Psychonomic Bulletin & Review
https://doi.org/10.3758/s13423-020-01792-x

Semantic memory: A review of methods, models, and current challenges

Abhilasha A. Kumar [1]

Abstract
Adult semantic memory has been traditionally conceptualized as a relatively static memory system that consists of knowledge about the world, concepts, and symbols. Considerable work in the past few decades has challenged this static view of semantic memory, and instead proposed a more fluid and flexible system that is sensitive to context, task demands, and perceptual and sensorimotor information from the environment. This paper (1) reviews traditional and modern computational models of semantic memory, within the umbrella of network (free association-based), feature (property generation norms-based), and distributional semantic (natural language corpora-based) models, (2) discusses the contribution of these models to important debates in the literature regarding knowledge representation (localist vs. distributed representations) and learning (error-free/Hebbian learning vs. error-driven/predictive learning), and (3) evaluates how modern computational models (neural network, retrieval-based, and topic models) are revisiting the traditional "static" conceptualization of semantic memory and tackling important challenges in semantic modeling such as addressing temporal, contextual, and attentional influences, as well as incorporating grounding and compositionality into semantic representations. The review also identifies new challenges regarding the abundance and availability of data, the generalization of semantic models to other languages, and the role of social interaction and collaboration in language learning and development. The concluding section advocates the need for integrating representational accounts of semantic memory with process-based accounts of cognitive behavior, as well as the need for explicit comparisons of computational models to human baselines in semantic tasks to adequately assess their psychological plausibility as models of human semantic memory.

Keywords Semantic memory · Distributional semantic models · Semantic networks · Neural networks · Language models

Introduction

What does it mean to know what an *ostrich* is? The question of how meaning is represented and organized by the human brain has been at the forefront of explorations in philosophy, psychology, linguistics, and computer science for centuries. Does knowing the meaning of an *ostrich* involve having a prototypical representation of an *ostrich* that has been created by averaging over multiple exposures to individual ostriches? Or does it instead involve extracting particular features that are characteristic of an *ostrich* (e.g., it is big, it is a bird, it does not fly, etc.) that are acquired via experience, and stored and activated upon encountering an *ostrich*? Further, is this knowledge stored through abstract and arbitrary symbols such as words, or is it grounded in sensorimotor interactions with the physical environment? The computation of meaning is fundamental to all cognition, and hence it is not surprising that considerable work has attempted to uncover the mechanisms that contribute to the construction of meaning from experience.

There have been several important historical seeds that have laid the groundwork for conceptualizing how meaning is learned and represented. One of the earliest attempts to study how meaning is represented was by Osgood (1952; also see Osgood, Suci, & Tannenbaum, 1957) through the use of the semantic differential technique. Osgood collected participants' ratings of concepts (e.g., *peace*) on several polar scales (e.g., hot-cold, good-bad, etc.), and using multidimensional scaling, showed that these ratings aligned themselves along three universal dimensions: evaluative (good-bad), potency (strong-weak), and activity (active-passive). Osgood's

✉ Abhilasha A. Kumar
abhilasha.kumar@wustl.edu

[1] Department of Psychological and Brain Sciences, Washington University in St. Louis, Campus Box 1125, One Brookings Drive, St. Louis, MO 63130, USA

Published online: 03 September 2020

work was important in the following two ways: (1) it introduced an empirical tool to study the nature of semantic representations; (2) it provided early evidence that the meaning of a concept or word may actually be distributed across several dimensions, in contrast to being represented through a localist representation, i.e., through a single dimension, feature, or node. As subsequently discussed, this contrast between localist and distributed meaning representations has led to different modeling approaches to understanding how meaning is learned and represented.

Another important milestone in the study of meaning was the formalization of the *distributional hypothesis* (Harris, 1970), best captured by the phrase "you shall know a word by the company it keeps" (Firth, 1957), which dates back to Wittgenstein's early intuitions (Wittgenstein, 1953) about meaning representation. The idea behind the distributional hypothesis is that meaning is learned by inferring how words co-occur in natural language. For example, *ostrich* and *egg* may become related because they frequently co-occur in natural language, whereas *ostrich* and *emu* may become related because they co-occur with similar words. This distributional principle has laid the groundwork for several decades of work in modeling the explicit nature of meaning representation. Importantly, despite the fact that several distributional models in the literature do make use of *distributed* representations, it is their learning process of extracting statistical redundancies from natural language that makes them *distributional* in nature.

Another historically significant event in the study of meaning was Tulving's (1972) classic distinction between *episodic* and *semantic* memory. Tulving proposed two subdivisions of declarative memory: *episodic* memory, consisting of memories of experiences linked to specific times and places (e.g., seeing an *ostrich* at the zoo last month), and *semantic* memory, storing general knowledge about the world and what verbal symbols (i.e., words) mean in an amodal (i.e., not linked to any specific modality) memory store (e.g., storing what an *ostrich* is, what it looks like, etc. through words). Although there are long-standing debates regarding the strong distinction between semantic and episodic memory (e.g., McKoon, Ratcliff, & Dell, 1986), this dissociation was supported by early neuropsychological studies of amnestic patients who were able to acquire new semantic knowledge without having any concrete memory for having learned this information (Gabrieli, Cohen, & Corkin, 1988; O'Kane, Kensinger, & Corkin, 2004). Indeed, the relative independence of these two types of memory systems has guided research efforts for many years, as is evidenced by early work on computational models of semantic memory. As described below, this perspective is beginning to change with the onset of more recent modeling perspectives.

These theoretical seeds have driven three distinct approaches to modeling the structure and organization of semantic memory: associative network models, distributional models, and feature-based models. Associative network models are models that represent words as individual nodes in a large memory network, such that words that are related in meaning are connected to each other through edges in the network (e.g., Collins & Loftus, 1975; Collins & Quillian, 1969). On the other hand, inspired by the distributional hypothesis, Distributional Semantic Models (DSMs) collectively refer to a class of models where the meaning of a word is learned by extracting statistical redundancies and co-occurrence patterns from natural language. Importantly, DSMs provide explicit mechanisms for how words or features for a concept may be learned from the natural environment. Finally, feature models assume that words are represented in memory as a *distributed* collection of binary features (e.g., *birds* have wings, whereas *cars* do not), and the correlation or overlap of these features determines the extent to which words have similar meanings (Smith, Shoben, & Rips, 1974; Tversky, 1977). Overall, the network-based, feature-based, and distributional approaches to semantic modeling have sparked important debates in the literature and informed our understanding of the different facets involved in the construction of meaning. Therefore, this review attempts to highlight important milestones in the study of semantic memory, identify challenges currently facing the field, and integrate traditional ideas with modern approaches to modeling semantic memory.

In a recent article, Günther, Rinaldi, and Marelli (2019) reviewed several common misconceptions about distributional semantic models and evaluated the cognitive plausibility of modern DSMs. Although the current review is somewhat similar in scope to Günther et al.'s work, the current paper has different aims. Specifically, this review is a comprehensive analysis of models of semantic memory across multiple fields and tasks and so is not focused only on DSMs. It ties together classic models in psychology (e.g., associative network models, standard DSMs, etc.) with current state-of-the-art models in machine learning (e.g., transformer neural networks, convolutional neural networks, etc.) to elucidate the potential psychological mechanisms that these fields posit to underlie semantic retrieval processes. Further, the present work reviews the literature on modern multimodal semantic models, compositional semantics, and newer retrieval-based models, and therefore assesses these newer models and applications from a psychological perspective. Therefore, while Günther et al.'s review serves the role of clarifying how DSMs may indeed represent a cognitively plausible account of how meaning is learned, the present review serves the role of presenting a more comprehensive assessment and synthesis of multiple classes of models, theories, and learning mechanisms, as well as drawing closer ties between the rapidly progressing machine-learning literature and the constraints imposed by psychological research on semantic memory –

two fields that have so far been only loosely connected to each other. Therefore, the goal of the present review is to survey the current state of the field by tying together work from psychology, computational linguistics, and computer science, and also identify new challenges to guide future empirical research in modeling semantic memory.

Overview

This review emphasizes five important areas of research in semantic memory. The first section presents a modern perspective on the classic issues of semantic memory representation and learning. Associative, feature-based, and distributional semantic models are introduced and discussed within the context of how these models speak to important debates that have emerged in the literature regarding semantic versus associative relationships, prediction, and co-occurrence. In particular, a distinction is drawn between distributional models that propose *error-free* versus *error-driven* learning mechanisms for constructing meaning representations, and the extent to which these models explain performance in empirical tasks. Overall, although empirical tasks have partly informed computational models of semantic memory, the *empirical* and *computational* approaches to studying semantic memory have developed somewhat independently. Therefore, the first section attempts to bridge this gap by integrating empirical findings from lexical decision, pronunciation, and categorization tasks, with modeling approaches such as large-scale associative semantic networks (e.g., De Deyne, Navarro, Perfors, Brysbaert, & Storms, 2019; Steyvers & Tenenbaum, 2005), error-free learning-based DSMs (e.g., Jones & Mewhort, 2007; Landauer & Dumais, 1997), as well as error-driven learning-based models (e.g., Mikolov, Chen, Corrado, & Dean, 2013).

The second section presents an overview of psychological research in favor of conceptualizing semantic memory as part of a broader integrated memory system (Jamieson, Avery, Johns, & Jones, 2018; Kwantes, 2005; Yee, Jones, & McRae, 2018). The idea of semantic memory representations being context-dependent is discussed, based on findings from episodic memory tasks, sentence processing, and eye-tracking studies (e.g., Yee & Thompson-Schill, 2016). These empirical findings are then integrated with modern approaches to modeling semantic memory as a dynamic system that is sensitive to contextual dependencies, and can account for polysemy and attentional influences through topic models (Griffiths, Steyvers, & Tenenbaum, 2007), recurrent neural networks (Elman, 1991; Peters et al., 2018), and attention-based neural networks (Devlin, Chang, Lee, & Toutanova, 2019; Vaswani et al., 2017). The remainder of the section discusses the psychological plausibility of a relatively new class of models (retrieval-based models, e.g., Jamieson et al., 2018) that question the need for "learning" meaning at all, and instead propose that semantic representations are merely a product of retrieval-based operations in response to a cue, therefore blurring the traditional distinction between semantic and episodic memory (Tulving, 1972).

The third section discusses the issue of *grounding*, and how sensorimotor input and environmental interactions contribute to the construction of meaning. First, empirical findings from sensorimotor priming and cross-modal priming studies are discussed, which challenge the static, amodal, lexical nature of semantic memory that has been the focus of the majority of computational semantic models. There is now accumulating evidence that meaning cannot be represented exclusively through abstract, amodal symbols such as words (Barsalou, 2016). Therefore, important critiques of amodal computational models are clarified in the extent to which these models represent psychologically plausible models of semantic memory that include perceptual motor systems. Next, state-of-the-art computational models such as convolutional neural networks (Collobert et al., 2011), feature-integrated DSMs (Andrews, Vigliocco, & Vinson, 2009; Howell, Jankowicz, & Becker, 2005; Jones & Recchia, 2010), and multimodal DSMs (Bruni, Tran, & Baroni, 2014; Lazaridou, Pham, & Baroni, 2015) are discussed within the context of how these models are incorporating non-linguistic information in the learning process and tackling the grounding problem.

The fourth section focuses on the issue of *compositionality*, i.e., how words can be effectively combined and scaled up to represent higher-order linguistic structures such as sentences, paragraphs, or even episodic events. In particular, some early approaches to modeling compositional structures like vector addition (Landauer & Dumais, 1997), frequent phrase extraction (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013), and finding linguistic patterns in sentences (Turney & Pantel, 2010) are discussed. The rest of the section focuses on modern approaches to representing higher-order structures through hierarchical tree-based neural networks (Socher et al., 2013) and modern recurrent neural networks (Elman & McRae, 2019; Franklin, Norman, Ranganath, Zacks, & Gershman, 2019).

The fifth and final section focuses on some open issues in semantic modeling, such as proposing models that can be applied to other languages, issues related to data abundance and availability, understanding the social and evolutionary roles of language, and finding mechanistic process-based accounts of model performance. These issues shed light on important next steps in the study of semantic memory and will be critical in advancing our understanding of how meaning is constructed and guides cognitive behavior.

Many tasks, many models

Before delving into the details of each of the sections, it is important to emphasize here that models of semantic memory are inextricably tied to the behaviors and tasks that they seek to explain. For example, associative network models and early feature-based models explained response latencies in sentence verification tasks (e.g., deciding whether "a canary is a bird" is true or false). Similarly, early semantic models accounted for higher-order semantic relationships that emerge out of similarity judgments (e.g., Osgood, Suci, & Tannenbaum, 1957), although several of these models have since been applied to other tasks. Indeed, the study of *meaning* has spanned a variety of tasks, models, and phenomena, including but not limited to semantic priming effects in lexical decision tasks (Balota & Lorch, 1986), false memory paradigms (Deese, 1959; Roediger & McDermott, 1995), sentence verification (Smith et al., 1974), sentence comprehension (Duffy, Morris, & Rayner, 1988; Rayner & Frazier, 1989), and argument reasoning (Niven & Kao, 2019) tasks. Importantly, the cognitive processes underlying the sentence verification task may vastly differ from those underlying similarity judgments, which in turn may also differ from the processes underlying other complex tasks like reading comprehension and argument reasoning, and it is unclear whether and *how* a model of semantic memory that can successfully explain behavior in one task would be able to explain behavior in an entirely different task.

Of course, the ultimate goal of the semantic modeling enterprise is to propose *one* model of semantic memory that can be flexibly applied to a variety of semantic tasks, in an attempt to mirror the flexible and complex ways in which humans use knowledge and language (see, e.g., Balota & Yap, 2006). However, it is important to underscore the need to separate *representational* accounts from *process*-based accounts in the field. Modern approaches to modeling the representational nature of semantic memory have come very far in describing the continuum in which meaning exists, i.e., from the lowest-level input in the form of sensory and perceptual information, to words that form the building blocks of language, to high-level structures like schemas and events. However, process models operating on these underlying semantic representations have not received the same kind of attention and have developed somewhat independently from the representation modeling movement. For example, although process models like the drift-diffusion model (Ratcliff & McKoon, 2008), the optimal foraging model (Hills, 2006), and the temporal context model (Howard & Kahana, 2002) have been applied to some semantic tasks like verbal fluency (Hills, Jones, & Todd, 2012), free association (Howard, Shankar, & Jagadisan, 2011), and semantic judgments (e.g., Pirrone, Marshall, & Stafford, 2017), their application to different tasks remains limited and most research has instead focused on representational issues. Ultimately, combining process-based accounts with representational accounts is going to be critical in addressing some of the current challenges in the field, an issue that is emphasized in the final section of this review.

I. Semantic memory representation and learning

How individuals represent knowledge of concepts is one of the most important questions in semantic memory research and cognitive science. Therefore, significant research on human semantic memory has focused on issues related to memory representation and given rise to three distinct classes of models: associative network models, feature-based models, and distributional semantic models. This section presents a broad overview of these models, and also discusses some important debates regarding memory representation that these models have sparked in the field. Another related fundamental question in semantic memory research is regarding the learning of concepts, and the remainder of this section focuses on semantic models that subscribe to two broad psychological mechanisms (error-free and error-driven learning) that have been posited to underlie the learning of meaning representations.

Semantic memory representation

Network-based approaches Network-based approaches to semantic memory have a long and rich tradition rooted in psychology and computer science. Collins and Quillian (1969) investigated how individuals navigate through semantic memory to verify the truth of sentences (e.g., the time taken to verify that a *shark* <has fins>), and found that retrieval times were most consistent with a hierarchically organized memory network (see Fig. 1), where nodes represented words, and links or edges represented semantic propositions about the words (e.g., *fish* was connected to *animal* by an "is a" link, and *fish* also had its own attributes such as <has fins> and <can swim>). The mechanistic account of these findings was through a spreading activation framework (Quillian, 1967, 1969), according to which individual nodes in the network are activated, which in turn leads to the activation of neighboring nodes, and the network is traversed until the desired node or proposition is reached and a response is made. Interestingly, the number of steps taken to traverse the path in the proposed memory network predicted the time taken to verify a sentence in the original Collins and Quillian (1969) model. However, the original model could not explain typicality effects (e.g., why individuals respond faster to "*robin* <is a> *bird*" compared to "*ostrich* <is a> *bird*"), and also encountered difficulties in explaining differences in latencies for "false" sentences (e.g., why individuals are slower to reject "*butterfly* <is a> bird" compared to "*dolphin* <is a> bird"). Collins and Loftus (1975) later

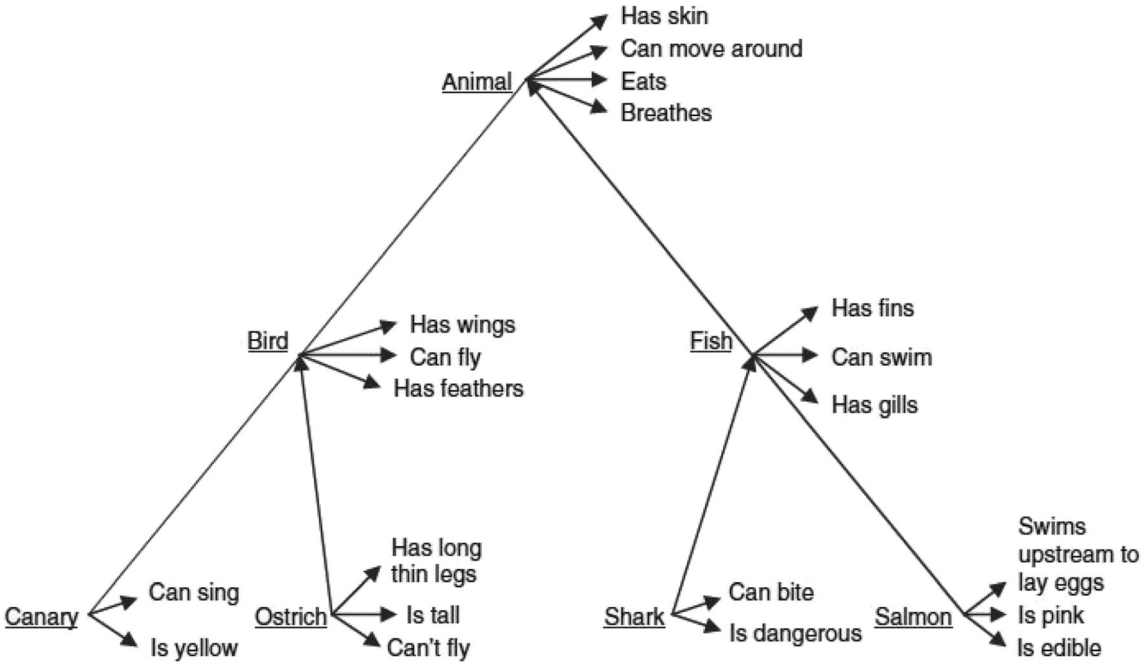

Fig. 1 Original network proposed by Collins and Quillian (1969). Reprinted from Balota and Coane (2008)

proposed a revised network model where links between words reflected the strength of the relationship, thereby eliminating the hierarchical structure from the original model to better account for behavioral patterns. This network/spreading activation framework was extensively applied to more general theories of language, memory, and problem solving (e.g., Anderson, 2000).

Computational network-based models of semantic memory have gained significant traction in the past decade, mainly due to the recent popularity of graph theoretical and network-science approaches to modeling cognitive processes (for a review, see Siew, Wulff, Beckage, & Kenett, 2018). Modern network-based approaches use large-scale databases to construct networks and capture large-scale relationships between nodes within the network. This approach has been used to empirically study the World Wide Web (Albert, Jeong, & Barabási, 2000; Barabási & Albert, 1999), biological systems (Watts & Strogatz, 1998), language (Steyvers & Tenenbaum, 2005; Vitevitch, Chan, & Goldstein, 2014), and personality and psychological disorders (for reviews, see Fried et al., 2017). Within the study of semantic memory, Steyvers and Tenenbaum (2005) pioneered this approach by constructing three different semantic networks using large-scale free-association norms (Nelson, McEvoy, & Schreiber, 2004), Roget's Thesaurus (Roget, 1911), and WordNet (Fellbaum, 1998; Miller, 1995). They presented several analyses to indicate that semantic networks possessed a "small-world structure," as indexed by high clustering coefficients (a parameter that estimates the likelihood that neighbors of two nodes will be neighbors themselves) and short average path lengths (a parameter that estimates the average number of steps from one node in the network to another), similar to several naturally occurring networks. Importantly, network metrics such as path length and clustering coefficients provide a quantitative way of estimating the large-scale organizational structure of semantic memory and the strength of relationships between words in a network (see Fig. 2), and have also proven to be successful in explaining performance across a wide variety of tasks, such as relatedness judgments (De Deyne & Storms, 2008; Kenett, Levi, Anaki, & Faust, 2017; Kumar, Balota, & Steyvers, 2019), verbal fluency (Abbott, Austerweil, & Griffiths, 2015; Zemla & Austerweil, 2018), and naming (Steyvers & Tenenbaum, 2005).Other work in this area has explored the influence of semantic network metrics on psychological disorders (Kenett, Gold, & Faust, 2016), creativity (Kenett, Anaki, & Faust, 2014), and personality (Beaty et al., 2016).

Despite the success of modern semantic networks at predicting cognitive performance, there is some skepticism in the field regarding the use of *free-association norms* to create network representations (Jones, Hills, & Todd, 2015; Siew et al., 2018). Specifically, it is not clear whether networks constructed from association norms are indeed a *representational* account of semantic memory or simply reflect the product of a retrieval-based *process* on an underlying representation of semantic memory. For example, producing the response *ostrich* to the word *emu* represents a retrieval-based process cued by the word *emu*, and may not necessarily reflect how the underlying representation of the words came to be closely associated in the first place. Therefore, it appears that associative network models lack an explicit mechanism through which concepts were learned in the first place.

