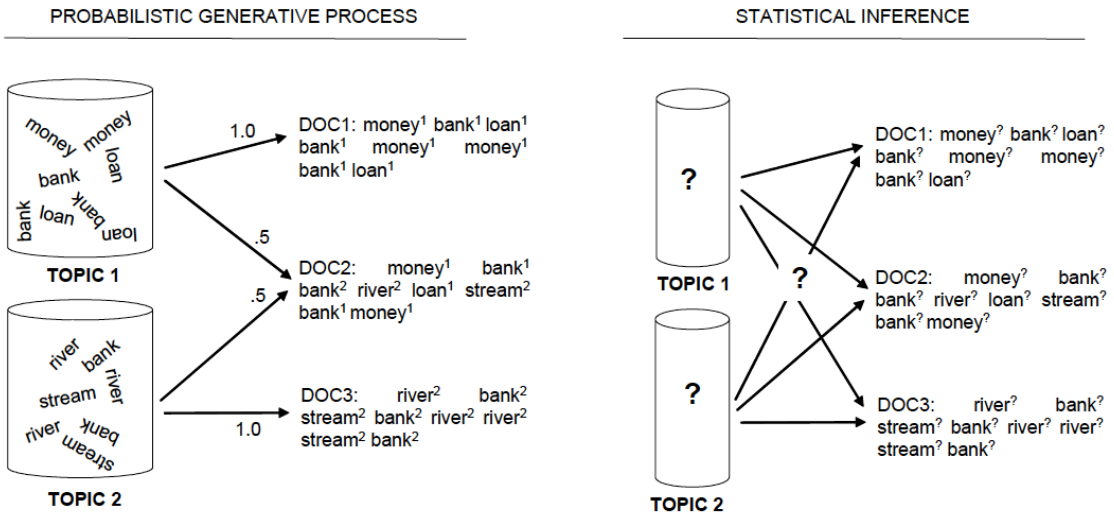

Figure 3. Illustration of the generative process (left) and the problem of statistical inference (right) underlying topic models. (Reproduced from Steyvers & Griffiths, 2007).

Textbox 1: So which model is right?

It is tempting to think of different distributional models as competing "brands." However, a potentially more fruitful approach is to consider each specific model as a point in a parameter space, as one would with other cognitive models. Each model is really just a particular set of decisions made to formalize the distributional theory of "knowing a word by the company it keeps" (Firth, 1957), and no single model has emerged victorious at accounting for the wide range of semantic behavioral data. Each model has its merits and shortcomings.

How should distributional models be compared? If a model is being proposed as a psychological model, it is important to identify the model's sub-processes. How do those sub-processes contribute to how the model works? How are they related to other psychological theories? And how do they contribute to the model's ability to predict behavioral data? For example, LSA and HAL vary in a large number of ways (see Table 1). Studies that perform simple model comparisons end up confounding these differences, leaving us unsure what underlying psychological claim is being tested.

Most model differences can be ascribed to one of three categories, each corresponding to important differences in the underlying psychological theory:

1. Representational Structure: What statistical information does the model pay attention to, and how is this information initially represented?
2. Representational Transformations: By what function are the representations transformed to produce a semantic space?
3. Comparison Process: How is the semantic space queried, and how is the semantic information, relations, or similarity used to model behavioral data?

The HAL model defines its representations in terms of a word-by-word co-occurrence matrix, whereas the LSA model defines its representation in terms of a word-counts-within-documents matrix. This difference corresponds to a long tradition of different psychological theories. HAL's word-word co-occurrences are akin to models that propose representations based on associations between specific stimuli (such as classical associationist theories of learning). In contrast, LSA's word-by-document representation proposes representations based associations between a stimuli and abstract pointers to the event in which those stimuli participate (similar to classic context association theories of learning).

A number of studies have begun comparing model performance as a function of differences in these subprocesses (e.g. Bullinaria & Levy, 2012; Shaoul & Westbury, 2010), but much more research is needed before any firm conclusions can be made.

Box 2: Semantic Memory Modeling Resources

A chapter on semantic models would seem incomplete without some code! Testing models of semantic memory has become much easier due to an increase in semantic modeling resources. There are now a wide variety of software packages that provide the ability to construct and test semantic models. The software packages vary in terms of their easy of installation and use, flexibility, and performance. In addition to the software packages, a limited number of web-based resources exist for doing simple comparisons online. You may test models on standardized datasets, train them on your own corpora for semantic exploration, or use them for generating stimuli.

Software Packages

- **HiDEx** (http://www.psych.ualberta.ca/~westburylab/downloads/HiDEx.download.html): A C++ implementation of the HAL model; it is useful for constructing large word-by-word co-occurrence matrices and testing a wide variety of possible parameters.
- **SuperMatrix** (http://semanticore.org/supermatrix/): A python implementation of a large number of semantic space model transformations (including PCA/SVD, Latent Dirichlet Allocation, and Random Vector Accumulation) on both word-by-word and word-by-document spaces. SuperMatrix was designed to emphasize the exchangeability of various sub-processes within semantic models (see Box 1), to allow isolation and testing the effects of specific model components.
- **GenSim** (http://radimrehurek.com/gensim/): A python module that is very fast and efficient for constructing and testing word-by-document models, including LSA (reduced using SVD) and Topics (reduced using Latent Dirichlet Allocation).
- **S-Space** (https://github.com/fozziethebeat/S-Space): A Java-based implementation of a large number of semantic space models, including HAL, LSA, BEAGLE, and COALS.
- **SEMMOD** (http://mall.psy.ohio-state.edu/wiki/index.php/Semantic_Models_Package_(SEMMOD)): A python package to implement and compare many of the most common semantic models.
- **Word-Similarity** (https://code.google.com/p/wordsimilarity/wiki/train): A tool to explore and visualize semantic spaces, displayed as directed graphical networks.

Web-Based Resources

- http://lsa.colorado.edu: The original LSA website provides the ability to explore Latent Semantic Analysis with a wide variety of different metrics, including word-word similarities, similarities of passages of text to individual words, and similarities of passages of texts to each other.
- http://semanticore.org: The Semanticore website is a web portal designed to bring data from many semantic models and psycholinguistic databases under one roof. Users can obtain frequency and co-occurrence statistics from a wide variety of corpora, as well as semantic similarities from a number of different semantic memory models, including HAL, LSA, BEAGLE, and Probabilistic Topics Models.