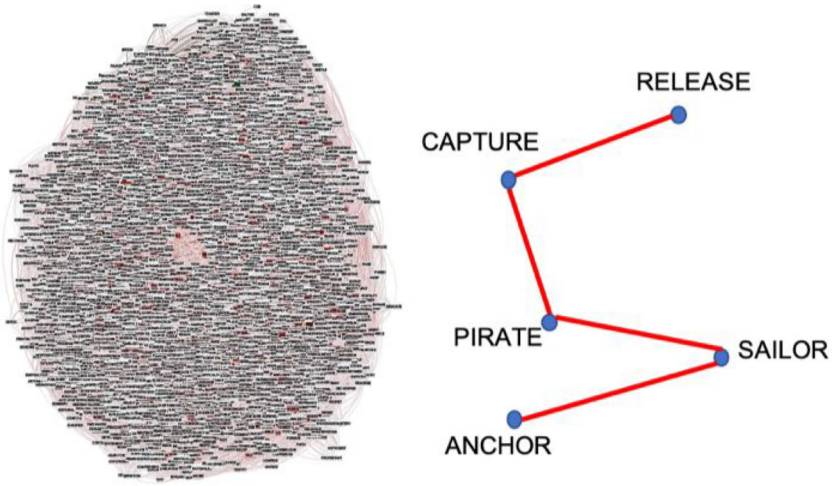

Fig. 2 Large-scale visualization of a directed semantic network created by Steyvers and Tenenbaum (2005) and shortest path between RELEASE to ANCHOR. Adapted from Kumar, Balota, and Steyvers (2019)

A recent example of this fundamental debate regarding the origin of the representation comes from research on the semantic fluency task, where participants are presented with a natural category label (e.g., "animals") and are required to generate as many exemplars from that category (e.g., *lion*, *tiger*, *elephant*…) as possible within a fixed time period. Hills, Jones, and Todd (2012) proposed that the temporal pattern of responses produced in the fluency task mimics optimal foraging techniques found among animals in natural environments. They provided a computational account of this search process based on the BEAGLE model (Jones & Mewhort, 2007). However, Abbott et al. (2015) contended that the behavioral patterns observed in the task could also be explained by a more parsimonious random walk on a network representation of semantic memory created from free-association norms. This led to a series of rebuttals from both camps (Jones, Hills, & Todd, 2015; Nematzadeh, Miscevic, & Stevenson, 2016), and continues to remain an open debate in the field (Avery & Jones, 2018). However, Jones, Hills, and Todd (2015) argued that while free-association norms are a useful proxy for memory representation, they remain an outcome variable from a search process on a representation and cannot be a pure measure of how semantic memory is organized. Indeed, Avery and Jones (2018) showed that when the input to the network and distributional space was controlled (i.e., both were constructed from text corpora), random walk and foraging-based models *both* explained semantic fluency data, although the foraging model outperformed several different random walk models. Of course, these findings are specific to the semantic fluency task and adequately controlled comparisons of network models to DSMs remain limited. However, this work raises the question of whether the success of association networks in explaining behavioral performance in cognitive tasks is a consequence of shared variance with the cognitive tasks themselves (e.g., fluency tasks can be thought of as association tasks constrained to a particular category) or truly reflects the structural representation of semantic memory, an issue that is discussed in detail in the section summary. Nonetheless, recent work in this area has focused on creating network representations using a learning model instead of behavioral data (Nematzadeh et al., 2016), and also advocated for alternative representations that incorporate such learning mechanisms and provide a computational account of how word associations might be *learned* in the first place.

Feature-based approaches Feature-based models depart from the traditional notion that a word has a localized representation (e.g., in an association network). The core idea behind feature models is that words are represented in memory as a collection of binary features (e.g., *birds* have *wings*, whereas *cars* do not), and the correlation or overlap of these features determines the extent to which words have similar meanings. Smith et al. (1974) proposed a feature-comparison model in which concepts had two types of semantic features: *defining* features that are shared by all concepts, and *characteristic* features that are specific to only some exemplars. For example, all birds <have wings> (defining feature) but not all birds <fly> (characteristic feature). Similarity between concepts in this model was computed through a feature comparison process, and the degree of overlap between the features of two concepts directly predicted sentence verification times, typicality effects, and differences in response times in responding to "false" sentences. This notion of featural overlap as an index of similarity was also central to the theory of feature matching proposed by Tversky (1977). Tversky viewed similarity as a set-theoretical matching function, such that the similarity between *a* and *b* could be conceptualized through a contrast model as a function of features that are common to both *a* and *b* (common features), and features that belong to *a* but not *b*, as well as features that belong to *b* but not *a*

(distinctive features). Tversky's contrast model successfully accounted for asymmetry in similarity judgments and judgments of difference for words, shapes, and letters.

Although early feature-based models of semantic memory set the groundwork for modern approaches to semantic modeling, none of the models had any systematic way of measuring these features (e.g., Smith et al., 1974, applied multidimensional scaling to similarity ratings to uncover underlying features). Later versions of feature-based models thus focused on explicitly coding these features into computational models by using norms from property-generation tasks (McRae, De Sa, & Seidenberg, 1997). To obtain these norms, participants were asked to list features for concepts (e.g., for the word *ostrich*, participants may list <is a> bird, <has wings>, <is heavy>, and <does not fly> as features), the idea being that these features constitute explicit knowledge participants have about a concept. McRae et al. then used these features to train a model using simple correlational learning algorithms (see next subsection) applied over a number of iterations, which enabled the network to settle into a stable state that represented a learned concept. A critical result of this modeling approach was that correlations among features predicted response latencies in feature-verification tasks in human participants as well as model simulations. Importantly, this approach highlighted how statistical regularities among features may be encoded in a memory representation over time. Subsequent work in this line of research demonstrated how feature correlations predicted differences in priming for living and nonliving things and explained typicality effects (McRae, 2004).

Despite the success of computational feature-based models, an important limitation common to *both* network and feature-based models was their inability to explain how knowledge of individual features or concepts was *learned* in the first place. For example, while feature-based models can explain that *ostrich* and *emu* are similar because both <have feathers>, how did an individual learn that <having feathers> is a feature that an *ostrich* or *emu* has? McRae et al. claimed that features were derived from repeated multimodal interactions with exemplars of a particular concept, but how this learning process might work in practice was missing from the implementation of these models. Still, feature-based models have been very useful in advancing our understanding of semantic memory structure, and the integration of feature-based information with modern machine-learning models continues to remain an active area of research (see Section III).

Distributional approaches Distributional Semantic Models (DSMs) refer to a class of models that provide explicit mechanisms for how words or features for a concept may be learned from the natural environment. Therefore, unlike associative network models or feature-based models, DSMs do not use free-association responses or feature norms, but instead build representations by directly extracting statistical regularities from a large natural language corpus (e.g., books, newspapers, online articles, etc.), the assumption being that large text corpora are a good proxy for the language that individuals are exposed to in their lifetime. The principle of extracting co-occurrence patterns and inferring associations between concepts/words from a large text-corpus is at the core of all DSMs, but exactly how these patterns are extracted has important implications for how these models conceptualize the learning process. Specifically, two distinct psychological mechanisms have been proposed to account for associative learning, broadly referred to as *error-free* and *error-driven* learning mechanisms. Error-free learning mechanisms refer to a class of psychological mechanisms that posit that learning occurs by identifying clusters of events that tend to co-occur in temporal proximity, and dates back to Hebb's (1949; also see McCulloch & Pitts, 1943) proposal of how individual neurons in the brain adjust their firing rates and activations in response to activations of other neighboring neurons. This Hebbian learning mechanism is at the heart of several classic and recent models of semantic memory, which are discussed in this section. On the other hand, error-driven learning mechanisms posit that learning is accomplished by predicting events in response to a stimulus, and then applying an error-correction mechanism to learn associations. Error-correction mechanisms often vary across learning models but broadly share principles with Rescorla and Wagner's (1972) model of animal cognition, where they described how learning may actually be driven by expectation error, instead of error-free associative learning (Rescorla, 1988). This section reviews DSMs that are consistent with the error-free and error-driven learning approaches to constructing meaning representations, and the summary section discusses the evidence in favor of and against each class of models.

Semantic memory learning

Error-free learning-based DSMs One of the earliest DSMs, the Hyperspace Analogue to Language (HAL; Lund & Burgess, 1996), built semantic representations by counting the co-occurrences of words within a sliding window of five to ten words, where co-occurrence between any two words was inversely proportional to the distance between the two words in that window. These local co-occurrences produced a word-by-word co-occurrence matrix that served as a spatial representation of meaning, such that words that were semantically related were closer in a high-dimensional space (see Fig. 3; *ear*, *eye*, and *nose* all acquire very similar representations). This relatively simple error-free learning mechanism was able to account for a wide variety of cognitive phenomena in tasks such as lexical decision and categorization (Li, Burgess, & Lund, 2000). However, HAL encountered difficulties in accounting for mediated priming effects (Livesay & Burgess,

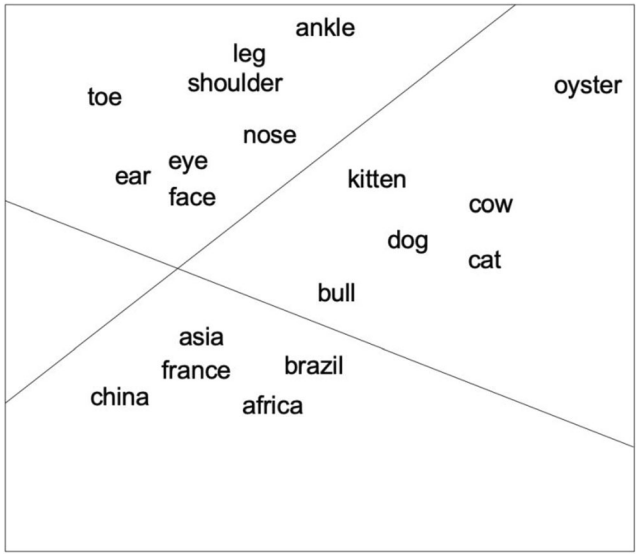

Fig. 3 The high-dimensional space produced by HAL from co-occurrence word vectors. Adapted from Lund and Burgess (1996)

1998; see section summary for details), which was considered as evidence in favor of semantic network models. However, despite its limitations, HAL was an important step in the ongoing development of DSMs.

Another popular distributional model that has been widely applied across cognitive science is Latent Semantic Analysis (LSA; Landauer & Dumais, 1997), a semantic model that has successfully explained performance in several cognitive tasks such as semantic similarity (Landauer & Dumais, 1997), discourse comprehension (Kintsch, 1998), and essay scoring (Landauer, Laham, Rehder, & Schreiner, 1997). LSA begins with a word-document matrix of a text corpus, where each row represents the frequency of a word in each corresponding document, which is clearly different from HAL's word-by-word co-occurrence matrix. Further, unlike HAL, LSA first transforms these simple frequency counts into log frequencies weighted by the word's overall importance over documents, to de-emphasize the influence of unimportant frequent words in the corpus. This transformed matrix is then factorized using truncated singular value decomposition, a factor-analytic technique used to infer latent dimensions from a multidimensional representation. The semantic representation of a word can then be conceptualized as an aggregate or *distributed* pattern across a few hundred dimensions. The construction of a word-by-document matrix and the dimensionality reduction step are central to LSA and have the important consequence of uncovering *global* or *indirect* relationships between words even if they never co-occurred with each other in the original context of documents. For example, *lion* and *stripes* may have never co-occurred within a sentence or document, but because they often occur in similar contexts of the word *tiger*, they would develop similar semantic representations. Importantly, the ability to infer latent dimensions and extend the context window from sentences to documents differentiates LSA from a model like HAL.

Despite its widespread application and success, LSA has been criticized on several grounds over the years, e.g., for ignoring word transitions (Perfetti, 1998), violating power laws of connectivity (Steyvers & Tenenbaum, 2005), and for the lack of a mechanism for learning incrementally (Jones, Willits, & Dennis, 2015). The last point is particularly important, as the LSA model assumes that meaning is learned and computed *after* a large amount of co-occurrence information is available (i.e., in the form of a word-by-document matrix). This is clearly unconvincing from a psychological standpoint and is often cited as a reason for distributional models being implausible psychological models (Hoffman, McClelland, & Lambon Ralph, 2018; Sloutsky, Yim, Yao, & Dennis, 2017). However, as Günther et al. (2019) have recently noted, this is an argument against batch-learning models like LSA, and not *distributional* models per se. In principle, LSA can learn incrementally by updating the co-occurrence matrix as each input is received and re-computing the latent dimensions (for a demonstration, see Olney, 2011), although this process would be computationally very expensive. In addition, there are several modern DSMs that are incremental learners and propose psychologically plausible accounts of semantic representation.

One such incremental approach involves developing random representations of words that slowly accumulate information about meaning through repeated exposure to words in a large text corpus. For example, Bound Encoding of the Aggregate Language Environment (BEAGLE; Jones & Mewhort, 2007) is a random vector accumulation model that gradually builds semantic representations as it processes text in sentence-sized context windows. BEAGLE begins by assigning a random, static environmental vector to a word the first time it is encountered in the corpus. This environmental vector does not change over different exposures of the word and is hypothesized to represent stable physical characteristics about the word. When words co-occur in a sentence, their environmental vectors are added to each other's representations, and, thus, their memory representations become similar over time. Further, even if two words never co-occur, they develop similar representations if they co-occur with the same words. This leads to the formation of higher-order relationships between words, without performing any LSA-type dimensionality reduction. Importantly, BEAGLE integrates this context-based information with word-order information using a technique called circular convolution (an effective method to combine two n-dimensional vectors into an associated vector of the same dimensions). BEAGLE computes order information by binding together all word chunks (formally called *n*-grams) that a particular word is part of (e.g., for the sentence "an ostrich flapped its wings", the two-gram convolution would bind the representations for <*an*, *ostrich*> and

<*ostrich*, *flapped*> together) and then summing this order vector with the word's context vector to compute the final semantic representation of the word. Thus, words that co-occur in similar contexts as well as in the same syntactic positions develop similar representations as the model acquires more experience through the corpus. BEAGLE outperforms several classic models of word representation (e.g., LSA and HAL), and explains performance on several complex tasks, such as mediated priming effects in lexical decision and pronunciation tasks, typicality effects in exemplar categorization, and reading times in stem completion tasks (Jones & Mewhort, 2007). Importantly, through the addition of environmental vectors of words whenever they co-occur, BEAGLE also indirectly infers relationships between words that did not directly co-occur. This process is similar in principle to inferring indirect co-occurrences across documents in LSA and can be thought of as an abstraction-based process applied to direct co-occurrence patterns, albeit through different mechanisms. Other incremental models use ideas similar to BEAGLE for accumulating semantic information over time, although they differ in their theoretical underpinnings (Howard et al., 2011; Sahlgren, Holst, & Kanerva, 2008) and the extent to which they integrate order information in the final representations (Kanerva, 2009). It is important to note here that the DSMs discussed so far (HAL, LSA, and BEAGLE) all share the principle of deriving meaning representations through error-free learning mechanisms, in the spirit of Hebbian associative learning. The following section discusses other DSMs that also produce rich semantic representations but are instead based on error-driven learning mechanisms or prediction.

Error-driven learning-based DSMs In contrast to error-free learning DSMs, a different approach to building semantic representations has focused on how representations may slowly develop through prediction and error-correction mechanisms. These models are also referred to as *connectionist* models and propose that meaning emerges through prediction-based weighted interactions between interconnected units (Rumelhart, Hinton, & McClelland, 1986). Most connectionist models typically consist of an *input* layer, an *output* layer, and one or more intervening units collectively called the *hidden* layers, each of which contains one or more "nodes" or units. Activating the nodes of the input layer (through an external stimulus) leads to activation or suppression of units connected to the input units, as a function of the weighted connection strengths between the units. Activation gradually reaches the output units, and the relationship between output units and input units is of primary interest. Learning in connectionist models (sometimes called *feed-forward* networks if there are no recurrent connections, see section II), can be accomplished in a *supervised* or *unsupervised* manner. In supervised learning, the network tries to maximize the likelihood of a desired goal or output for a given set of input units by predicting outputs at every iteration. The weights of the signals are thus adjusted to minimize the error between the target output and the network's output, through error backpropagation (Rumelhart, Hinton, & Williams, 1988). In unsupervised learning, weights within the network are adjusted based on the inherent structure of the data, which is used to inform the model about prediction errors (e.g., Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013).

Rumelhart and Todd (1993) proposed one of the first feed-forward connectionist models of semantic memory. To train the network, all facts represented in a traditional semantic network (e.g., Collins & Quillian, 1969) were first converted to input-output training pairs (e.g., the fact *bird* <has wings> was converted to term 1: *bird* – relation: has – term 2: wings). Then, the network learned semantic representations in a supervised manner, by turning on the input and relation units, and backpropagating the error from predicted output units through two hidden layers. For example, the words *oak* and *pine* acquired a similar pattern of activation across the hidden units because their node-relations pairs were similar during training. Additionally, the network was able to hierarchically learn information about new concepts (e.g., adding the *sparrow* <is a> *bird* link in the model formed a new representation for *sparrow* that also included relations like <has wings>, <can fly>, etc.). Connectionist networks are sometimes also called *neural networks* (NNs) to emphasize that connectionist models (old and new) are inspired by neurobiology and attempt to model how the brain might process incoming input and perform a particular task, although this is a very loose analogy and modern researchers do not view neural networks as accurate models of the brain (Bengio, Goodfellow, & Courville, 2015).

A feed-forward NN model, word2vec, proposed by researchers at Google (Mikolov, Chen, et al., 2013) has gained immense popularity in the last few years due to its impressive performance on a variety of semantic tasks. Word2vec is a two-layer NN model that has two versions: continuous bag-of-words (CBOW) and skip-gram. The objective of the CBOW model is to predict a target word, given four context words before and after the intended word, using a classifier. The skip-gram model reverses this objective and attempts to predict the surrounding context words, given an input word (see Figs. 4 and 5). In this way, word2vec trains the network on a surrogate task and iteratively improves the word representations or "embeddings" (represented via the hidden layer units) formed during this process by computing stochastic gradient descent, a common technique to compute prediction error for backpropagation in NN models. Further, word2vec tweaks several *hyperparameters* to achieve optimal performance. For example, it uses dynamic context windows so that words that are more distant from the target word are sampled less frequently in the prediction task. Additionally, word2vec de-emphasizes the role of frequent words by discarding

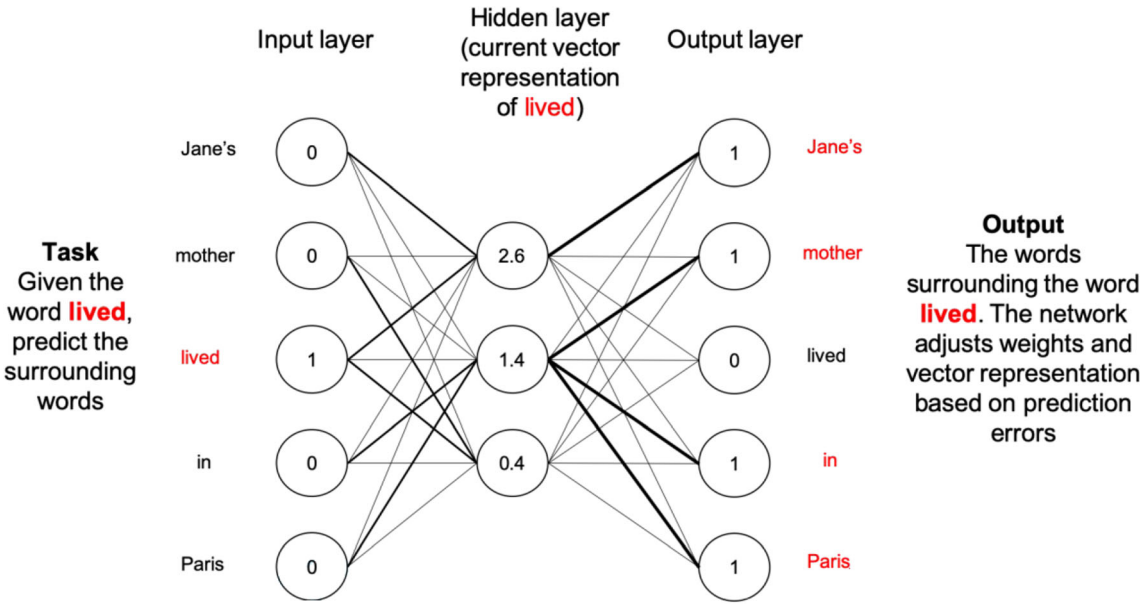

Fig. 4 A depiction of the skip-gram version of the word2vec model architecture. The model is creating a vector representation for the word *lived* by predicting its surrounding words in the sentence "Jane's mother lived in Paris." The weights of the hidden layer represent the vector representation for the word *lived*, as the model performs the prediction task and adjusts the weights based on the prediction error. Adapted from Günther et al. (2019)

frequent words above a threshold with some probability. Finally, to refine representations, word2vec uses negative sampling, by which the model randomly samples a set of unrelated words and learns to suppress these words during prediction. These sophisticated techniques allow word2vec to develop very rich semantic representations. For example, word2vec is able to solve verbal analogy problems, e.g., man: king :: woman: ??, through simple vector arithmetic (but see Chen, Peterson, & Griffiths, 2017), and also model human similarity judgments. This indicates that the representations acquired by word2vec are sensitive to complex higher-order semantic relationships, a characteristic that had not been previously observed or demonstrated in other NN models. Further, word2vec is a very weakly supervised (or unsupervised) learning algorithm, as it does not require labeled or annotated data but only sequential text (i.e., sentences) to generate the word embeddings. word2vec's pretrained embeddings have proven to be useful inputs for several downstream natural language-processing tasks (Collobert & Weston, 2008) and have inspired several other embedding models. For example, fastText (Bojanowski, Grave, Joulin, & Mikolov, 2017) is a word2vec-type NN that incorporates character-level information (i.e., n-grams) in the learning process, which leads to more fine-grained representations for rare words and words that are not in the training corpus. However, the psychological validity of some of the hyperparameters used by word2vec has been called into question by some researchers. For example, Johns, Mewhort, and Jones (2019) recently investigated how negative sampling, which appears to be psychologically unintuitive, affects semantic representations. They argued that negative sampling simply establishes a more accurate base rate of word occurrence and proposed solutions to integrate base-rate information into BEAGLE without the need to randomly sample unrelated words or even a prediction mechanism. However, as discussed in subsequent sections, prediction appears to be a central mechanism in certain tasks that involve sequential dependencies, and it is possible that NN models based on prediction are indeed capturing these long-term dependencies.

Another modern distributional model, Global Vectors (GloVe), which was recently introduced by Pennington, Socher, and Manning (2014), shares similarities with both error-free and NN-based error-driven models of word representation. Similar to several DSMs, GloVe begins with a

Probability and Ratio	*k* = *solid*	*k* = *gas*	*k* = *water*	*k* = *fashion*
P(*k* | *ice*)	1.9×10^{-4}	6.6×10^{-5}	3.0×10^{-3}	1.7×10^{-5}
P(*k* | *steam*)	2.2×10^{-5}	7.8×10^{-4}	2.2×10^{-3}	1.8×10^{-5}
P(*k* | *ice*)/ P(*k* | *steam*)	8.9	8.5×10^{-2}	1.36	0.96

Fig. 5 Ratio of co-occurrence probabilities for ice and steam, as described in Pennington et al. (2014)

word-by-word co-occurrence matrix. But, instead of using raw counts as a starting point, GloVe estimates the ratio of co-occurrence probabilities between words. To give an example used by the authors, based on statistics from text corpora, *ice* co-occurs more frequently with *solid* than it does with *gas*, whereas *steam* co-occurs more frequently with *gas* than it does with *solid*. Further, both words (*ice* and *steam*) co-occur with their shared property *water* frequently, and both co-occur with the unrelated word *fashion* infrequently. The critical insight that GloVE capitalizes on is that words like *water* and *fashion* are non-discriminative, but the words *gas* and *solid* are important in differentiating between *ice* and *steam*. The *ratio* of probabilities highlights these differences, such that large values (much greater than 1) correspond to properties specific to *ice*, and small values (much less than 1) correspond to properties specific of *steam* (see Fig. 4). In this way, co-occurrence ratios successfully capture abstract concepts such as thermodynamic phases. GloVe aims to *predict* the logarithm of these co-occurrence ratios between words using a regression model, in the same spirit as factorizing the logarithm of the co-occurrence matrix in LSA. Therefore, while incorporating global information in the learning process (similar to LSA), GloVe also uses error-driven mechanisms to minimize the cost function from the regression model (using a modified version of stochastic gradient descent, similar to word2vec), and therefore represents a type of hybrid model. Further, to de-emphasize the overt influence of frequent and rare words, GloVe penalizes words with very high and low frequency (similar to importance weighting in LSA). The final abstracted representations or "embeddings" that emerge from the GloVe model are particularly sensitive to higher-order semantic relationships, and the GloVe model has been shown to perform remarkably well at analogy tasks, word similarity judgments, and named entity recognition (Pennington et al., 2014), although there is little consensus in the field regarding the relative performance of GloVe against strictly prediction-based models (e.g., word2vec; see Baroni, Dinu, & Kruszewski, 2014; Levy & Goldberg, 2014)

Summary

This section provided a detailed overview of traditional and recent computational models of semantic memory and highlighted the core ideas that have inspired the field in the past few decades with respect to semantic memory representation and learning. While several models draw inspiration from psychological principles, the differences between them certainly have implications for the extent to which they explain behavior. This summary focuses on the extent to which associative network and feature-based models, as well as error-free and error-driven learning-based DSMs speak to important debates regarding association, direct and indirect patterns of co-occurrence, and prediction.

Semantic versus associative relationships Within the network-based conceptualization of semantic memory, concepts that are related to each other are directly connected (e.g., *ostrich* and *emu* have a direct link). An important insight that follows from this line of reasoning is that if *ostrich* and *emu* are indeed related, then processing one of the words should facilitate processing for the other word. This was indeed the observation made by Meyer and Schvaneveldt (1971), who reported the first semantic priming study, where they found that individuals were faster to make lexical decisions (deciding whether a presented stimulus was a word or non-word) for semantically related (e.g., *ostrich-emu*) word pairs, compared to unrelated word pairs (e.g., *apple-emu*). Given that individuals were not required to access the semantic relationship between words to make the lexical decision, these findings suggested that the task potentially reflected automatic retrieval processes operating on underlying semantic representations (also see Neely, 1977). The semantic priming paradigm has since become the most widely applied task in cognitive psychology to examine semantic representation and processes (for reviews, see Hutchison, 2003; Lucas, 2000; Neely, 1977).

An important debate that arose within the semantic priming literature was regarding the nature of the relationship that produces the semantic priming effect as well as the basis for connecting edges in a semantic network. Specifically, does processing the word *ostrich* facilitate the processing of the word *emu* due to the associative strength of connections between *ostrich* and *emu*, or because the semantic features that form the concepts of *ostrich* and *emu* largely overlap? As discussed earlier, *associative* relations are thought to reflect contiguous associations that individuals likely infer from natural language (e.g., *ostrich-egg*). Traditionally, such associative relationships have been operationalized through responses in a free-association task (e.g., De Deyne et al., 2019; Nelson et al., 2004). On the other hand, *semantic* relations have traditionally included only category coordinates or concepts with similar features (e.g., *ostrich-emu*; Hutchison, 2003; Lucas, 2000). Given these different operationalizations, some researchers have attempted to isolate pure "semantic" priming effects by selecting items that are semantically related (i.e., share category membership; Fischler, 1977; Lupker, 1984; Thompson-Schill, Kurtz, & Gabrieli, 1998) but not associatively related (i.e., based on free-association norms), although these attempts have not been successful. Specifically, there appear to be discrepancies in how associative strength is defined and the locus of these priming effects. For example, in a meta-analytic review, Lucas (2000) concluded that semantic priming effects can indeed be found in the absence of associations, arguing for the existence of "pure" semantic effects. In contrast, Hutchison (2003) revisited the same studies and

argued that both associative and semantic relatedness can produce priming, and the effects largely depend on the type of semantic relation being investigated as well as the task demands (also see Balota & Paul, 1996).

Another line of research in support of associative influences underlying semantic priming comes from studies on mediated priming. In a typical experiment, the prime (e.g., *lion*) is related to the target (e.g., *stripes*) only through a mediator (e.g., *tiger*), which is not presented during the task. The critical finding is that robust priming effects are observed in pronunciation and lexical decision tasks for mediated word pairs that do not share any obvious semantic relationship or featural overlap (Balota & Lorch, 1986; Livesay & Burgess, 1998; McNamara & Altarriba, 1988). Traditionally, mediated priming effects have been explained through an associative-network based account of semantic representation (e.g., Balota & Lorch, 1986), where, consistent with a spreading activation mechanism, activation from the prime node (e.g., *lion*) spreads to the mediator node in the network (e.g., *tiger*), which in turn activates the related target node (e.g., *stripes*). Recent computational network models have supported this conceptualization of semantic memory as an *associative* network. For example, Kenett et al. (2017) constructed a Hebrew network based on correlations of responses in a free-association task, and showed that network path lengths in this Hebrew network successfully predicted the time taken by participants to decide whether two words were related or unrelated, for directly related (e.g., *bus-car*) and relatively distant word pairs (e.g., *cheater-carpet*). More recently, Kumar, Balota, and Steyvers (2019) replicated Kenett et al.'s work in a much larger corpus in English, and also showed that undirected and directed networks created by Steyvers and Tenenbaum (2005) also account for such distant priming effects.

While network models provide a straightforward account for mediated (and distant) priming, such effects were traditionally considered a core challenge for feature-based and distributional semantic models (Hutchison, 2003; Masson, 1995; Plaut & Booth, 2000). The argument was that in feature-based representations that conceptualize word meaning through the presence or absence of features, *lion* and *stripes* would not overlap because *lions* do not have *stripes*. Similarly, in distributional models, at least some early evidence from the HAL model suggested that mediated word pairs neither co-occur nor have similar high-dimensional vector representations (Livesay & Burgess, 1998), which was taken as evidence against a distributional representation of semantic memory. However, other distributional models such as LSA and BEAGLE have since been able to account for mediated priming effects (e.g., Chwilla & Kolk, 2002; Hutchison, 2003; Jones, Kintsch, & Mewhort, 2006; Jones & Mewhort, 2007; Kumar, Balota, & Steyvers, 2019). In fact, Jones et al. (2006) showed that HAL's greater focus on "semantic" relationships contributes to its inability to account for mediated priming effects, which are more "associative" in nature (also see Sahlgren, 2008). However, LSA and other DSMs that subscribe to a broader conceptualization of meaning that includes *both* local "associative" as well as *global* "semantic" relationships are indeed able to account for mediated priming effects. The counterargument is that mediated priming may simply reflect weak semantic relationships between words (McKoon & Ratcliff, 1992), which can indeed be learned from statistical regularities in natural language. Thus, even though *lion* and *stripes* may have never co-occurred, newer semantic models that capitalize on higher-order indirect relationships are able to extract similar vectors for these words and produce the same priming effects without the need for a mediator or a spreading activation mechanism (Jones et al., 2006).

Therefore, an important takeaway from these studies on clarifying the locus of semantic priming effects is that the traditional distinction between *associative* and *semantic* relations may need to be revisited. Importantly, the operationalization of associative relations through free-association norms has further complicated this distinction, as *only* responses that are produced in free-association tasks have been traditionally considered to be associative in nature. However, free association responses may themselves reflect a wide variety of semantic relations (McRae, Khalkhali, & Hare, 2012; see also Guida & Lenci, 2007) that can produce different types of semantic priming (Hutchison, 2003). Indeed, as McRae et al. (2012) noted, several of the associative level relations examined in previous work (e.g., Lucas, 2000) could in fact be considered semantically related in the broad sense (e.g., scene, feature, and script relations). Within this view, it is unclear exactly how *associative* relations operationalized in this way can be truly separated from *semantic* relations, or conversely, how *semantic* relations could truly be considered any different from simple *associative* co-occurrence. In fact, it is unlikely that words are purely associative or purely semantically related. As McNamara (2005) noted, "Having devoted a fair amount of time perusing free-association norms, I challenge anyone to find two highly associated words that are not semantically related in some plausible way" (McNamara, 2005; pp. 86). Furthermore, the traditional notion of what constitutes a "semantic" relationship has changed and is no longer limited to only coordinate or feature-based overlap, as is evidenced by the DSMs discussed in this section. Therefore, it appears that *both* associative relationships as well as coordinate/feature relationships now fall within the broader umbrella of what is considered semantic memory.

There is one possible way to reconcile the historical distinction between what are considered traditionally associative and "semantic" relationships. Some relationships may be simply dependent on *direct* and *local* co-occurrence of words in natural language (e.g., *ostrich* and *egg* frequently co-occur in

natural language), whereas other relationships may in fact emerge from *indirect* co-occurrence (e.g., *ostrich* and *emu* do not co-occur with each other, but tend to co-occur with similar words). Within this view, traditionally "associative" relationships may reflect more direct co-occurrence patterns, whereas traditionally "semantic" relationships, or coordinate/featural relations, may reflect more indirect co-occurrence patterns. As discussed in this section, DSMs often distinguish between and differentially emphasize these two types of relationships (i.e., direct vs. indirect co-occurrences; see Jones et al., 2006), which has important implications for the extent to which these models speak to this debate between associative vs. truly semantic relationships. The combined evidence from the semantic priming literature and computational modeling literature suggests that the formation of direct associations is most likely an *initial* step in the computation of meaning. However, it also appears that the complex semantic memory system does not simply rely on these direct associations but also applies additional learning mechanisms (vector accumulation, abstraction, etc.) to derive other meaningful, indirect semantic relationships. Implementing such global processes allows modern distributional models to develop more fine-grained semantic representations that capture different types of relationships (direct and indirect). However, there do appear to be important differences in the underlying mechanisms of meaning construction posited by different DSMs. Further, there is also some concern in the field regarding the reliance on pure linguistic corpora to construct meaning representations (De Deyne, Perfors, & Navarro, 2016), an issue that is closely related to assessing the role of associative networks and feature-based models in understanding semantic memory, as discussed below. Furthermore, it is also unlikely that any semantic relationships are *purely* direct or indirect and may instead fall on a continuum, which echoes the arguments posed by Hutchison (2003) and Balota and Paul (1996) regarding semantic versus associative relationships.

Value of associative networks and feature-based models

Another important part of this debate on associative relationships is the representational issues posed by association network models and feature-based models. As discussed earlier, the validity of associative semantic networks and feature-based models as accurate models of semantic memory has been called into question (Jones, Hills, & Todd, 2015) due to the lack of explicit mechanisms for learning relationships between words. One important observation from this work is that the debate is less about the underlying *structure* (network-based/localist or distributed) and more about the *input* contributing to the resulting structure. Networks and feature lists in and of themselves are simply tools to represent a particular set of data, similar to high-dimensional vector spaces. As such, cosines in vector spaces can be converted to step-based distances that form a network using cosine thresholds (e.g., Gruenenfelder, Recchia, Rubin, & Jones, 2016; Steyvers & Tenenbaum, 2005) or a binary list of features (similar to "dimensions" in DSMs). Therefore, the critical difference between associative networks/feature-based models and DSMs is not that the former is a network/list and the latter is a vector space, but rather the fact that associative networks are constructed from free-association responses, feature-based models use property norms, and DSMs learn from text corpora. Therefore, as discussed earlier, the success of associative networks (or feature-based models) in explaining behavioral performance in cognitive tasks could be a consequence of shared variance with the cognitive tasks themselves. However, associative networks also explain performance in tasks that are arguably *not* based solely on retrieving associations or features – for example, progressive demasking (Kumar, Balota, & Steyvers, 2019), similarity judgments (Richie, Zou, & Bhatia, 2019), and the remote triads task where participants are asked to choose the most related pair among a set of three nouns (De Deyne, Perfors, & Navarro, 2016). This points to the possibility that the part of the variance explained by associative networks or feature-based models may in fact be *meaningful* variance that distributional models are unable to capture, instead of entirely being *shared* task-based variance.

To the extent that DSMs are limited by the corpora they are trained on (Recchia & Jones, 2009), it is possible that the responses from free-association tasks and property-generation norms capture some non-linguistic aspects of meaning that are missing from standard DSMs, for example, imagery, emotion, perception, etc. Therefore, even though it is unlikely that associative networks and feature-based models are a *complete* account of semantic memory, the free-association and property-generation *norms* that they are constructed from are likely useful baselines to compare DSMs against, because they include different types of relationships that go beyond those observable in textual corpora (De Deyne, Perfors, & Navarro, 2016). To that end, Gruenenfelder et al. (2016) compared three distributional models (LSA, BEAGLE, and Topic models) and one simple associative model and indicated that only a hybrid model that combined contextual similarity and associative networks successfully predicted the graph theoretic properties of free-association norms (also see Richie, White, Bhatia, & Hout, 2019). Therefore, associative networks and feature-based models can potentially capture complementary information compared to standard distributional models, and may provide additional cues about the features and associations other than co-occurrence that may constitute meaning. For instance, there is evidence to show that perceptual features such as *size*, *color*, and *texture* that are readily apparent to humans and may be used to infer semantic relationships, are not effectively captured by co-occurrence statistics derived from natural language corpora (e.g., Baroni & Lenci, 2008; see Section III),

suggesting that semantic memory may in fact go beyond simple co-occurrence. Indeed, as discussed in Section III, multimodal and feature-integrated DSMs that use different linguistic and non-linguistic sources of information to learn semantic representations are currently a thriving area of research and are slowly changing the conceptualization of what constitutes semantic memory (e.g., Bruni et al., 2014; Lazaridou et al., 2015).

Error-free versus error-driven learning Prediction is another contentious issue in semantic modeling that has gained a considerable amount of traction in recent years, and the traditional distinction between error-free Hebbian learning and error-driven Rescorla-Wagner-type learning has been carried over to debates between different DSMs in the literature. In particular, DSMs that are based on extracting temporally contiguous associations via error-free learning mechanisms to derive word meanings (e.g., HAL, LSA, BEAGLE, etc.) have been referred to as "count-based" models in computational linguistics and natural language processing, and have been contrasted against DSMs that employ a prediction-based mechanism to learn representations (e.g., word2vec, fastText, etc.), often referred to as "predict" models. It is important to note here that the count versus predict distinction is somewhat artificial and misleading, because even prediction-based DSMs effectively use co-occurrence *counts* of words from natural language corpora to generate predictions. The important difference between these models is therefore not that one class of models counts co-occurrences whereas the other predicts them, but in fact that one class of models employs an error-free Hebbian learning process whereas the other class of models employs a prediction-based error-driven learning process to learn direct and indirect associations between words. Nonetheless, in an influential paper, Baroni et al. (2014) compared 36 "count-based" or error-free learning-based DSMs to 48 "predict" or error-driven learning-based DSMs and concluded that error-driven learning-based (predictive) models significantly outperformed their Hebbian learning-based counterparts in a large battery of semantic tasks. Additionally, Mandera, Keuleers, and Brysbaert (2017) compared the relative performance of error-free learning-based DSMs (LSA and HAL-type) and error-driven learning-based models (CBOW and skip-gram versions of word2vec) on semantic priming tasks (Hutchison et al., 2013) and concluded that predictive models provided a better fit to the data. They also argued that predictive models are psychologically more plausible because they employ error-driven learning mechanisms consistent with principles posited by Rescorla and Wagner (1972) and are computationally more compact.

However, the argument that predictive models employ psychologically plausible learning mechanisms is incomplete, because error-free learning-based DSMs *also* employ equally plausible learning mechanisms, consistent with Hebbian learning principles. Further, there is also some evidence challenging the resounding success of predictive models. Asr, Willits, and Jones (2016) compared an error-free learning-based model (similar to HAL), a random vector accumulation model (similar to BEAGLE), and word2vec in their ability to acquire semantic categories when trained on child-directed speech data. Their results indicated that when the corpus was scaled down to stimulus available to children, the HAL-like model outperformed word2vec. Other work has also found little to no advantage of predictive models over error-free learning-based models (De Deyne, Perfors, & Navarro, 2016; Recchia & Nulty, 2017). Additionally, Levy, Goldberg, and Dagan (2015) showed that hyperparameters like window sizes, subsampling, and negative sampling can significantly affect performance, and it is not the case that predictive models are always superior to error-free learning-based models.

Collectively, these results point to two possibilities. First, it is possible that large amounts of training data (e.g., a billion words) and hyperparameter tuning (e.g., subsampling or negative sampling) are the main factors contributing to predictive models showing the reported gains in performance compared to their Hebbian learning counterparts. To address this possibility, Levy and Goldberg (2014) compared the computational algorithms underlying error-free learning-based models and predictive models and showed that the skip-gram word2vec model implicitly factorizes the word-context matrix, similar to several error-free learning-based models such as LSA. Therefore, it does appear that predictive models and error-free learning-based models may not be as different as initially conceived, and both approaches may actually converge on the same set of psychological principles. Second, it is possible that predictive models are indeed capturing a basic error-driven learning mechanism that humans use to perform certain types of complex tasks that require keeping track of sequential dependencies, such as sentence processing, reading comprehension, and event segmentation. Subsequent sections in this review discuss how state-of-the-art approaches specifically aimed at explaining performance in such complex semantic tasks are indeed variants or extensions of this prediction-based approach, suggesting that these models currently represent a promising and psychologically intuitive approach to semantic representation.

Language is clearly an extremely complex behavior, and even though modern DSMs like word2vec and GloVe that are trained on vast amounts of data successfully explain performance across a variety of tasks, adequate accounts of how humans generate sufficiently rich semantic representations with arguably lesser "data" are still missing from the field. Further, there appears to be relatively little work examining how newly trained models on smaller datasets (e.g., child-directed speech) compare to children's actual performance on semantic tasks. The majority of the work in machine learning and natural language processing

has focused on building models that outperform other models, or how the models compare to task benchmarks for only young adult populations. Therefore, it remains unclear how the mechanisms proposed by these models compare to the language acquisition and representation processes in humans, although subsequent sections make the case that recent attempts towards incorporating multimodal information, and temporal and attentional influences are making significant strides in this direction. Ultimately, it is possible that humans use multiple levels of representation and more than one mechanism to produce and maintain flexible semantic representations that can be widely applied across a wide range of tasks, and a brief review of how empirical work on context, attention, perception, and action has informed semantic models will provide a finer understanding on some of these issues.

II. Contextual and Retrieval-Based Semantic Memory

Despite the traditional notion of semantic memory being a "static" store of verbal knowledge about concepts, accumulating evidence within the past few decades suggests that semantic memory may actually be context-dependent. Consider the meaning of the word *ostrich*. Does the conceptualization of what the word *ostrich* means change when an individual is thinking about the size of different birds versus the types of eggs one could use to make an omelet? Although intuitively it appears that there is one "static" representation of *ostrich* that remains unchanged across different contexts, considerable evidence on the time course of sentence processing suggests otherwise. In particular, a large body of work has investigated how semantic representations come "online" during sentence comprehension and the extent to which these representations depend on the surrounding context. For example, there is evidence to show that the surrounding sentential context and the frequency of meaning may influence lexical access for ambiguous words (e.g., *bark* has a tree and sound-related meaning) at different timepoints (Swinney, 1979; Tabossi, Colombo, & Job, 1987). Furthermore, extensive work by Rayner and colleagues on eye movements in reading has shown that the frequency of different meanings of a word, the bias in the linguistic context, and preceding modifiers can modulate the extent to which multiple meanings of a word are automatically activated (Binder, 2003; Binder & Rayner, 1998; Duffy et al., 1988; Pacht & Rayner, 1993; Rayner, Cook, Juhasz, & Frazier, 2006; Rayner & Frazier, 1989). Collectively, this work is consistent with the two-process theories of attention (Neely, 1977; Posner & Snyder, 1975), according to which a *fast, automatic activation* process, as well as a *slow, conscious attention mechanism* are both at play during language-related tasks. The two-process theory can clearly account for findings like "automatic" facilitation in lexical decisions for words related to the dominant meaning of the ambiguous word in the presence of biasing context (Tabossi et al., 1987), and longer "conscious attentional" fixations on the ambiguous word when the context emphasizes the non-dominant meaning (Pacht & Rayner, 1993).

Another aspect of language processing is the ability to consciously attend to different parts of incoming linguistic input to form inferences on the fly. One line of evidence that speaks to this behavior comes from empirical work on reading and speech processing using the N400 component of event-related brain potentials (ERPs). The N400 component is thought to reflect contextual semantic processing, and sentences ending in unexpected words have been shown to elicit greater N400 amplitude compared to expected words, given a sentential context (e.g., Block & Baldwin, 2010; Federmeier & Kutas, 1999; Kutas & Hillyard, 1980). This body of work suggests that sentential context and semantic memory structure interact during sentence processing (see Federmeier & Kutas, 1999). Other work has examined the influence of local attention, context, and cognitive control during sentence comprehension. In an eye-tracking paradigm, Nozari, Trueswell, and Thompson-Schill (2016) had participants listen to a sentence (e.g., "She will cage the red lobster") as they viewed four colorless drawings. The drawings contained a local attractor (e.g., *cherry*) that was compatible with the closest adjective (e.g., *red*) but not the overall context, or an adjective-incompatible object (e.g., *igloo*). Context was manipulated by providing a verb that was highly constraining (e.g., *cage*) or non-constraining (e.g., *describe*). The results indicated that participants fixated on the local attractor in both constraining and non-constraining contexts, compared to incompatible control words, although fixation was smaller in more constrained contexts. Collectively, this work indicates that linguistic context and attentional processes interact and shape semantic memory representations, providing further evidence for automatic and attentional components (Neely, 1977; Posner & Snyder, 1975) involved in language processing.

Given these findings and the automatic-attentional framework, it is important to investigate how computational models of semantic memory handle ambiguity resolution (i.e., multiple meanings) and attentional influences, and depart from the traditional notion of a context-free "static" semantic memory store. Critically, DSMs that assume a static semantic memory store (e.g., LSA, GloVe, etc.) cannot straightforwardly account for the different contexts under which multiple meanings of a word are activated and suppressed, or how attending to specific linguistic contexts can influence the degree to which other related words are activated in the memory network. The following sections will further elaborate on this issue of ambiguity resolution and review some recent literature on modeling contextually dependent semantic representations.

Ambiguity resolution in error-free learning-based DSMs

Virtually all DSMs discussed so far construct a *single* representation of a word's meaning by aggregating statistical regularities across documents or contexts. This approach suffers from the drawback of collapsing multiple senses of a word into an "average" representation. For example, the homonym *bark* would be represented as a weighted average of its two meanings (the sound and the trunk), leading to a representation that is more biased towards the more dominant sense of the word. Homonyms (e.g., *bark*) and polysemes (e.g., *newspaper* may refer to the physical object or a national daily) represent over 40% of all English words (Britton, 1978; Durkin & Manning, 1989), and because DSMs do not appropriately model the non-dominant sense of a word, they tend to underperform in disambiguation tasks and also cannot appropriately model the behavior observed in sentence-processing tasks (e.g., Swinney, 1979). Indeed, Griffiths et al. (2007) have argued that the inability to model representations for polysemes and homonyms is a core challenge and may represent a key falsification criterion for certain distributional models (also see Jones, 2018). Early distributional models like LSA and HAL recognized this limitation of collapsing a word's meaning into a single representation. Landauer (2001) noted that LSA is indeed able to disambiguate word meanings when given surrounding context, i.e., neighboring words (for similar arguments see Burgess, 2001). To that end, Kintsch (2001) proposed an algorithm operating on LSA vectors that examined the local context around the target word to compute different senses of the word. While the approach of applying a process model over and above the core distributional model could be criticized, it is important to note that meaning is necessarily *distributed* across several dimensions in DSMs and therefore any process model operating on these vectors is using only information already contained within the vectors (see Günther et al., 2019, for a similar argument).

An alternative proposal to model semantic memory and also account for multiple meanings was put forth by Blei, Ng, and Jordan (2003) and Griffiths et al. (2007) in the form of *topic models* of semantic memory. In topic models, word meanings are represented as a distribution over a set of meaningful probabilistic topics, where the content of a topic is determined by the words to which it assigns high probabilities. For example, high probabilities for the words *desk*, *paper*, *board*, and *teacher* might indicate that the topic refers to a *classroom*, whereas high probabilities for the words *board*, *flight*, *bus*, and *baggage* might indicate that the topic refers to *travel*. Thus, in contrast to geometric DSMs where a word is represented as a point in a high-dimensional space, words (e.g., *board*) can have multiple representations across the different topics (e.g., *classroom*, *travel*) in a topic model. Importantly, topic models take the same word-document matrix as input as LSA and uncover latent "topics" in the same spirit of uncovering latent dimensions through an abstraction-based mechanism that goes over and above simply counting direct co-occurrences, albeit through different mechanisms, based on Markov Chain Monte Carlo methods (Griffiths & Steyvers, 2002, 2003, 2004). Topic models successfully account for free-association norms that show violations of symmetry, triangle inequality, and neighborhood structure (Tversky, 1977) that are problematic for other DSMs (but see Jones et al., 2018) and also outperform LSA in disambiguation, word prediction, and gist extraction tasks (Griffiths et al., 2007). However, the original architecture of topic models involved setting priors and specifying the number of topics a priori, which could lead to the possibility of experimenter bias in modeling (Jones, Willits, & Dennis, 2015). Further, the original topic model was essentially a "bag-of-words" model and did not capitalize on the sequential dependencies in natural language, like other DSMs (e.g., BEAGLE). Recent work by Andrews and Vigliocco (2010) has extended the topic model to incorporate word-order information, yielding more fine-grained linguistic representations that are sensitive to higher-order semantic relationships. Additionally, given that topic models represent word meanings as a distribution over a set of topics, they naturally account for multiple senses of a word without the need for an explicit process model, unlike other DSMs such as LSA or HAL (Griffiths et al., 2007).

Therefore, it appears that when DSMs are provided with appropriate context vectors through their representation (e.g., topic models) or additional assumptions (e.g., LSA), they are indeed able to account for patterns of polysemy and homonymy. Additionally, there has been a recent movement in natural language processing to build distributional models that can naturally tackle homonymy and polysemy. For example, Reisinger and Mooney (2010) used a clustering approach to construct sense-specific word embeddings that were successfully able to account for word similarity in isolation and within a sentential context. In their model, a word's contexts were clustered to produce different groups of similar context vectors, and these context vectors were then averaged into sense-specific vectors for the different clusters. A slightly different clustering approach was taken by Li and Jurafsky (2015), where the sense clusters and embeddings were jointly learned using a Bayesian non-parametric framework. Their model used the Chinese Restaurant Process, according to which a new sense vector for a word was computed when evidence from the context (e.g., neighboring and co-occurring words) suggested that it was sufficiently different from the existing senses. Li and Jurafsky indicated that their model successfully outperformed traditional embeddings on semantic relatedness tasks. Other work in this area has employed multilingual distributional information to generate different senses for words (Upadhyay, Chang, Taddy, Kalai,

& Zou, 2017), although the use of multiple languages to uncover word senses does not appear to be a psychologically plausible proposal for how humans derive word senses from language. Importantly, several of these recent approaches rely on error-free learning-based mechanisms to construct semantic representations that are sensitive to context. The following section describes some recent work in machine learning that has focused on error-driven learning mechanisms that can also adequately account for contextually-dependent semantic representations.

Ambiguity resolution in predictive DSMs

One particular drawback of multi-sense embeddings discussed above is that the meaning of a word can vary across multiple sentential contexts and enumerating all the different senses for a particular word can be both subjective (Westbury, 2016) and computationally expensive. For example, the word *star* can refer to its astronomical meaning, a film star, a rockstar, as well as an asterisk among other things, and the surrounding linguistic context itself may be more informative in understanding the meaning of the word *star*, instead of trying to enumerate all the different senses of *star*, which was the goal of multi-sense embeddings. The idea of using the sentential context itself to derive a word's meaning was first proposed in Elman's (1990) seminal work on the Simple Recurrent Network (SRN), where a set of *context* units that contained the previous hidden state of the neural network model served as "memory" for the next cycle. In this way, the internal representations that the SRN learned were sensitive to previously encountered linguistic context. This simple recurrent architecture successfully predicted word sequences, grammatical classes, and constituent structure in language (Elman, 1990, 1991). Modern Recurrent Neural Networks (RNNs) build upon the intuitions of the SRN and come in two architectures: Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU). LSTMs introduced the idea of memory cells, i.e., a vector that could preserve error signals over time and overcome the problem of vanishing error signals over long sequences (Hochreiter & Schmidhuber, 1997). Access to the memory cells is controlled through gates in LSTMs, where gate values are linear combinations of the current input and the previous model state. GRUs also have a gated architecture but differ in the number of gates and how they combine the hidden states (Olah, 2019). LSTMs and GRUs are currently the most successful types of RNNs and have been extensively applied to construct contextually sensitive, compositional (discussed in Section IV) models of semantic memory.

The RNN approach inspired Peters et al. (2018) to construct Embeddings from Language Models (ELMo), a modern version of recurrent neural networks (RNNs). Peters et al.'s ELMo model uses a bidirectional LSTM combined with a traditional NN language model to construct contextual word embeddings. Specifically, instead of explicitly training to predict predefined or empirically determined sense clusters, ELMo first tries to predict words in a sentence going sequentially forward and then backward, utilizing recurrent connections through a two-layer LSTM. The embeddings returned from these "pretrained" forward and backward LSTMs are then combined with a task-specific NN model to construct a task-specific representation (see Fig. 6). One key innovation in the ELMo model is that instead of only using the topmost layer produced by the LSTM, it computes a weighed linear combination of all three layers of the LSTM to construct the final semantic representation. The logic behind using all layers of the LSTM in ELMo is that this process yields very rich word representations, where higher-level LSTM states capture contextual aspects of word meaning and lower-level states capture syntax and parts of speech. Peters et al. showed that ELMo's unique architecture is successfully able to outperform other models in complex tasks like question answering, coreference resolution, and sentiment analysis among others. The success of recent recurrent models such as ELMo in tackling multiple senses of words represents a significant leap forward in modeling contextualized semantic representations.

Modern RNNs such as ELMo have been successful at predicting complex behavior because of their ability to incorporate previous states into semantic representations. However, one limitation of RNNs is that they encode the entire input sequence at once, which slows down processing and becomes problematic for extremely long sequences. For example, consider the task of text summarization, where the input is a body of text, and the task of the model is to paraphrase the original text. Intuitively, the model should be able to "attend" to specific parts of the text and create smaller "summaries" that effectively paraphrase the entire passage. This intuition inspired the *attention mechanism*, where "attention" could be focused on a subset of the original input units by weighting the input words based on positional and semantic information. The model would then predict target words based on relevant parts of the input sequence. Bahdanau, Cho, and Bengio (2014) first applied the attention mechanism to machine translation using two separate RNNs to first encode the input sequence and then used an attention head to explicitly focus on relevant words to generate the translated outputs. "Attention" was focused on specific words by computing an alignment score, to determine which input states were most relevant for the current time step and combining these weighted input states into a context vector. This context vector was then combined with the previous state of the model to generate the predicted output. Bahdanau et al. showed that the attention mechanism was able to outperform previous models in machine translation (e.g., Cho et al., 2014), especially for longer sentences.

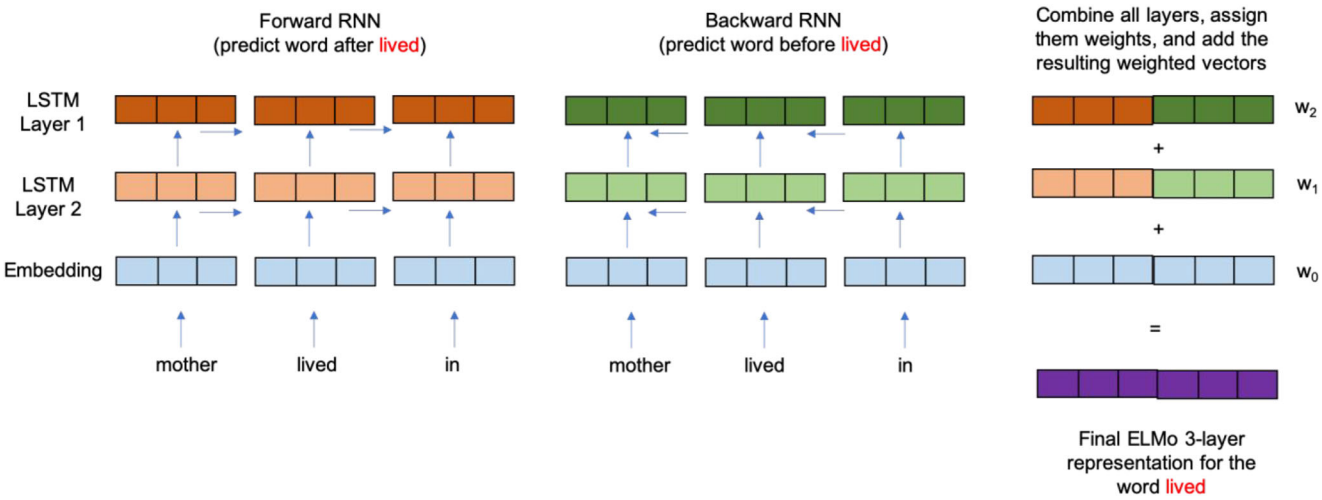

Fig. 6 A depiction of the ELMo architecture. The hidden layers of two long short-term memory networks (LSTMs; forward and backward) are first concatenated, followed by a weighted sum of the hidden layers with the embedding layer, resulting in the final three-layer representation for a particular word. Adapted from Alammar (2018)

Attention NNs are now at the heart of several state-of-the-art language models, like Google's Transformer (Vaswani et al., 2017), BERT (Devlin et al., 2019), OpenAI's GPT-2 (Radford et al., 2019) and GPT-3 (Brown et al., 2020), and Facebook's RoBERTa (Liu et al., 2019). Two key innovations in these new attention-based NNs have led to remarkable performance improvements in language-processing tasks. First, these models are being trained on a much larger scale than ever before, allowing them to learn from a billion iterations and over several days (e.g., Radford et al., 2019). Second, modern attention-NNs entirely eliminate the sequential recurrent connections that were central to RNNs. Instead, these models use multiple layers of attention and positional information to process words in parallel. In this way, they are able to focus attention on multiple words at a time to perform the task at hand. For example, Google's BERT model assigns position vectors to each word in a sentence. These position vectors are then updated using *attention* vectors, which represent a weighted sum of position vectors of other words and depend upon how strongly each position contributes to the word's representation. Specifically, attention vectors are computed using a compatibility function (similar to an alignment score in Bahdanau et al., 2014), which assigns a score to each pair of words indicating how strongly they should attend to one another. These computations iterate over several layers and iterations with the dual goal of predicting masked words in a sentence (e.g., I went to the [mask] to buy a [mask] of milk; predict *store* and *carton*) as well as deciding whether one sentence (e.g., They were out of reduced fat [mask], so I bought [mask] milk) is a valid continuation of another sentence (e.g., I went to the store to buy a carton of milk). By computing errors bidirectionally and updating the position and attention vectors with each iteration, BERT's word vectors are influenced by other words' vectors and tend to develop contextually dependent word embeddings. For example, the representation of the word *ostrich* in the BERT model would be different when it is in a sentence about birds (e.g., *ostriches* and emus are large birds) versus food (*ostrich* eggs can be used to make omelets), due to the different position and attention vectors contributing to these two representations. Importantly, the architecture of BERT allows it to be flexibly finetuned and applied to any semantic task, while still using the basic attention-based mechanism. This framework has turned out to be remarkably efficient and models based on the general Transformer architecture (e.g., BERT, RoBERTa, GPT-2, & GPT-3) outperform LSTM-based recurrent approaches in semantic tasks such as sentiment analysis (Socher et al., 2013), sentence acceptability judgments (Warstadt, Singh, & Bowman, 2018), and even tasks that are dependent on semantic and world knowledge, such as the Winograd Schema Challenge (Levesque, Davis, & Morgenstern, 2012) or novel language generation (Brown et al., 2020). However, considerable work is beginning to evaluate these models using more rigorous test cases and starting to question whether these models are actually learning anything meaningful (e.g., Brown et al., 2020; Niven & Kao, 2019), an issue that is discussed in detail in Section V.

Although the technical complexity of attention-based NNs makes it difficult to understand the underlying *mechanisms* contributing to their impressive success, some recent work has attempted to demystify these models (e.g., Clark, Khandelwal, Levy, & Manning, 2019; Coenen et al., 2019; Michel, Levy, & Neubig, 2019; Tenney, Das, & Pavlick, 2019). For example, Clark et al. (2019) recently showed that BERT's attention heads actually attend to meaningful semantic and syntactic information in sentences, such as determiners, objects of verbs, and co-referent mentions (see Fig. 7), suggesting that these models may indeed be capturing

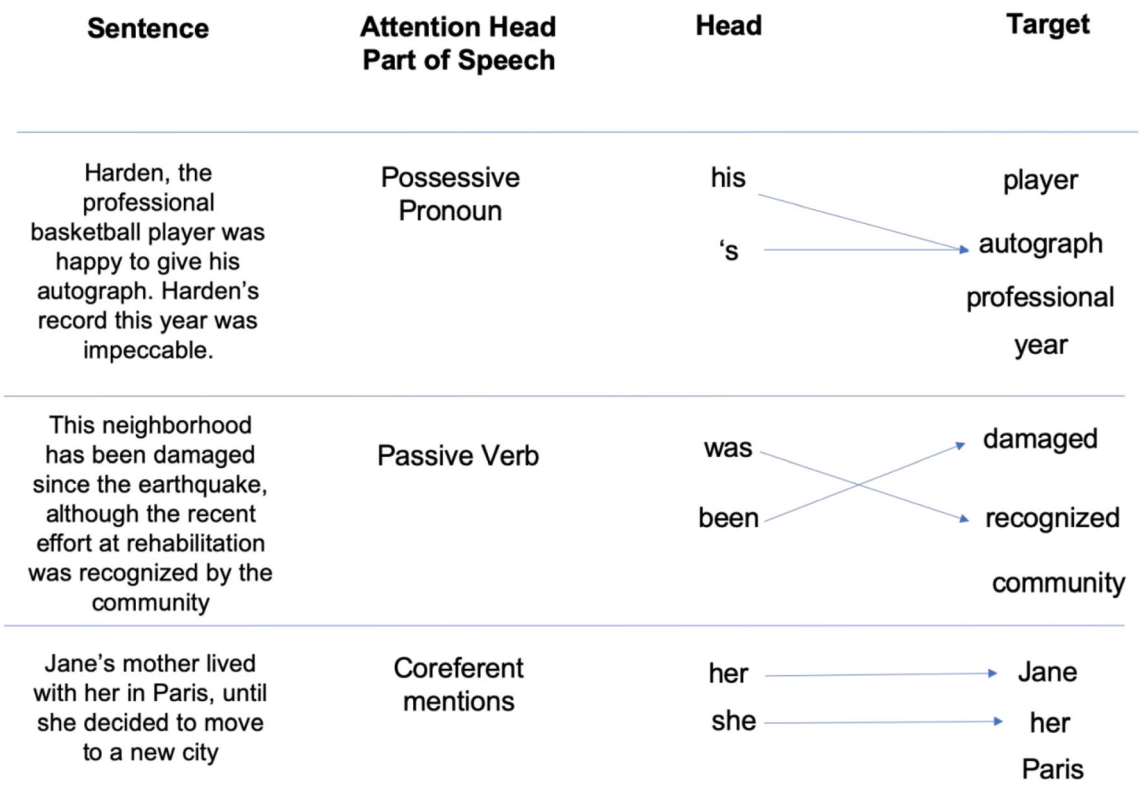

Fig. 7 BERT attention heads that correspond to linguistic phenomena like attending to noun phrases and verbs. Arrows indicate specific relationships that the heads are attending to within each sentence. Adapted from Clark et al. (2019)

meaningful linguistic knowledge, which may be driving their performance. Further, some recent evidence also shows that BERT successfully captures phrase-level representations, indicating that BERT may indeed have the ability to model compositional structures (Jawahar, Sagot, & Seddah, 2019), although this work is currently in its nascent stages. Furthermore, it remains unclear how this conceptualization of attention fits with the automatic-attentional framework (Neely, 1977). Demystifying the inner workings of attention NNs and focusing on process-based accounts of how computational models may explain cognitive phenomena clearly represents the next step towards integrating these recent computational advances with empirical work in cognitive psychology.

Collectively, these recent approaches to construct contextually sensitive semantic representations (through recurrent and attention-based NNs) are showing unprecedented success at addressing the bottlenecks regarding polysemy, attentional influences, and context that were considered problematic for earlier DSMs. An important insight that is common to both contextualized RNNs and attention-based NNs discussed above is the idea of *contextualized* semantic representations, a notion that is certainly at odds with the traditional conceptualization of context-free semantic memory. Indeed, the following section discusses a new class of models take this notion a step further by entirely eliminating the need for learning representations or "semantic memory" and propose that *all* meaning representations may in fact be retrieval-based, therefore blurring the historical distinction between episodic and semantic memory.

Retrieval-based models of semantic memory

Tulving's (1972) episodic-semantic dichotomy inspired foundational research on semantic memory and laid the groundwork for conceptualizing semantic memory as a static memory store of facts and verbal knowledge that was distinct from episodic memory, which was linked to events situated in specific times and places. However, some recent attempts at modeling semantic memory have taken a different perspective on how meaning representations are constructed. Retrieval-based models challenge the strict distinction between semantic and episodic memory, by constructing semantic representations through retrieval-based processes operating on episodic experiences. Retrieval-based models are based on Hintzman's (1988) MINERVA 2 model, which was originally proposed to explain how individuals learn to categorize concepts. Hintzman argued that humans store all instances or episodes that they experience, and that categorization of a new concept is simply a weighted function of its similarity to these stored instances at the time of retrieval. In other words, each episodic experience lays down a trace, which implies that if an item is

presented multiple times, it has multiple traces. At the time of retrieval, traces are activated in proportion to its similarity with the retrieval cue or probe. For example, an individual may have seen an *ostrich* in pictures or at the zoo multiple times and would store each of these instances in memory. The next time an *ostrich*-like bird is encountered by this individual, they would match the features of this bird to a weighted sum of all stored instances of *ostrich* and compute the similarity between these features to decide whether the new bird is indeed an *ostrich*. Hintzman's work was crucial in developing the exemplar theory of categorization, which is often contrasted against the prototype theory of categorization (Rosch & Mervis, 1975), which suggests that individuals "learn" or generate an abstract prototypical representation of a concept (e.g., *ostrich*) and compare new examples to this prototype to organize concepts into categories. Importantly, Hintzman's model rejected the need for a strong distinction between episodic and semantic memory (Tulving, 1972) and has inspired a class of models of semantic memory often referred to as *retrieval-based models*.

Kwantes (2005) proposed a retrieval-based alternative to LSA-type distributional models by computing semantic representations "on the fly" from a term-document matrix of episodic experiences. Based on principles from Hintzman's (1988) MINERVA 2 model, in Kwantes' model, each word has a context vector (i.e., memory trace) associated with it, which contains its frequency of occurrence within each document of the training corpus. When a word is encountered in the environment, it is used as a cue to retrieve the context vector, which activates the traces of all words in lexical memory. The activation of a trace is directly proportional to the contextual similarity between their context vectors. Memory traces are then weighted by their activations and summed across the context vectors to construct the final semantic representation of the target word. The resulting semantic representations from Kwantes' model successfully captured higher-order semantic relationships, similar to LSA, without the need for storing, abstracting, or learning these representations at the time of encoding.

Modern retrieval-based models have been successful at explaining complex linguistic and behavioral phenomena, such as grammatical constraints (Johns & Jones, 2015) and free association (Howard et al., 2011), and certainly represent a significant departure from the models discussed thus far. For example, Howard et al. (2011) proposed a model that constructed semantic representations using temporal context. Instead of defining context in terms of a sentence or document like most DSMs, the Predictive Temporal Context Model (pTCM; see also Howard & Kahana, 2002) proposes a continuous representation of temporal context that gradually changes over time. Items in the pTCM are activated to the extent that their encoded context overlaps with the context that is cued. Further, context is also used to *predict* items that are likely to appear next, and the semantic representation of an item is the collection of prediction vectors in which it appears over time. These previously *learned* prediction vectors also contribute to the word's future representations. Howard et al. showed that the pTCM successfully simulates human performance in word-association tasks and is able to capture long-range dependencies in language that are problematic for other DSMs. In its core principles of constructing representations from episodic contexts, the pTCM is similar to other retrieval-based models, but its ability to *learn* from previous states and gradually accumulate information also shares similarities with the SRN (Elman, 1990), BEAGLE (Jones & Mewhort, 2007), and some of the recent error-driven learning DSMs discussed in Section II (e.g., word2vec, ELMo, etc.).

More recently, Jamieson, Avery, Johns, and Jones et al. (2018) proposed an instance-based theory of semantic memory, also based on MINERVA 2. In their model, word contexts are stored as *n*-dimensional vectors representing multiple instances in episodic memory. Memory of a document (or conversation) is the sum of all word vectors, and a "memory" vector stores all documents in a single vector. A word's meaning is retrieved by cueing the memory vector with a probe, which activates each trace in proportion to its similarity to the probe. The aggregate of all activated traces is called an echo, where the contribution of a trace is directly weighted by its activation. The retrieved echo, in response to a probe, is assumed to represent a word's meaning. Therefore, the model exhibits "context sensitivity" by comparing the activations of the retrieval probe with the activations of other traces in memory, thus producing context-dependent semantic representations without any mechanism for learning these representations. For example, Jamieson et al. showed that for the homograph *break* (with three senses, related to stopping, smashing, and news reporting), when their model is provided with a disambiguating context using a joint probe (e.g., *break*/*car*), the retrieved representation (or "echo") is more similar to the word *stop*, compared to the words *report* and *smash*, thus producing a context-dependent semantic representation of the word *break*. Therefore, Jamieson et al.'s model successfully accounts for some findings pertaining to ambiguity resolution that have been difficult to accommodate within traditional DSM-based accounts and proposes that meaning is created "on the fly" and in response to a retrieval cue, an idea that is certainly inconsistent with traditional semantic models.

Summary

Although it is well understood that prior knowledge or semantic memory influences how individuals perceive events (e.g., Bransford & Johnson, 1972; Deese, 1959; Roediger & McDermott, 1995), the notion that semantic memory may itself be influenced by episodic events is relatively recent. This section discussed how the conceptualization of semantic

memory of being an independent and static memory store is slowly changing, in light of evidence that context shapes the structure of semantic memory. Retrieval-based models represent an important departure from the traditional notions about semantic memory, and instead propose that the meaning of a word is computed "on the fly" at retrieval, and do not subscribe to the idea of storing or learning a static semantic representation of a concept. This conceptualization is clearly at odds with traditional accounts of semantic memory and hearkens back to the distinction between prototype and exemplar theories of categorization briefly eluded to earlier. Specifically, in the computational models of semantic memory discussed so far (with the exception of retrieval-based models), the idea of inferring indirect co-occurrences and/or latent dimensions, i.e., *learning through abstraction* emerges as a core mechanism contributing to the construction of meaning. This idea of abstraction has also been central to computational models that have been applied to understand category structure. Specifically, *prototype* theories (Rosch, Mervis, Gray, Johnson, & Boyes-Braem, 1976; Rosch & Lloyd, 1978; also see Posner & Keele, 1968) posit that as individual concepts are experienced, humans gradually develop a prototypical representation that contains the most useful and representative information about that category. This notion of constructing an abstracted, prototypical representation is at the heart of several computational models of semantic memory discussed in this review. For example, both LSA and BEAGLE construct an "average" prototypical semantic representation from individual linguistic experiences. Of course, LSA uses a term-document matrix and singular value decomposition whereas BEAGLE learns meaning by incrementally combining co-occurrence and order information to compute a composite representation, but both models represent a word as a single point (prototype) in a high-dimensional space. Retrieval-based models, on the other hand, are inspired by Hintzman's work and the exemplar theory of categorization and assume that semantic representations are constructed in response to retrieval cues and reject the idea of prototypical representations or abstraction-like learning processes occurring at the time of encoding. Given the success of retrieval-based models at tackling ambiguity and several other linguistic phenomena, these models clearly represent a powerful proposal for how meaning is constructed.

However, before abstraction (at encoding) can be rejected as a plausible mechanism underlying meaning computation, retrieval-based models need to address several bottlenecks, only one of which is computational complexity. Jones et al. (2018) recently noted that computational constraints should not influence our preference of traditional prototype models over exemplar-based models, especially since exemplar models have provided better fits to categorization task data, compared to prototype models (Ashby & Maddox, 1993; Nosofsky, 1988; Stanton, Nosofsky, & Zaki, 2002). However, implementation is a core test for theoretical models and retrieval-based models must be able to explain how the brain manages this computational overhead. Specifically, retrieval-based models argue against any type of "semantic memory" at all and instead propose that semantic representations are *created* "on the fly" when words or concepts are encountered within a particular context. As discussed earlier, while there is evidence to suggest that the representations likely change with every new encounter (e.g., for a review, see Yee et al., 2018), it is still unclear why the brain would create a fresh new representation for a particular concept "on the fly" each time that concept is encountered, and not "learn" something about the concept from previous encounters that could aid future processing. It seems more psychologically plausible that the brain learns and maintains a semantic representation (stored via changes in synaptic activity; see Mayford, Siegelbaum, & Kandel, 2012) that is subsequently finetuned or modified with each new incoming encounter – a proposal that is closer to the mechanisms underlying recurrent and attention-NNs discussed earlier in this section. Furthermore, in light of findings that top-down information or previous knowledge does in fact guide cognitive behavior (e.g., Bransford & Johnson, 1972; Deese, 1959; Roediger & McDermott, 1995) and bottom-up processes interact with top-down processes (Neisser, 1976), the proposal that there may not be any existing semantic structures in place at all certainly requires more investigation.

It is important to note here that individual traces for episodic events may indeed need to be stored by the system for other cognitive tasks, but the argument here is that retrieving the meaning of a concept need not necessarily require the storage of every individual experience or trace. For example, consider the simple memory task of remembering a list of words: *train*, *ostrich*, *lemon*, and *truth*. Encoding a representation of this event likely involves laying down a trace of this experience in memory. However, retrieval-based models would posit that the representation of the word *ostrich* in this context would in fact be a weighted sum of every other time the word or concept of *ostrich* has been experienced, all of which have been stored in memory. This conceptualization seems unnecessary, especially given that other DSMs that instead use more compact learning-based representations have been fairly successful at simulating performance in semantic as well as non-semantic tasks (for a model of LSA-type semantic structures applied to free recall tasks, see Polyn, Norman, & Kahana, 2009).

Additionally, it appears that retrieval-based models currently lack a complete account of how long-term sequential dependencies, sentential context, and multimodal information might simultaneously influence the computation of meaning. For example, how does multimodal information about an object get stored in retrieval-based models – does each individual sensorimotor encounter also leave its own trace in memory and contribute to the "context-specific" representation or is

the scope of "context" limited to patterns of co-occurrence? Further, it remains unclear how representations derived from retrieval-based models differ from representations derived from modern RNNs and attention-based NNs, which also propose contextualized representations. It appears that these classes of models share similarities in their fundamental claim that the retrieval context determines the representation of a concept or word, although retrieval-based models do not subscribe to any particular learning mechanism (with the exception of Howard et al.' s predictive pTCM model), whereas RNNs and attention-NNs are based on error-driven learning mechanisms. Specifically, RNNs and attention-NNs learn via prediction and incrementally build semantic representations, whereas retrieval-based models instead propose that representations are constructed solely at the time of retrieval, without any learning occurring at the time of exposure or encoding. Furthermore, while RNNs and attention-NNs take word order and positional information (e.g., bidirectionality in BERT) into account within their definition of "context" when constructing semantic representations, it appears that recent retrieval-based models currently lack mechanisms to incorporate word order into their representations (e.g., Jamieson et al., 2018), even though this may simply be a practical limitation at this point.

Finally, it is unclear how retrieval-based models would scale up to sentences, paragraphs, and other higher-order structures like events, issues that are being successfully addressed by other learning-based DSMs (see Sections III and IV). Clearly, more research is needed to adequately assess the relative performance of retrieval-based models, compared to state-of-the-art learning-based models of semantic memory currently being widely applied in the literature to a large collection of semantic (and non-semantic) tasks. Collectively, it seems most likely that humans store individual exemplars in some form (e.g., a distributed pattern of activation) or at least to some extent (e.g., storing only traces above a certain threshold of stable activation), but also learn a prototypical representation as consistent exemplars are experienced, which facilitates faster top-down processing (for a similar argument, see Yee et al., 2018) in cognitive tasks, although this issue clearly needs to be explored further.

The central idea that emerged in this section is that semantic memory representations may indeed vary across contexts. The accumulating evidence that meaning rapidly changes with linguistic context certainly necessitates models that can incorporate this flexibility into word representations. Attention-based NNs like BERT and GPT-2/3 represent a promising step towards constructing such contextualized, attention-based representations and appear to be consistent with the automatic and attentional components of language processing (Neely, 1977), although more work is needed to clarify *how* these models compute meaningful representations that can be flexibly applied across different tasks. The success of attention-based NNs is truly impressive on one hand but also cause for concern on the other. First, it is remarkable that the underlying mechanisms proposed by these models at least appear to be psychologically intuitive and consistent with empirical work showing that attentional processes and predictive signals do indeed contribute to semantic task performance (e.g., Nozari et al., 2016). However, if the ultimate goal is to build models that explain and mirror human cognition, the issues of scale and complexity cannot be ignored. Current state-of-the-art models operate at a scale of word exposure that is much larger than what young adults are typically exposed to (De Deyne, Perfors, & Navarro, 2016; Lake, Ullman, Tenenbaum, & Gershman, 2017). Therefore, exactly how humans perform the same semantic tasks without the large amounts of data available to these models remains unknown. One line of reasoning is that while humans have lesser *linguistic* input compared to the corpora that modern semantic models are trained on, humans instead have access to a plethora of *non-linguistic* sensory and environmental input, which is likely contributing to their semantic representations. Indeed, the following section discusses how conceptualizing semantic memory as a multimodal system sensitive to perceptual input represents the next big paradigm shift in the study of semantic memory.

III. Grounding Models of Semantic Memory

Virtually all distributional and network-based semantic models rely on large text corpora or databases to construct semantic representations. Consequently, a consistent and powerful criticism of distributional semantic models comes from the grounded cognition movement (Barsalou, 2016), which rejects the idea that meaning can be represented through abstract and amodal symbols like words in a language. Instead, grounded cognition researchers posit that sensorimotor modalities, the environment, and the body all contribute and play a functional role in cognitive processing, and by extension, the construction of meaning. Grounded (or embodied) cognition is a rather broad enterprise that attempts to redefine the study of cognition (Matheson & Barsalou, 2018). Within the domain of semantic memory, distributional models in particular have been criticized because they derive semantic representations from only linguistic texts and are not grounded in perception and action, leading to the symbol grounding problem (Harnad, 1990; Searle, 1980), i.e., how can the meaning of a word (e.g., an *ostrich*) be grounded only in other words (e.g., *big*, *bird*, etc.) that are further grounded in more words?

While there is no one theory of grounded cognition (Matheson & Barsalou, 2018), the central tenet common to several of them is that the body, brain, and physical environment dynamically interact to produce meaning and cognitive behavior. For example, based on Barsalou's account

(Barsalou, 1999, 2003, 2008), when an individual first encounters an object or experience (e.g., a *knife*), it is stored in the modalities (e.g., its shape in the visual modality, its sharpness in the tactile modality, etc.) and the sensorimotor system (e.g., how it is used as a weapon or kitchen utensil). Repeated co-occurrences of physical stimulations result in functional associations (likely mediated by associative Hebbian learning and/or connectionist mechanisms) that form a multimodal representation of the object or experience (Matheson & Barsalou, 2018). Features of these representations are activated through recurrent connections, which produces a simulation of past experiences. These simulations not only guide an individual's ongoing behavior retroactively (e.g., how to dice onions with a *knife*), but also proactively influence their future or imagined plans of action (e.g., how one might use a *knife* in a fight). Simulations are assumed to be neither conscious nor complete (Barsalou, 2003; Barsalou & Wiemer-Hastings, 2005), and are sensitive to cognitive and social contexts (Lebois, Wilson-Mendenhall, & Barsalou, 2015).

There is some empirical support for the grounded cognition perspective from sensorimotor priming studies. In particular, there is substantial evidence that modality-specific neural information is activated during language-processing tasks. For example, it has been demonstrated that reading verbs like *kick* (corresponding to feet) or *pick* (corresponding to hand) activates the motor cortex in a somatotopic fashion (Pulvermüller, 2005), passive reading of taste-related words (e.g., *salt*) activates gustatory cortices (Barros-Loscertales et al., 2011), and verifying modality-specific properties of words (e.g., color, taste, sound, and touch) activates the corresponding sensory brain regions (Goldberg, Perfetti, & Schneider, 2006). However, whether the activation of modality-specific information is *incidental* to the task and simply a result of post-representation processes, or actually *part* of the semantic representation itself is an important question. Support for the latter argument comes from studies showing that transcranial stimulation of areas in the premotor cortex related to the hand facilitates lexical decision performance for hand-related action words (Willems, Labruna, D'Esposito, Ivry, & Casasanto, 2011), Parkinson's patients show selective impairment in comprehending motor action words (Fernandino et al., 2013), and damage to brain regions supporting object-related action can hinder access to *knowledge* about how objects are manipulated (Yee, Chrysikou, Hoffman, & Thompson-Schill, 2013). Yee et al. also showed that when individuals performed a concurrent manual task while naming pictures, there was more naming interference for objects that are more manually used (e.g., *pencils*), compared to objects that are not typically manually used (e.g., *tigers*). Furthermore, Yee, Huffstetler, and Thompson-Schill (2011) used a visual eye-tracking paradigm to show that as an object unfolds over time (e.g., auditorily hearing *frisbee*), particular features (e.g., form-related) come online in a temporally constrained fashion and can influence eye fixation times for related words (e.g., e.g., participants fixated longer on *pizza*, because *frisbee* and *pizza* are both round). Taken together, these findings suggest that semantic memory representations are accessed in a dynamic way during tasks and different perceptual features of these representations may be accessed at different timepoints, suggesting a more flexible and fluid conceptualization (also see Yee, Lahiri, & Kotzor, 2017) of semantic memory that can change as a function of task. Therefore, it is important to evaluate whether computational models of semantic memory can indeed encode these rich, non-linguistic features as part of their representations.

It is important to note here that while the sensorimotor studies discussed above provide support for the grounded cognition argument, these studies are often limited in scope to processing sensorimotor words and do not make specific predictions about the direction of effects (Matheson & Barsalou, 2018; Matheson, White, & McMullen, 2015). For example, although several studies show that modality-specific information is activated during behavioral tasks, it remains unclear whether this activation leads to facilitation or inhibition within a cognitive task. Indeed, both types of findings are taken to support the grounded cognition view, therefore leading to a lack of specificity in predictions regarding the role of modality-specific information (Matheson et al., 2015), although some recent work has proposed that timing of activation may be critical in determining how modality-specific activation influences cognitive performance (Matheson & Barsalou, 2018). Another strong critique of the grounded cognition view is that it has difficulties accounting for how abstract concepts (e.g., *love*, *freedom* etc.) that do not have any grounding in perceptual experience are acquired or can possibly be simulated (Dove, 2011). Some researchers have attempted to "ground" abstract concepts in metaphors (Lakoff & Johnson, 1999), emotional or internal states (Vigliocco et al., 2013), or temporally distributed events and situations (Barsalou & Wiemer-Hastings, 2005), but the mechanistic account for the acquisition of abstract concepts is still an active area of research. Finally, there is a dearth of formal models that provide specific mechanisms by which features acquired by the sensorimotor system might be combined into a coherent concept. Some accounts suggest that semantic representations may be created by patterns of synchronized neural activity, which may represent different sensorimotor information (Schneider, Debener, Oostenveld, & Engel, 2008). Other work has suggested that certain regions of the cortex may serve as "hubs" or "convergence zones" that combine features into coherent representations (Patterson, Nestor, & Rogers, 2007), and may reflect temporally synchronous activity within areas to which the features belong (Damasio, 1989). However, comparisons of such approaches to DSMs remain limited due to the lack of formal grounded models, although there have been some recent attempts at modeling perceptual schemas (Pezzulo & Calvi, 2011) and Hebbian learning (Garagnani & Pulvermüller, 2016).

Proponents of the grounded cognition view have also presented empirical (Glenberg & Robertson, 2000; Rubinstein, Levi, Schwartz, & Rappoport, 2015) and theoretical criticisms (Barsalou, 2003; Perfetti, 1998) of DSMs over the years. For example, Glenberg and Robertson (2000) reported three experiments to argue that high-dimensional space models like LSA/HAL are inadequate theories of meaning, because they fail to distinguish between sensible (e.g., filling an old sweater with leaves) and nonsensical sentences (e.g., filling an old sweater with water) based on cosine similarity between words (but see Burgess, 2000). Some recent work also shows that traditional DSMs trained solely on linguistic corpora do indeed lack salient features and attributes of concepts. Baroni and Lenci (2008) compared a model analogous to LSA with attributes derived from McRae, Cree, Seidenberg, and McNorgan (2005) and an image-based dataset. They provided evidence that DSMs entirely miss external (e.g., a *car* <has wheels>) and surface level (e.g., a *banana* <is yellow>) properties of objects, and instead focus on taxonomic (e.g., *cat-dog*) and situational relations (e.g., *spoon-bowl*), which are more frequently encountered in natural language. More recently, Rubinstein et al. (2015) evaluated four computational models, including word2vec and GloVE, and showed that DSMs are poor at classifying attributive properties (e.g., an *elephant* <is large>), but relatively good at classifying taxonomic properties (e.g., *apple* <is a> *fruit*) identified by human subjects in a property generation task (also see Collell & Moens, 2016; Lucy & Gauthier, 2017).

Collectively, these studies appear to underscore the intuitions of the grounded cognition researchers that semantic models based solely on linguistic sources do not produce sufficiently rich representations. While this is true, it is important to realize here that the failure of DSMs to encode these perceptual features is a function of the training corpora they are exposed to, i.e., a *practical* limitation, and not necessarily a *theoretical* one. Early DSMs were trained on linguistic corpora not because it was intrinsic to the theoretical assumptions made by the models, but because text corpora were easily available (for more fleshed-out arguments on this issue, see Burgess, 2000; Günther et al., 2019; Landauer & Dumais, 1997). Therefore, the more important question is whether DSMs can be adequately trained to derive statistical regularities from other sources of information (e.g., visual, haptic, auditory etc.), and whether such DSMs can effectively incorporate these signals to construct "grounded" semantic representations.

Grounding DSMs through feature integration

The lack of grounding in standard DSMs led to a resurging interest in early feature-based models (McRae et al., 1997; Smith et al., 1974). As discussed earlier, early feature-based models represented words as a collection of binary features (e.g., *birds* have wings, whereas *cars* do not), and words with similar meanings had greater overlap in their constituent features (McCloskey & Glucksberg, 1979; Smith et al., 1974; Tversky, 1977), although these early models did not have explicit mechanisms to account for how features were learned in the first place. However, one important strength of feature-based models was that the features encoded could directly be interpreted as placeholders for grounded sensorimotor experiences (Baroni & Lenci, 2008). For example, the representation of a *banana* is distributed across several hundred dimensions in a distributional approach, and these dimensions may or may not be interpretable (Jones, Willits, & Dennis, 2015), but the perceptual experience of the *banana's* color being *yellow* can be directly encoded in feature-based models (e.g., *banana* <is yellow>).

However, it is important to note here that, again, the fact that features can be verbalized and are more interpretable compared to dimensions in a DSM is a result of the features having been extracted from property generation norms, compared to textual corpora. Therefore, it is possible that some of the information captured by property generation norms may already be encoded in DSMs, albeit through less interpretable dimensions. Indeed, a systematic comparison of feature-based and distributional models by Riordan and Jones (2011) demonstrated that representations derived from DSMs produced comparable categorical structure to feature representations generated by humans, and the type of information encoded by both types of models was highly correlated but also complementary. For example, DSMs gave more weight to actions and situations (e.g., *eat*, *fly*, *swim*) that are frequently encountered in the linguistic environment, whereas feature-based representations were better are capturing object-specific features (e.g., <is yellow>, <made of metal>) that potentially reflected early sensorimotor experiences with objects. Riordan and Jones argued that children may be more likely to initially extract information from sensorimotor experiences. However, as they acquire more linguistic experience, they may shift to extracting the redundant information from the distributional structure of language and rely on perception for only novel concepts or the unique sources of information it provides. This idea is consistent with the symbol interdependency hypothesis (Louwerse, 2011), which proposes that while words must be grounded in the sensorimotor action and perception, they also maintain rich connections with each other at the symbolic level, which allows for more efficient language processing by making it possible to skip grounded simulations when unnecessary. The notion that both sources of information are critical to the construction of meaning presents a promising approach to reconciling distributional models with the grounded cognition view of language (for similar accounts, see Barsalou, Santos, Simmons, & Wilson, 2008; Paivio, 1991).

Recent work in computational modeling has attempted to integrate featural information with distributional information

to enrich semantic representations. For example, Andrews et al. (2009) used a Bayesian probabilistic topic model to jointly model semantic representations using experiential feature-based (e.g., an *ostrich* <is big>, <does not fly>, <has feathers> etc.) and linguistic (e.g., *ostrich* and *emu* co-occur) data as complementary sources of information. Further, Vigliocco, Meteyard, Andrews, and Kousta (2009) argued that affective and internal states can serve as another data source that could potentially enrich semantic representations, particularly for abstract concepts that lack sensorimotor associations (Kousta, Vigliocco, Vinson, Andrews, & Del Campo, 2011). The information integration approach has also been applied to other types of DSMs. For example, Jones and Recchia (2010) integrated feature-based information with BEAGLE to show that temporal linguistic information plays a critical role in generating accurate semantic representations. Johns and Jones (2012) have also explored the integration of perceptual information with linguistic information based on simple associative mechanisms, borrowing principles from Hintzman's (1988) MINERVA architecture and Kwantes' (2005) model. Their model provided a proof of concept that perceptually rich semantic representations may be constructed by grounding them in already formed or learned representations of other words (accessible via feature norms). This notion of grounding representations in previously learned words has also been explored by Howell et al. (2005) using a recurrent NN model. Using a modified version of the Elman's (1990) SRN with two additional output layers for noun and verb features, Howell et al. trained the model to map phonetically presented input words (nouns) to semantic features and perform a grammatical word prediction task. Howell et al. argued that this type of learning mechanism could be applied to simulate a "propagation of grounding" effect, where sensorimotor information from early, concrete words acquired by children feeds into semantic representations of novel words, although this proposal was not formally tested in the paper. Other work on integrating featural information has explored training a recurrent NN model with sensorimotor feature inputs and patterns of co-occurrence to account for a wide variety of behavioral patterns consistent with normal and impaired semantic cognition (Hoffman et al., 2018), implementing a feedforward NN to apply feature learning to a simple word-word co-occurrence model (Durda, Buchanan, & Caron, 2009) and using feature-based vectors as input to a random-vector accumulation model (Vigliocco, Vinson, Lewis, & Garrett, 2004).

Multimodal DSMs

Despite their considerable success, an important limitation of feature-integrated distributional models is that the perceptual features available are often restricted to small datasets (e.g., 541 concrete nouns from McRae et al., 2005), although some recent work has attempted to collect a larger dataset of feature norms (e.g., 4436 concepts; Buchanan, Valentine, & Maxwell, 2019). Moreover, the features produced in property generation tasks are potentially prone to saliency biases (e.g., hardly any participant will produce the feature <has a head> for a *dog* because having a head is not salient or distinctive), and thus can only serve as an incomplete proxy for all the features encoded by the brain. To address these concerns, Bruni et al. (2014) applied advanced computer vision techniques to automatically extract visual and linguistic features from multimodal corpora to construct multimodal distributional semantic representations. Using a technique called "bag-of-visual-words" (Sivic & Zisserman, 2003), the model discretized visual images and produced visual units comparable to words in a text document. The resulting image matrix was then concatenated with a textual matrix constructed from a natural language corpus using singular value decomposition to yield a multimodal semantic representation. Bruni et al. showed that this model was superior to a purely text-based approach and successfully predicted semantic relations between related words (e.g., *ostrich-emu*) and clustering of words into superordinate concepts (e.g., *ostrich-bird*).

This multimodal approach to semantic representation is currently a thriving area of research (Feng & Lapata, 2010; Kiela & Bottou, 2014; Lazaridou et al., 2015; Silberer & Lapata, 2012, 2014). Advances in the machine-learning community have majorly contributed to accelerating the development of these models. In particular, Convolutional Neural Networks (CNNs) were introduced as a powerful and robust approach for automatically extracting meaningful information from images, visual scenes, and longer text sequences. The central idea behind CNNs is to apply a non-linear function (a "filter") to a sliding window of the full chunk of information, e.g., pixels in an image, words in a sentence, etc. The filter transforms the larger window of information into a fixed *d*-dimensional vector, which captures the important properties of the pixels or words in that window. Convolution is followed by a "pooling" step, where vectors from different windows are combined into a single *d*-dimensional vector, by taking the maximum or average value of each of the *d*-dimensions across the windows. This process extracts the most important features from a larger set of pixels (see Fig. 8), or the most informative *k*-grams in a long sentence. CNNs have been flexibly applied to different semantic tasks like sentiment analysis and machine translation (Collobert et al., 2011; Kalchbrenner, Grefenstette, & Blunsom, 2014), and are currently being used to develop multimodal semantic models.

Kiela and Bottou (2014) applied CNNs to extract the most meaningful features from images from a large image database (ImageNet; Deng et al., 2009) and then concatenated these image vectors with linguistic word2vec vectors to produce superior semantic representations compared to Bruni et al. (2014); also see Silberer & Lapata, 2014). Lazaridou et al.

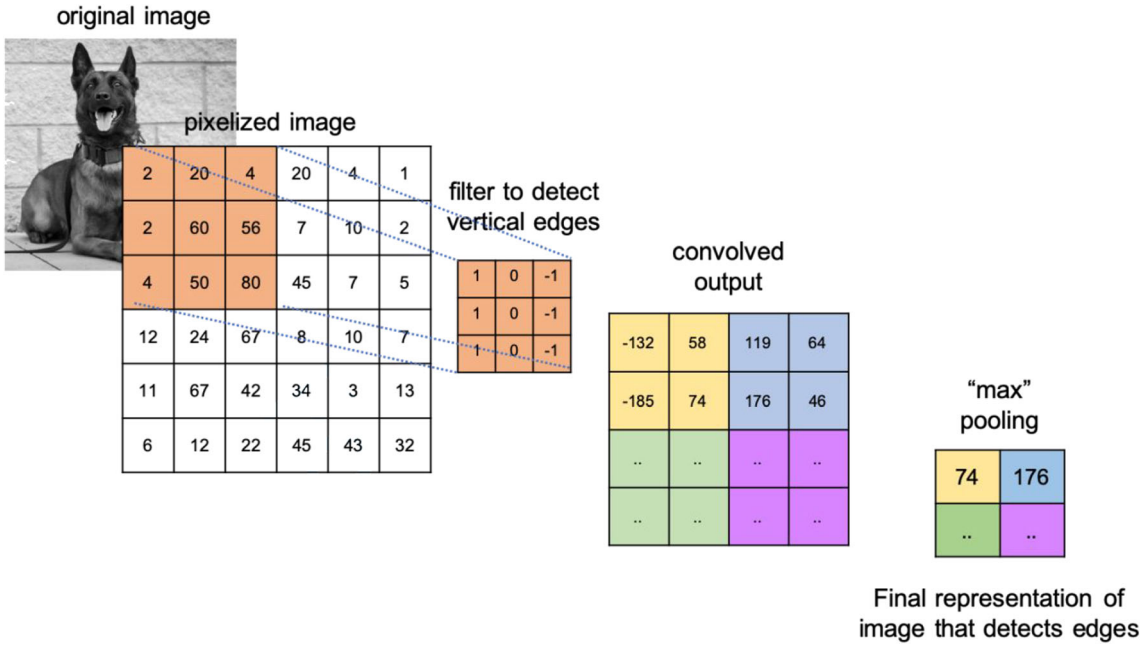

Fig. 8 A depiction of a typical convolutional neural network that detects vertical edges in an image. A sliding filter is multiplied with the pixelized image to produce a matrix, and then a pooling step combines results from the convolved output into a smaller matrix by selecting the maximum value from each 2 × 2 sub-matrix in the convolved matrix. This final 2 × 2 matrix represents the final representation of the image highlighting the vertical edges

(2015) constructed a multimodal word2vec model that was trained to jointly learn visual and semantic representations for a subset of words (using image-based CNNs and word2vec), and this learning was then generalized to the entire corpus, thus echoing Howell et al.'s (2005) intuitions of "propagation of grounding." Lazaridou et al. also demonstrated how the learning of abstract words might be grounded in concrete scenes (e.g., *freedom* might be the inferred concept from a scene of a person raising their hands in a protest), an intuitively powerful proposal that can potentially demystify the acquisition of abstract concepts but clearly needs further exploration.

There is also some work within the domain of associative network models of semantic memory that has focused on integrating different sources of information to construct the semantic networks. One particular line of research has investigated combining word-association norms with featural information, co-occurrence information, and phonological similarity to form multiplex networks (Stella, Beckage, & Brede, 2017; Stella, Beckage, Brede, & De Domenico, 2018). Stella et al. (2017) demonstrated that the "layers" in such a multiplex network differentially influence language acquisition, with all layers contributing equally initially but the association layer overtaking the word learning process with time. This proposal is similar to the ideas presented earlier regarding how perceptual or sensorimotor experience might be important for grounding words acquired earlier, and words acquired later might benefit from and derive their representations through semantic associations with these early experiences (Howell et al., 2005; Riordan & Jones, 2011). In this sense, one can think of phonological information and featural information providing the necessary grounding to early acquired concepts. This "grounding" then propagates and enriches semantic associations, which are easier to access as the vocabulary size increases and individuals develop more complex semantic representations.

Summary

Given the success of integrated and multimodal DSMs memory that use state-of-the-art modeling techniques to incorporate other modalities to augment linguistic representations, it appears that the claim that semantic models are "amodal" and "ungrounded" may need to be revisited. Indeed, the fact that multimodal semantic models can adequately encode perceptual features (Bruni et al., 2014; Kiela & Bottou, 2014) and can approximate human judgments of taxonomic and visual similarity (Lazaridou et al., 2015), suggests that the limitations of previous models (e.g., LSA, HAL etc.) were more practical than theoretical. Of course, incorporating other modalities besides vision is critical to this enterprise, and although there have been some efforts to integrate sound and olfactory data into semantic representations (Kiela, Bulat, & Clark, 2015; Kiela & Clark, 2015; Lopopolo & Miltenburg, 2015), these approaches are limited by the availability of large datasets that capture other aspects of embodiment that may be critical for meaning construction, such as touch, emotion, and

taste. Investing resources in collecting and archiving multimodal datasets (e.g., video data) is an important next step for advancing research in semantic modeling and broadening our understanding of the many facets that contribute to the construction of meaning.

IV. Compositional Semantic Representations

An additional aspect of extending our understanding of meaning by incorporating other sources of information is that meaning may be situated within and as part of higher-order semantic structures like sentence models, event models, or schemas. Indeed, language is inherently compositional in that morphemes combine to form words, words combine to form phrases, and phrases combine to form sentences. Moreover, behavioral evidence from sentential priming studies indicates that the meaning of words depends on complex syntactic relations (Morris, 1994). Further, it is well known that the meaning of a sentence itself is not merely the sum of the words it contains. For example, the sentence "John loves Mary" has a different meaning to "Mary loves John," despite both sentences having the same words. Thus, it is important to consider how compositionality can be incorporated into and inform existing models of semantic memory.

Compositional linguistic approaches

Associative network models do not have any explicit way of modeling compositionality, as they propose representations at the word level that cannot be straightforwardly scaled to higher-order semantic structures. On the other hand, distributional models have attempted to build compositionality into semantic representations by assigning roles to different entities in sentences (e.g., in "Mary loves John," Mary is the *lover* and John is the *lovee*; Dennis, 2004, 2005), treating frequent phrases as single units and deriving phrase-based representations (e.g., treating proper names like *New York* as a single unit; Bannard, Baldwin, & Lascarides, 2003; Mikolov, Sutskever, et al., 2013) or forming pair-pattern matrices (e.g., encoding words that fulfil the pattern *X cuts Y,* i.e., *mason*: *stone*; Turney & Pantel, 2010). However, these approaches were either not scalable for longer phrases or lacked the ability to model constituent parts separately (Mitchell & Lapata, 2010). Vector addition (or averaging) is another common method of combining distributional semantic representations for different words to form higher-order vectors (Landauer & Dumais, 1997), but this method is insensitive to word order and syntax and produces a blend that does not appropriately extract meaningful information from the constituent words (Mitchell & Lapata, 2010).

An alternative method of combining word-level vectors is through a matrix multiplication technique called *tensor products*. Tensor products are a way of computing pairwise products of the component word vector elements (Clark, Coecke, & Sadrzadeh, 2008; Clark & Pulman, 2007; Widdows, 2008), but this approach suffers from the curse of dimensionality, i.e., the resulting product matrix becomes very large as more individual vectors are combined. *Circular convolution* is a special case of tensor products that compresses the resulting product of individual word vectors into the same dimensionality (e.g., Jones & Mewhort, 2007). In a systematic review, Mitchell and Lapata (2010) examined several compositional functions applied onto a simple high-dimensional space model and a topic model space in a phrase similarity rating task (judging similarity for phrases like *vast amount-large amount*, *start work-begin career*, *good place-high point*, etc.). Specifically, they examined how different methods of combining word-level vectors (e.g., addition, multiplication, pairwise multiplication using tensor products, circular convolution, etc.) compared in their ability to explain performance in the phrase similarity task. Their findings indicated that *dilation* (a function that amplified some dimensions of a word when combined with another word, by differentially weighting the vector products between the two words) performed consistently well in both spaces, and circular convolution was the least successful in judging phrase similarity. This work sheds light on how simple compositional operations (like tensor products or circular convolution) may not sufficiently mimic human behavior in compositional tasks and may require modeling more complex interactions between words (i.e., functions that emphasize different aspects of a word).

Recent efforts in the machine-learning community have also attempted to tackle semantic compositionality using Recursive NNs. *Recursive* NNs represent a generalization of recurrent NNs that, given a syntactic parse-tree representation of a sentence, can generate hierarchical tree-like semantic representations by combining individual words in a recursive manner (conditional on how probable the composition would be). For example, Socher, Huval, Manning, and Ng (2012) proposed a recursive NN to compute compositional meaning representations. In their model, each word is assigned a vector that captures its meaning and also a matrix that contains information about how it modifies the meaning of another word. This representation for each word is then recursively combined with other words using a non-linear composition function (an extension of work by Mitchell & Lapata, 2010). For example, in the first iteration, the words *very* and *good* may be combined into a representation (e.g., *very good*), which would recursively be combined with *movie* to produce the final representation (e.g., *very good movie*). Socher et al. showed that this model successfully learned propositional logic, how adverbs and adjectives modified nouns, sentiment classification, and complex semantic relationships (also see Socher et al., 2013). Other work in this area has explored multiplication-

based models (Yessenalina & Cardie, 2011), LSTM models (Zhu, Sobhani, & Guo, 2016), and paraphrase-supervised models (Saluja, Dyer, & Ruvini, 2018). Collectively, this research indicates that modeling the sentence structure through NN models and recursively applying composition functions can indeed produce compositional semantic representations that are achieving state-of-the-art performance in some semantic tasks.

Compositional Event Representations

Another critical aspect of modeling compositionality is being able to extend representations at the word or sentence level to higher-level cognitive structures like events or situations. The notion of schemas as a higher-level, structured representation of knowledge has been shown to guide language comprehension (Schank & Abelson, 1977; for reviews, see Rumelhart, 1991) and event memory (Bower, Black, & Turner, 1979; Hard, Tversky, & Lang, 2006). The past few years have seen promising advances in the field of event cognition (Elman & McRae, 2019; Franklin et al., 2019; Reynolds, Zacks, & Braver, 2007; Schapiro, Rogers, Cordova, Turk-Browne, & Botvinick, 2013). Importantly, while most event-based accounts have been conceptual, recent computational models have attempted to explicitly specify processes that might govern event knowledge. For example, Elman and McRae (2019) recently proposed a recurrent NN model of event knowledge, trained on activity sequences that make up events. An activity was defined as a collection of agents, patients, actions, instruments, states, and contexts, each of which were supplied as inputs to the network. The task of the network was to learn the internal structure of an activity (i.e., which features correlate with a particular activity) and also predict the next activity in sequence. Elman and McRae showed that this network was able to infer the co-occurrence dynamics of activities, and also predict sequential activity sequences for new events. For example, when presented with the activity sequence, "The crowd looks around. The skater goes to the podium. The audience applauds. The skater receives a ___", the network activated the words *podium* and *medal* after the fourth sentence ("The skater receives a") because both of these are contextually appropriate (receiving an award at the *podium* and receiving a *medal*), although *medal* was more activated than *podium* as it was more appropriate within that context. This behavior of the model was strikingly consistent with N400 amplitudes observed for the same types of sentences in an ERP study (Metusalem et al., 2012), indicating that the model was able to make predictive inferences like human participants.

Franklin et al. (2019) recently proposed a probabilistic model of event cognition. In their model, each visual scene had a distributed vector representation, encoding the features that are relevant to the scene, which were learned using an unsupervised CNN. Additionally, scenes contained relational information that linked specific roles to specific fillers via circular convolution. A four-layer fully connected NN with Gated Recurrent Units (GRUs; a type of recurrent NN) was then trained to predict successive scenes in the model. Using the Chinese Restaurant Process, at each timepoint, the model evaluated its prediction error to decide if its current event representation was still a good fit. If the prediction error was high, the model chose whether it should switch to a different previously-learned event representation or create an entirely new event representation, by tuning parameters to evaluate total number of events and event durations. Franklin et al. showed that their model successfully learned complex event dynamics and simulated a wide variety of empirical phenomena. For example, the model's ability to predict event boundaries from unannotated video data (Zacks, Kurby, Eisenberg, & Haroutunian, 2011) of a person completing everyday tasks like washing dishes, was highly correlated with grouped participant data and also produced similar levels of prediction error across event boundaries as human participants.

Summary

This section reviewed some early and recent work at modeling compositionality, by building higher-level representations such as sentences and events, through lower-level units such as words or discrete time points in video data. One important limitation of the event models described above is that they are not models of *semantic memory* per se, in that they neither contain rich semantic representations as input (Franklin et al., 2019), nor do they explicitly model how linguistic or perceptual input might be integrated to learn concepts (Elman & McRae, 2019). Therefore, while there have been advances in modeling word and sentence-level semantic representations (Sections I and II), and at the same time, there has been work on modeling how individuals experience events (Section IV), there appears to be a gap in the literature as far as integrating word-level semantic structures with event-level representations is concerned. Given the advances in language modeling discussed in this review, the integration of structured semantic knowledge (e.g., recursive NNs), multimodal semantic models, and models of event knowledge discussed in this review represents a promising avenue for future research that would enhance our understanding of how semantic memory is organized to represent higher-level knowledge structures. Another promising line of research in the direction of bridging this gap comes from the artificial intelligence literature, where neural network agents are being trained to learn language in a simulated grid world full of perceptual and linguistic information (Bahdanau et al., 2018; Hermann et al., 2017) using reinforcement learning principles. Indeed, McClelland, Hill, Rudolph, Baldridge, and Schütze (2019) recently advocated the need to situate language within a larger cognitive system. Conceptualizing semantic memory as part of a broader

integrated memory system consisting of objects, situations, and the social world is certainly important for the success of the semantic modeling enterprise.

V. Open Issues and Future Directions

The question of how concepts are represented, stored, and retrieved is fundamental to the study of all cognition. Over the past few decades, advances in the fields of psychology, computational linguistics, and computer science have truly transformed the study of semantic memory. This paper reviewed classic and modern models of semantic memory that have attempted to provide explicit accounts of how semantic knowledge may be acquired, maintained, and used in cognitive tasks to guide behavior. Table 1 presents a short summary of the different types of models discussed in this review, along with their basic underlying mechanisms. In this concluding section, some open questions and potential avenues for future research in the field of semantic modeling will be discussed.

Data availability and abundance

Within the context of semantic modeling, data is a double-edged sword. On one hand, the availability of training data in the form of large text corpora such as Wikipedia articles, Google News corpora, etc. has led to an explosion of models such as word2vec (Mikolov, Chen, et al., 2013), fastText (Bojanowski et al., 2017), GLoVe (Pennington et al., 2014), and ELMo (Peters et al., 2018), which have outperformed several standard models of semantic memory traditionally trained on lesser data. Additionally, with the advent of computational resources to quickly process even larger volumes of data using parallel computing, models such as BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020) are achieving unprecedented success in language tasks like question answering, reading comprehension, and language generation. At the same time, however, criticisms of ungrounded distributional models have led to the emergence of a new class of "grounded" distributional models. These models automatically derive non-linguistic information from other modalities like vision and speech using convolutional neural networks (CNNs) to construct richer representations of concepts. Even so, these grounded models are limited by the availability of multimodal sources of data, and consequently there have been recent efforts at advocating the need for constructing larger databases of multimodal data (Günther et al., 2019).

On the other hand, training models on more data is only part of the solution. As discussed earlier, if models trained on several gigabytes of data perform as well as young adults who were exposed to far fewer training examples, it tells us little about human language and cognition. The field currently lacks systematic accounts for how humans can flexible use language in different ways with the impoverished data they are exposed to. For example, children can generalize their knowledge of concepts fairly easily from relatively sparse data when learning language, and only require a few examples of a concept before they understand its meaning (Carey & Bartlett, 1978; Landau, Smith, & Jones, 1988; Xu & Tenenbaum, 2007). Furthermore, both children and young adults can rapidly learn new information from a single training example, a phenomenon referred to as *one-shot learning*. To address this particular challenge, several researchers are now building models than can exhibit *few-shot learning*, i.e., learning concepts from only a few examples, or *zero-shot learning*, i.e., generalizing already acquired information to never-seen before data. Some of these approaches utilize pretrained models like GPT-2 and GPT-3 trained on very large datasets and generalizing their architecture to new tasks (Brown et al., 2020; Radford et al., 2019). While this approach is promising, it appears to be circular because it still uses vast amounts of data to build the initial pretrained representations. Other work in this area has attempted to implement one-shot learning using Bayesian generative principles (Lake, Salakhutdinov, & Tenenbaum, 2015), and it remains to be seen how probabilistic semantic representations account for the generative and creative nature of human language.

Errors and degradation in language processing

Another striking aspect of the human language system is its tendency to break down and produce errors during cognitive tasks. Analyzing errors in language tasks provides important cues about the mechanics of the language system. Indeed, there is considerable work on tip-of-the-tongue experiences (James & Burke, 2000; Kumar, Balota, Habbert, Scaltritti, & Maddox, 2019), speech errors (Dell, 1990), errors in reading (Clay, 1968), language deficits (Hodges & Patterson, 2007; Shallice, 1988), and age-related differences in language tasks (Abrams & Farrell, 2011), to suggest that the cognitive system is prone to interference, degradation, and variability. However, computational accounts for how language may be influenced by interference or degradation remain limited. Early connectionist models did provide ways of lesioning the network to account for neuropsychological deficits such as dyslexia (Hinton & Shallice, 1991; Plaut & Shallice, 1993) and category-specific semantic deficits (Farah & McClelland, 2013), and this general approach has recently been extended to train a recurrent NN based on sensorimotor and co-occurrence-based information and simulate behavioral patterns observed in patients of semantic dementia and semantic aphasia (Hoffman et al., 2018). However, current state-of-the-art language models like word2vec, BERT, and GPT-2 or GPT-3 do not provide explicit accounts for how neuropsychological deficits may arise, or how systematic speech and

Table 1 Modern computational models of semantic memory

Group	Model class	Representative papers	Input	Mechanism
Network-based	Association networks	De Deyne & Storms, 2008; Kenett et al., 2011; Steyvers & Tenenbaum, 2005	Free-association norms	Words connected by edges (semantic relationships)
	Multiplex networks	Stella et al., 2017, 2018	Free-association norms, features, co-occurrence, phonology	Integrate different sources to produce multi-level network
Feature-based	Feature-integrated models	Andrews et al. (2009); Jones and Recchia (2010); Howell et al. (2005)	Explicitly coded features & words	Overlap of features determines semantic overlap between words
Distributional Semantic Models (DSMs)	Error-free (Hebbian) Learning-based DSMs	Jones & Mewhort, 2007 (BEAGLE); Landauer & Dumais, 1997 (LSA); Lund & Burgess, 1996 (HAL)	Words in a text corpus	Co-occurrence matrix, often transformed by SVD or random vector accumulation
	Error-driven Learning-based (Predictive) DSMs	Neural embedding models (Mikolov, Chen, et al., 2013, Mikolov, Sutskever, et al., 2013 (word2vec), Bojanowski et al., 2017 (fastText))	Words in a text corpus	Train neural network (NN)-based word vectors to perform semantic task
		Topic Models (Blei et al. (2003); Griffiths et al., 2007)	Words in a text corpus	Word-by-document matrix, with dimensionality reduction
		Convolutional Neural Networks (CNNs; Collobert et al., 2011; Kalchbrenner et al., 2014)	Natural images, scenes, or text	Extract features from input using matrix operations and NNs
		Recurrent NNs (e.g., LSTMs, GRUs, Recursive NNs; Peters et al., 2018 (ELMo); Socher et al., 2013)	Word sequences or syntactic parse trees	Store previous state of NN model as recurrent connections to inform future predictions
		Attention-NNs (e.g., Bahdanau et al., 2014; Vaswani et al., 2017; Radford et al., 2019 (GPT-2); Brown et al., 2020 (GPT-3); Devlin et al., 2019 (BERT))	Words in a text corpus	Use attention "heads" or vectors in NNs to focus on different parts of input while minimizing error
	Retrieval-based	Jamieson et al., 2018; Kwantes, 2005;	Words in a text corpus	Store each individual word occurrence and perform abstraction-at-retrieval cued by a probe
	Hybrid DSMs	GloVe (Pennington et al., 2014)	Words in a text corpus	Uses global co-occurrence and prediction
		pTCM (Howard et al., 2011)	Words in a text corpus	Uses retrieval-based operations and prediction vectors
		Multimodal DSMs (e.g., Kiela & Bottou, 2014; Lazaridou et al., 2015)	Words and images	Combine image and textual embeddings

reading errors are produced. Furthermore, while there is considerable empirical work investigating age-related differences in language-processing tasks (e.g., speech errors, picture naming performance, lexical retrieval, etc.), it is unclear how current semantic models would account for these age-related changes, although some recent work has compared the semantic network structure between older and younger adults (Dubossarsky, De Deyne, & Hills, 2017; Wulff, Hills, & Mata, 2018). Indeed, the deterministic nature of modern machine-learning models is drastically different from the stochastic nature of human language that is prone to errors and variability (Kurach et al., 2019). Computational accounts of how the language system produces and recovers from errors will be an important part of building machine-learning models that can mimic human language.

Communication, social collaboration, and evolution

Another important aspect of language learning is that humans actively learn from each other and through interactions with their social counterparts, whereas the majority of computational language models assume that learners are simply processing incoming information in a passive manner (Günther et al., 2019). Indeed, there is now ample evidence to suggest that language evolved through natural selection for the purposes of gathering and sharing information (Pinker, 2003, p. 27; DeVore & Tooby, 1987), thereby allowing for personal experiences and episodic information to be shared among humans (Corballis, 2017a, 2017b). Consequently, understanding how artificial and human learners may communicate and collaborate in complex tasks is currently an active area of research. For example, some recent work in natural language processing has attempted to model interactions and search processes in collaborative language games, such as Codenames (Kumar, Steyvers, & Balota, under review; Shen, Hofer, Felbo, & Levy, 2018, also see Kim, Ruzmaykin, Truong, & Summerville, 2019), Password (Xu & Kemp, 2010), and navigational games (Wang, Liang, & Manning, 2016), and suggested that speakers and listeners do indeed calibrate their responses based on feedback from their conversational partner. Another body of work currently being led by technology giants like Google and OpenAI is focused on modeling interactions in multiplayer games like football (Kurach et al., 2019) and Dota 2 (OpenAI, 2019). This work is primarily based on *reinforcement learning* principles, where the goal is to train neural network agents to interact with their environment and perform complex tasks (Sutton & Barto, 1998). Although these research efforts are less language-focused, deep reinforcement learning models have also been proposed to specifically investigate language learning. For example, Li et al. (2016) trained a conversational agent using reinforcement learning, and a reward metric based on whether the dialogues generated by the model were easily answerable, informative, and coherent. Other learning-based models have used adversarial training, a method by which a model is trained to produce responses that would be indistinguishable from human responses (Li et al., 2017), a modern version of the Turing test (also see Spranger, Pauw, Loetzsch, & Steels, 2012). However, these recent attempts are still focused on *independent* learning, whereas psychological and linguistic research suggests that language evolved for purposes of sharing information, which likely has implications for how language is learned in the first place. Clearly, this line of work is currently in its nascent stages and requires additional research to fully understand and model the role of communication and collaboration in developing semantic knowledge.

Multilingual semantic models

A computational model can only be considered a model of semantic memory if it can be broadly applied to *any* semantic memory system and does not depend on the specific language of training. Therefore, an important challenge for computational semantic models is to be able to generalize the basic mechanisms of building semantic representations from English corpora to other languages. Some recent work has applied character-level CNNs to learn the rich morphological structure of languages like Arabic, French, and Russian (Kim, Jernite, Sontag, & Rush, 2016; also see Botha & Blunsom, 2014; Luong, Socher, & Manning, 2013). These approaches clearly suggest that pure word-level models that have occupied centerstage in the English language modeling community may not work as well in other languages, and subword information may in fact be critical in the language learning process. More recent embeddings like *fastText* (Bojanowski et al., 2017) that are trained on sub-lexical units are a promising step in this direction. Furthermore, constructing multilingual word embeddings that can represent words from multiple languages in a single distributional space is currently a thriving area of research in the machine-learning community (e.g., Chen & Cardie, 2018; Lample, Conneau, Ranzato, Denoyer, & Jégou, 2018). Overall, evaluating modern machine-learning models on other languages can provide important insights about language learning and is therefore critical to the success of the language modeling enterprise.

Revisiting benchmarks for semantic models

A critical issue that has not received adequate attention in the semantic modeling field is the quality and nature of benchmark test datasets that are often considered the final word for comparing state-of-the-art machine-learning-based language models. The General Language Understanding Evaluation (GLUE; Wang et al., 2018) benchmark was recently proposed as a collection of language-based task datasets, including the

Corpus of Linguistic Acceptability (CoLA; Warstadt et al., 2018), the Stanford Sentiment Treebank (Socher et al., 2013), and the Winograd Schema Challenge (Levesque, Davis, & Morgenstern, 2012), among a total of 11 language tasks. Other popular benchmarks in the field include decaNLP (McCann, Keskar, Xiong, & Socher, 2018), the Stanford Question Answering Dataset (SQuAD; Rajpurkar et al., 2018), Word Similarity Test Collection (WordSim-33; Finkelstein et al., 2002) among others. While these benchmarks offer a standardized method of comparing performance across models, several of the tasks included within these benchmark datasets either consist of crowdsourced information collected from an unknown number of participants (e.g., SQuAD), scores or annotations based on very few human participants (e.g., 16 participants assessed similarity for 200 word-pairs in the WordSim-33 dataset), or sometimes datasets with no established human benchmark (e.g., the GLUE Diagnostic dataset, Wang et al., 2018). This is in contrast to more psychologically motivated models (e.g., semantic network models, BEAGLE, Temporal Context Model, etc.), where model performance is often compared against human baselines, for example in predicting accuracy or response latencies to perform a particular task, or through large-scale normed databases of human performance in semantic tasks (e.g., English Lexicon Project; Balota et al., 2007; Semantic Priming Project; Hutchison et al., 2013). Therefore, to evaluate whether state-of-the-art machine learning models like ELMo, BERT, and GPT-2 are indeed plausible psychological models of semantic memory, it is important to not only establish human baselines for benchmark tasks in the machine-learning community, but also explicitly compare model performance to human baselines in both accuracy and response times.

There have been some recent efforts in this direction. For example, Bender (2015) tested over 400 Amazon Mechanical Turk users on the Winograd Schema Challenge (a task that requires the use of world knowledge, commonsense reasoning and anaphora resolution) and provided quantitative baselines for accuracy and response times that should provide useful benchmarks to compare machine-learning models in the extent to which they explain human behavior (also see Morgenstern, Davis, & Ortiz, 2016). Further, Chen et al. (2017) compared the performance of the word2vec model against human baselines of solving analogies using relational similarity judgments to show that word2vec successfully captures only a subset of analogy relations. Additionally, Lazaridou, Marelli, and Baroni (2017) recently compared the performance of their multimodal skip-gram model (Lazaridou et al., 2015) against human relatedness judgments to visual and word cues for newly learned concepts to show that the model performed very similar to human participants. Despite these promising studies, such efforts remain limited due to the goals of machine learning often being *application*-focused and the goals of psychology being *explanation*-focused. Explicitly comparing model performance to behavioral task performance represents an important next step towards reconciling these two fields, and also combining representational and process-based accounts of how semantic memory guides cognitive behavior.

Prioritizing mechanistic accounts

Despite the lack of systematic comparisons to human baselines, an important takeaway that emerges from this review is that several state-of-the-art language models such as word2vec (Mikolov, Chen, et al., 2013, Mikolov, Sutskever, et al., 2013), ELMo (Peters et al., 2018), BERT (Devlin et al., 2019), GPT-2 (Radford et al., 2019), and GPT-3 (Brown et al., 2020) do indeed show impressive performance across a wide variety of semantic tasks such as summarization, question answering, and sentiment analysis. However, despite their success, relatively little is known about *how* these models are able to produce this complex behavior, and exactly *what* is being learned by them in their process of building semantic representations. Indeed, there is some skepticism in the field about whether these models are truly *learning* something meaningful or simply exploiting spurious statistical cues in language, which may or may not reflect human learning. For example, Niven and Kao (2019) recently evaluated BERT's performance in a complex argument-reasoning comprehension task, where world knowledge was critical for evaluating a particular claim. For example, to evaluate the strength of the claim "Google is not a harmful monopoly," an individual may *reason* that "people can choose not to use Google," and also provide the additional *warrant* that "other search engines do not redirect to Google" to argue in favor of the claim. On the other hand, if the *alternative*, "all other search engines redirect to Google" is true, then the claim would be false. Niven and Kao found that BERT was able to achieve state-of-the-art performance with 77% accuracy in this task, without any explicit world knowledge. For example, knowing what a monopoly might mean in this context (i.e., restricting consumer choices) and that Google is a search engine are critical pieces of knowledge required to evaluate the claim. Further analysis showed that BERT was simply exploiting statistical cues in the *warrant* (i.e., the word "not") to evaluate the claim, and once this cue was removed through an adversarial test dataset, BERT's performance dropped to chance levels (53%). The authors concluded that BERT was not able to learn anything *meaningful* about argument comprehension, even though the model performed better than other LSTM and vector-based models and was only a few points below the human baseline on the original task (also see Zellers, Holtzman, Bisk, Farhadi, & Choi, 2019, for a similar demonstration on a commonsense-based inference task).

These results are especially important if state-of-the-art models like word2vec, ELMo, BERT or GPT-2/3 are to be

considered plausible models of semantic memory in any manner and certainly underscore the need to focus on *mechanistic* accounts of model behavior. Understanding *how* machine-learning models arrive at answers to complex semantic problems is as important as simply evaluating how many questions the model was able to answer. Humans not only extract complex statistical regularities from natural language and the environment, but also form semantic structures of world knowledge that influence their behavior in tasks like complex inference and argument reasoning. Therefore, explicitly testing machine-learning models on the specific *knowledge* they have acquired will become extremely important in ensuring that the models are truly *learning* meaning and not simply exhibiting the "Clever Hans" effect (Heinzerling, 2019). To that end, explicit *process-based accounts* that shed light on the cognitive processes operating on underlying semantic representations across different semantic tasks may be useful in evaluating the psychological plausibility of different models. For instance, while distributional models perform well on a broad range of semantic tasks on average (Bullinaria & Levy, 2007; Mandera et al., 2017), it is unclear why their performance is better on tasks like synonym detection (Bullinaria & Levy, 2007) and similarity judgments (Bruni et al., 2014) and worse for semantic priming effects (Hutchison, Balota, Cortese, & Watson, 2008; Mandera et al., 2017), free association (Griffiths et al., 2007; Kenett et al., 2017), and complex inference tasks (Niven & Kao, 2019). A promising step towards understanding *how* distributional models may dynamically influence task performance was taken by Rotaru, Vigliocco, and Frank (2018), who recently showed that combining semantic network-based representations derived from LSA, GloVe, and word2vec with a dynamic spreading-activation framework significantly improved the predictive power of the models on semantic tasks. In light of this work, testing competing process-based models (e.g., spreading activation, drift-diffusion, temporal context, etc.) and structural or representational accounts of semantic memory (e.g., prediction-based, topic models, etc.) represents the next step in fully understanding how *structure* and *processes* interact to produce complex behavior.

Conclusion

The nature of knowledge representation and the processes used to retrieve that knowledge in response to a given task will continue to be the center of considerable theoretical and empirical work across multiple fields including philosophy, linguistics, psychology, computer science, and cognitive neuroscience. The ultimate goal of semantic modeling is to propose *one* architecture that can simultaneously integrate perceptual and linguistic input to form meaningful semantic representations, which in turn naturally scales up to higher-order semantic structures, and also performs well in a wide range of cognitive tasks. Given the recent advances in developing multimodal DSMs, interpretable and generative topic models, and attention-based semantic models, this goal at least appears to be achievable. However, some important challenges still need to be addressed before the field will be able to integrate these approaches and design a unified architecture. For example, addressing challenges like one-shot learning, language-related errors and deficits, the role of social interactions, and the lack of process-based accounts will be important in furthering research in the field. Although the current modeling enterprise has come very far in decoding the statistical regularities humans use to learn meaning from the linguistic and perceptual environment, no single model has been successfully able to account for the flexible and innumerable ways in which humans acquire and retrieve knowledge. Ultimately, integrating lessons learned from behavioral studies showing the interaction of world knowledge, linguistic and environmental context, and attention in complex cognitive tasks with computational techniques that focus on quantifying association, abstraction, and prediction will be critical in developing a complete theory of language.

Author's Note I sincerely thank David A. Balota, Jeffrey M. Zacks, Michael N. Jones, and Ian G. Dobbins for their extremely insightful feedback and helpful comments on earlier versions of the manuscript.

Open Practices Statement Given the theoretical nature of this review, no data or program code is available. However, Table 1 provides a succinct summary of the key models discussed in this review.

References

Abbott, J. T., Austerweil, J. L., & Griffiths, T. L. (2015). Random walks on semantic networks can resemble optimal foraging. In *Neural Information Processing Systems Conference. 22*(3). 558. American Psychological Association.

Abrams, L., & Farrell, M. T. (2011). Language processing in normal aging. *The Handbook of Psycholinguistic and Cognitive processes: Perspectives in Communication Sisorders*, 49–73.

Alammar, J. (2018). The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning). Retrieved from http://jalammar.github.io/illustrated-bert/.

Albert, R., Jeong, H., & Barabási, A. L. (2000). Error and attack tolerance of complex networks. *Nature*, *406*(6794), 378.

Anderson, J. R. (2000). *Learning and Memory: An Integrated Approach*. John Wiley & Sons Inc.

Andrews, M., & Vigliocco, G. (2010). The hidden Markov topic model: A probabilistic model of semantic representation. *Topics in Cognitive Science*, *2*(1), 101–113.

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological Review*, *116*(3), 463.

Ashby, F. G., & Maddox, W. T. (1993). Relations between prototype, exemplar, and decision bound models of categorization. *Journal of Mathematical Psychology*, *37*(3), 372–400.

Asr, F. T., Willits, J., & Jones, M. (2016). Comparing Predictive and Co-occurrence Based Models of Lexical Semantics Trained on Child-directed Speech. *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Avery, J., Jones, M.N. (2018). Comparing models of semantic fluency: Do humans forage optimally, or walk randomly? In *Proceedings of the 40th Annual Meeting of the Cognitive Science Society*. 118–123.

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv: 1409.0473*.

Bahdanau, D., Hill, F., Leike, J., Hughes, E., Hosseini, A., Kohli, P., & Grefenstette, E. (2018). Learning to understand goal specifications by modelling reward. *arXiv preprint arXiv:1806.01946*.

Balota, D. A., & Coane, J. H. (2008). Semantic memory. In Byrne JH, Eichenbaum H, Mwenzel R, Roediger III HL, Sweatt D (Eds.). *Learning and Memory: A Comprehensive Reference* (pp. 511–34). Amsterdam: Elsevier.

Balota, D. A., & Lorch, R. F. (1986). Depth of automatic spreading activation: Mediated priming effects in pronunciation but not in lexical decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(3), 336.

Balota, D. A., & Paul, S. T. (1996). Summation of activation: Evidence from multiple primes that converge and diverge within semantic memory. *Journal of Experimental Psychology: Learning, Memory & Cognition*, *22*, 827–845.

Balota, D. A., & Yap, M. J. (2006). Attentional control and flexible lexical processing: Explorations of the magic moment of word recognition. In S. Andrews (Ed.), *From inkmarks to ideas: Current issues in Lexical processing*, 229–258.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., … Treiman, R. (2007). The English lexicon project. *Behavior Research Methods*, *39*(3), 445–459.

Bannard, C., Baldwin, T., & Lascarides, A. (2003). A statistical approach to the semantics of verb-particles. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment* (pp. 65–72).

Barabási, A. L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*(5439), 509–512.

Baroni, M., Dinu, G., & Kruszewski, G. (2014). Don't count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (Vol. 1, pp. 238–247).

Baroni, M., & Lenci, A. (2008). Concepts and properties in word spaces. *Italian Journal of Linguistics*, *20*(1), 55–88.

Barros-Loscertales, A., González, J., Pulvermüller, F., Ventura-Campos, N., Bustamante, J. C., Costumero, V., … Ávila, C. (2011). Reading salt activates gustatory brain regions: fMRI evidence for semantic grounding in a novel sensory modality. *Cerebral Cortex*, *22*(11), 2554–2563.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain sciences*, *22*(4), 577–660.

Barsalou, L. W. (2003). Abstraction in perceptual symbol systems. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 358(1435), 1177–1187.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology, 59*, 617–645.

Barsalou, L. W. (2016). On staying grounded and avoiding quixotic dead ends. *Psychonomic Bulletin & Review, 23*(4), 1122–1142.

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. *Symbols, Embodiment, and Meaning*, 245–283.

Barsalou, L. W., & Wiemer-Hastings, K. (2005). Situating abstract concepts. *Grounding Cognition: The role of Perception and Action in Memory, Language, and Thought*, 129–163.

Beaty, R. E., Kaufman, S. B., Benedek, M., Jung, R. E., Kenett, Y. N., Jauk, E., … Silvia, P. J. (2016). Personality and complex brain networks: The role of openness to experience in default network efficiency. *Human Brain Mapping*, *37*(2), 773–779.

Bender, D. (2015). Establishing a Human Baseline for the Winograd Schema Challenge. In *MAICS* (pp. 39–45). Retrieved from https://pdfs.semanticscholar.org/1346/3717354ab61348a0141ebd3b0fdf28e91af8.pdf.

Bengio, Y., Goodfellow, I. J., & Courville, A. (2015). Deep learning, book in preparation for MIT press (2015).

Binder, K. S. (2003). Sentential and discourse topic effects on lexical ambiguity processing: An eye movement examination. *Memory & Cognition*, *31*(5), 690–702.

Binder, K. S., & Rayner, K. (1998). Contextual strength does not modulate the subordinate bias effect: Evidence from eye fixations and self-paced reading. *Psychonomic Bulletin & Review*, *5*(2), 271–276.

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

Block, C. K., & Baldwin, C. L. (2010). Cloze probability and completion norms for 498 sentences: Behavioral and neural validation using event-related potentials. *Behavior Research Methods*, *42*(3), 665–670.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Botha, J., & Blunsom, P. (2014). Compositional morphology for word representations and language modelling. In *International Conference on Machine Learning* (pp. 1899–1907). Retrieved from http://proceedings.mlr.press/v32/botha14.pdf.

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, *11*(2), 177–220.

Bransford, J. D., & Johnson, M. K. (1972). Contextual prerequisites for understanding: Some investigations of comprehension and recall. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 717–726.

Britton, B. K. (1978). Lexical ambiguity of words used in English text. *Behavior Research Methods & Instrumentation, 10*, 1–7. https://doi.org/10.3758/BF03205079.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Agarwal, S. (2020). Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*. Retrieved from https://arxiv.org/pdf/2005.14165.pdf.

Bruni, E., Tran, N. K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47.

Buchanan, E. M., Valentine, K. D., & Maxwell, N. P. (2019). English semantic feature production norms: An extended database of 4436 concepts. *Behavior Research Methods*, 1–15

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*(3), 510–526.

Burgess, C. (2000). Theory and operational definitions in computational memory models: A response to Glenberg and Robertson. *Journal of Memory and Language*, *43*(3), 402–408.

Burgess, C. (2001). Representing and resolving semantic ambiguity: A contribution from high-dimensional memory modeling. *On the consequences of meaning selection: Perspectives on resolving lexical ambiguity*, 233–260.

Carey, S., & Bartlett, E. (1978). Acquiring a single new word. *Papers and Reports on Child Language Development*, *15*, 17–29

Chen, D., Peterson, J. C., & Griffiths, T. L. (2017). Evaluating vector-space models of analogy. In *Proceedings of the 39th Annual Conference of the Cognitive Science Society*. Retrieved from https://arxiv.org/abs/1705.04416.

Chen, X., Cardie, C. (2018). Unsupervised Multilingual Word Embeddings. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing *(EMNLP 2018)*. Retrieved from https://arxiv.org/abs/1808.08933.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*. Retrieved from https://arxiv.org/abs/1406.1078.

Chwilla, D. J., & Kolk, H. H. (2002). Three-step priming in lexical decision. *Memory & Cognition*, *30*(2), 217–225.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What Does BERT Look At? An Analysis of BERT's Attention. *arXiv preprint arXiv:1906.04341*. Retrieved from https://arxiv.org/abs/1906.04341.

Clark, S., Coecke, B., & Sadrzadeh, M. (2008). A compositional distributional model of meaning. In *Proceedings of the Second Quantum Interaction Symposium* (QI-2008) (pp. 133–140).

Clark, S., & Pulman, S. (2007). Combining symbolic and distributional models of meaning. Retrieved from https://www.aaai.org/Papers/Symposia/Spring/2007/SS-07-08/SS07-08-008.pdf.

Clay, M. M. (1968). A syntactic analysis of reading errors. *Journal of Verbal Learning and Verbal Behavior*, *7*(2), 434–438.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F., & Wattenberg, M. (2019). Visualizing and measuring the geometry of bert. arXiv preprint arXiv:1906.02715. Retrieved from https://arxiv.org/pdf/1906.02715.pdf.

Collell, G., & Moens, M. F. (2016). Is an image worth more than a thousand words? on the fine-grain semantic differences between visual and linguistic representations. In *Proceedings of the 26th International Conference on Computational Linguistics* (pp. 2807–2817). ACL.

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

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

Collobert, R., & Weston, J. (2008). A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*,160–167. ACM.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., & Kuksa, P. (2011). Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, *12*, 2493–2537.

Corballis, M. C. (2017a). Language evolution: a changing perspective. *Trends in Cognitive Sciences*, *21*(4), 229–236.

Corballis, M. C. (2017b). The evolution of language. In J. Call, G. M. Burghardt, I. M. Pepperberg, C. T. Snowdon, & T. Zentall (Eds.), *APA handbooks in psychology®. APA handbook of comparative psychology: Basic concepts, methods, neural substrate, and behavior* (p. 273–297). American Psychological Association. https://doi.org/10.1037/0000011-014

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems-level proposal for the neural substrates of recall and recognition. *Cognition*, *33*(1–2), 25–62.

De Deyne, S., Navarro, D. J., Perfors, A., Brysbaert, M., & Storms, G. (2019). The "Small World of Words" English word association norms for over 12,000 cue words. *Behavior Research Methods*, *51*(3), 987–1006.

De Deyne, S., Perfors, A., & Navarro, D. J. (2016). Predicting human similarity judgments with distributional models: The value of word associations. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (pp. 1861–1870).

De Deyne, S., & Storms, G. (2008). Word associations: Network and semantic properties. *Behavior Research Methods*, *40*(1), 213–231.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology*, *58*(1), 17.

Dell, G. S. (1990). Effects of frequency and vocabulary type on phonological speech errors. *Language and Cognitive Processes*, *5*(4), 313–349.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In 2009 IEEE conference on computer vision and pattern recognition (pp. 248–255). IEEE.

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*(2), 145–193.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

DeVore, I., & Tooby, J. (1987). The reconstruction of hominid behavioral evolution through strategic modeling. *The Evolution of Human Behavior: Primate Models, edited by WG Kinzey*, 183–237.

Dove, G. (2011). On the need for embodied and dis-embodied cognition. *Frontiers in Psychology*, *1*, 242.

Dubossarsky, H., De Deyne, S., & Hills, T. T. (2017). Quantifying the structure of free association networks across the life span. *Developmental Psychology*, *53*(8), 1560.

Duffy, S. A., Morris, R. K., & Rayner, K. (1988). Lexical ambiguity and fixation times in reading. *Journal of Memory and Language*, *27*(4), 429–446.

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(4), 1210–1223.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science, 14*(2), 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning, 7*(2-3), 195–225.

Elman, J. L., & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126*(2), 252.

Farah, M. J., & McClelland, J. L. (2013). A computational model of semantic memory impairment: Modality specificity and emergent category specificity (Journal of Experimental Psychology: General, 120 (4), 339–357). In *Exploring Cognition: Damaged Brains and Neural Networks* (pp. 79–110). Psychology Press.

Federmeier, K. D., & Kutas, M. (1999). A rose by any other name: Long-term memory structure and sentence processing. *Journal of Memory and Language*, *41*(4), 469–495.

Fellbaum, C. (Ed.). (1998). *WordNet, an electronic lexical database*. Cambridge, MA: MIT Press.

Feng, Y., & Lapata, M. (2010). Visual information in semantic representation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 91–99). Association for Computational Linguistics.

Fernandino, L., Conant, L. L., Binder, J. R., Blindauer, K., Hiner, B., Spangler, K., & Desai, R. H. (2013). Where is the action? Action sentence processing in Parkinson's disease. *Neuropsychologia*, *51*(8), 1510–1517.

Finkelstein, L., Gabrilovich, E., Matias, Y., Rivlin, E., Solan, Z., Wolfman, G., & Ruppin, E. (2002). Placing search in context: The

concept revisited. *ACM Transactions on information systems*, *20*(1), 116–131.

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

Fischler, I. (1977). Semantic facilitation without association in a lexical decision task. *Memory & Cognition*, *5*, 335–339.

Franklin, N., Norman, K. A., Ranganath, C., Zacks, J. M., & Gershman, S. J. (2019). Structured event memory: a neuro-symbolic model of event cognition. *BioRxiv*, 541607. Retrieved from https://www.biorxiv.org/content/biorxiv/early/2019/02/05/541607.full.pdf.

Fried, E. I., van Borkulo, C. D., Cramer, A. O. J., Boschloo, L., Schoevers, R. A., & Borsboom, D (2017). Mental disorders as networks of problems: a review of recent insights. *Social Psychiatry and Psychiatric Epidemiology*, *52*(1), 1–10.

Gabrieli, J. D., Cohen, N. J., & Corkin, S. (1988). The impaired learning of semantic knowledge following bilateral medial temporal-lobe resection. *Brain and Cognition*, *7*(2), 157–177.

Garagnani, M., & Pulvermüller, F. (2016). Conceptual grounding of language in action and perception: a neurocomputational model of the emergence of category specificity and semantic hubs. *European Journal of Neuroscience*, *43*(6), 721–737.

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

Goldberg, R. F., Perfetti, C. A., & Schneider, W. (2006). Perceptual knowledge retrieval activates sensory brain regions. *Journal of Neuroscience*, *26*(18), 4917–4921.

Griffiths, T. L., & Steyvers, M. (2002). A probabilistic approach to semantic representation. In *Proceedings of the Annual meeting of the Cognitive Science Society*, *24*(24).

Griffiths, T. L., & Steyvers, M. (2003). Prediction and semantic association. In *Advances in Neural Information Processing Systems*, 11–18.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(1), 5228–5235.

Griffiths, T. L., Steyvers, M., & Tenenbaum, J. B. (2007). Topics in semantic representation. *Psychological Review, 114*(2), 211.

Gruenenfelder, T. M., Recchia, G., Rubin, T., & Jones, M. N. (2016). Graph-theoretic properties of networks based on word association norms: implications for models of lexical semantic memory. *Cognitive Science*, *40*(6), 1460–1495.

Guida, A., & Lenci, A. (2007). Semantic properties of word associations to Italian verbs. *Italian Journal of Linguistics*, *19*(2), 293–326.

Günther, F., Rinaldi, L., & Marelli, M. (2019). Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science*, *14*(6), 1006–1033.

Hard, B. M., Tversky, B., & Lang, D. S. (2006). Making sense of abstract events: Building event schemas. *Memory & Cognition*, *34*(6), 1221–1235.

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1–3), 335–346.

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

Hebb, D. (1949). The organization of learning. Cambridge, MA: MIT Press.

Heinzerling, B. (2019). NLP's Clever Hans Moment has Arrived. Retrieved from https://thegradient.pub/nlps-clever-hans-moment-has-arrived/.

Hermann, K. M., Hill, F., Green, S., Wang, F., Faulkner, R., Soyer, H., … Wainwright, M. (2017). Grounded language learning in a simulated 3d world. *arXiv preprint arXiv:1706.06551*. Retrieved from https://arxiv.org/abs/1706.06551.

Hills, T. T. (2006). Animal foraging and the evolution of goal-directed cognition. *Cognitive Science*, *30*(1), 3–41.

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431.

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

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review, 95*(4), 528.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Hodges, J. R., & Patterson, K. (2007). Semantic dementia: a unique clinicopathological syndrome. *The Lancet Neurology*, *6*(11), 1004–1014.

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological Review*, *125*(3), 293.

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. (2011). Constructing semantic representations from a gradually changing representation of temporal context. Topics in Cognitive Science, 3(1), 48–73.

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning. *Journal of Memory and Language*, 53(2), 258–276.

Hutchison, K. A. (2003). Is semantic priming due to association strength or feature overlap? A microanalytic review. *Psychonomic Bulletin & Review*, 10(4), 785–813.

Hutchison, K. A., Balota, D. A., Neely, J. H., Cortese, M. J., Cohen-Shikora, E. R., Tse, C. S., … Buchanan, E. (2013). The semantic priming project. *Behavior Research Methods*, *45*(4), 1099–1114.

James, L. E., & Burke, D. M. (2000). Phonological priming effects on word retrieval and tip-of-the-tongue experiences in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(6), 1378.

Jamieson, R. K., Avery, J. E., Johns, B. T., & Jones, M. N. (2018). An instance theory of semantic memory. *Computational Brain & Behavior*, 1(2), 119–136.

Jawahar, G., Sagot, B., Seddah, D. What does BERT learn about the structure of language?. ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Jul 2019, Florence, Italy. ffhal-02131630f. Retrieved from https://hal.inria.fr/hal-02131630/document.

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

Johns, B. T., & Jones, M. N. (2015). Generating structure from experience: A retrieval-based model of language processing. *Canadian Journal of Experimental Psychology/Revue canadienne de psychologie expérimentale*, *69*(3), 233.

Johns, B. T., Mewhort, D. J., & Jones, M. N. (2019). The Role of Negative Information in Distributional Semantic Learning. *Cognitive Science*, *43*(5), e12730.

Jones, M., & Recchia, G. (2010). You can't wear a coat rack: A binding framework to avoid illusory feature migrations in perceptually grounded semantic models. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 32, No. 32).

Jones, M. N. (2018). When does abstraction occur in semantic memory: insights from distributional models. *Language, Cognition and Neuroscience*, 1–9.

Jones, M. N., Gruenenfelder, T. M., & Recchia, G. (2018). In defense of spatial models of semantic representation. *New Ideas in Psychology*, *50*, 54–60.

Jones, M. N., Hills, T. T., & Todd, P. M. (2015). Hidden processes in structural representations: A reply to Abbott, Austerweil, and Griffiths (2015). *Psychological Review*, *122*(3), 570–574. doi: https://doi.org/10.1037/a0039248

Jones, M. N., Kintsch, W., & Mewhort, D. J. (2006). High-dimensional semantic space accounts of priming. *Journal of Memory and Language*, *55*(4), 534–552.

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

Jones, M. N., Willits, J., & Dennis, S. (2015). Models of semantic memory. *Oxford Handbook of Mathematical and Computational Psychology*, 232–254.

Kalchbrenner, N., Grefenstette, E., & Blunsom, P. (2014). A convolutional neural network for modelling sentences. *arXiv preprint arXiv:*1404.2188.

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

Kenett, Y. N., Anaki, D., & Faust, M. (2014). Investigating the structure of semantic networks in low and high creative persons. *Frontiers in Human Neuroscience, 8*, 407.

Kenett, Y. N., Gold, R., & Faust, M. (2016). The hyper-modular associative mind: a computational analysis of associative responses of persons with Asperger syndrome. *Language and Speech, 59*(3), 297–317.

Kenett, Y. N., Kenett, D. Y., Ben-Jacob, E., & Faust, M. (2011). Global and local features of semantic networks: Evidence from the Hebrew mental lexicon. *PloS one, 6*(8), e23912.

Kenett, Y. N., Levi, E., Anaki, D., & Faust, M. (2017). The semantic distance task: Quantifying semantic distance with semantic network path length. Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(9), 1470.

Kiela, D., & Bottou, L. (2014). Learning image embeddings using convolutional neural networks for improved multi-modal semantics. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 36–45).

Kiela, D., Bulat, L., & Clark, S. (2015). Grounding semantics in olfactory perception. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (pp. 231–236). Beijing, China: ACL.

Kiela, D., & Clark, S. (2015). Multi-and cross-modal semantics beyond vision: Grounding in auditory perception. In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing *(EMNLP 2015)* (pp. 2461–2470). Lisbon, Portugal: ACL.

Kim, A., Ruzmaykin, M., Truong, A., & Summerville, A. (2019). Cooperation and Codenames: Understanding Natural Language Processing via Codenames. In *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment* (Vol. 15, No. 1, pp. 160–166).

Kim, Y., Jernite, Y., Sontag, D., & Rush, A. M. (2016). Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*. Retrieved from https://www.aaai.org/ocs/index.php/AAAI/AAAI16/paper/viewPaper/12489.

Kintsch, W. (2001). Predication. *Cognitive Science, 25*(2), 173–202.

Kintsch, W., & Walter Kintsch, C. (1998). Comprehension: A paradigm for cognition. Cambridge university press.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14.

Kumar, A. A., Balota, D. A., Habbert, J., Scaltritti, M., & Maddox, G. B. (2019). Converging semantic and phonological information in lexical retrieval and selection in young and older adults. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. *45*(12), 2267–2289. https://doi.org/10.1037/xlm0000699

Kumar, A.A., Balota, D.A., Steyvers, M. (2019). Distant Concept Connectivity in Network-Based and Spatial Word Representations. In *Proceedings of the 41st Annual Meeting of the Cognitive Science Society*, 1348–1354.

Kumar, A.A., Steyvers, M., & Balota, D.A. (under review). Investigating Semantic Memory Retrieval in a Cooperative Word Game.

Kurach, K., Raichuk, A., Stańczyk, P., Zając, M., Bachem, O., Espeholt, L., … Gelly, S. (2019). Google research football: A novel reinforcement learning environment. *arXiv preprint arXiv:1907.11180*. Retrieved from https://arxiv.org/abs/1907.11180.

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*(2), 99–116.

Kwantes, P. J. (2005). Using context to build semantics. *Psychonomic Bulletin & Review*, *12*(4), 703–710.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain sciences*, *40*.

Lakoff, G., & Johnson, M. (1999). *Philosophy in the Flesh* (Vol. 4). New York: Basic books.

Lample, G., Conneau, A., Ranzato, M. A., Denoyer, L., & Jégou, H. (2018). Word translation without parallel data. In *International Conference on Learning Representations*. Retrieved from https://openreview.net/forum?id=H196sainb.

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

Landauer, T. K. (2001). Single representations of multiple meanings in latent semantic analysis.

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

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 *Proceedings of the 19th annual meeting of the Cognitive Science Society* (pp. 412–417).

Lazaridou, A., Marelli, M., & Baroni, M. (2017). Multimodal word meaning induction from minimal exposure to natural text. *Cognitive Science*, *41*, 677–705.

Lazaridou, A., Pham, N. T., & Baroni, M. (2015). Combining language and vision with a multimodal skip-gram model. *arXiv preprint arXiv:1501.02598*.

Lebois, L. A., Wilson-Mendenhall, C. D., & Barsalou, L. W. (2015). Are automatic conceptual cores the gold standard of semantic processing? The context-dependence of spatial meaning in grounded congruency effects. *Cognitive Science*, *39*(8), 1764–1801.

Levesque, H., Davis, E., & Morgenstern, L. (2012). The winograd schema challenge. In *Thirteenth International Conference on the Principles of Knowledge Representation and Reasoning*. Retrieved from https://www.aaai.org/ocs/index.php/KR/KR12/paper/viewPaper/4492.

Levy, O., & Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems* (pp. 2177–2185).

Levy, O., Goldberg, Y., & Dagan, I. (2015). Improving distributional similarity with lessons learned from word embeddings. *Transactions of the Association for Computational Linguistics*, *3*, 211–225.

Li, J., & Jurafsky, D. (2015). Do multi-sense embeddings improve natural language understanding?. *arXiv preprint arXiv:1506.01070*.

Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., & Jurafsky, D. (2016). Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*. Retrieved from https://arxiv.org/abs/1606.01541.

Li, J., Monroe, W., Shi, T., Jean, S., Ritter, A., & Jurafsky, D. (2017). Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*. Retrieved from https://arxiv.org/abs/1701.06547.

Li, P., Burgess, C., & Lund, K. (2000). The acquisition of word meaning through global lexical co-occurrences. In *Proceedings of the Thirtieth Annual Child Language Research Forum*, 166–178.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., … Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*. Retrieved from https://arxiv.org/abs/1907.11692.

Livesay, K., & Burgess, C. (1998). Mediated priming in high-dimensional semantic space: No effect of direct semantic relationships or co-occurrence. *Brain and Cognition, 37*(1), 102–105.

Lopopolo, A., & Miltenburg, E. (2015). Sound-based distributional models. In *Proceedings of the 11th International Conference on Computational Semantics* (pp. 70–75).

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302.

Lucas, M. (2000). Semantic priming without association: A meta-analytic review. *Psychonomic Bulletin & Review*, 7(4), 618–630.

Lucy, L., & Gauthier, J. (2017). Are distributional representations ready for the real world? Evaluating word vectors for grounded perceptual meaning. *arXiv preprint arXiv:1705.11168*.

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

Luong, T., Socher, R., & Manning, C. (2013). Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning* (pp. 104–113). Retrieved from https://www.aclweb.org/anthology/W13-3512/.

Lupker, S. J. (1984). Semantic priming without association: A second look. *Journal of Verbal Learning and Verbal Behavior*, *23*, 709–733.

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language, 92*, 57–78.

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

Matheson, H., White, N., & McMullen, P. (2015). Accessing embodied object representations from vision: A review. *Psychological Bulletin*, *141*(3), 511.

Matheson, H. E., & Barsalou, L. W. (2018). Embodiment and grounding in cognitive neuroscience. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, *3*, 1–27.

Mayford, M., Siegelbaum, S. A., & Kandel, E. R. (2012). Synapses and memory storage. *Cold Spring Harbor Perspectives in Biology*, *4*(6), a005751.

McCann, B., Keskar, N. S., Xiong, C., & Socher, R. (2018). The natural language decathlon: Multitask learning as question answering. *arXiv preprint arXiv:1806.08730*. Retrieved from https://arxiv.org/abs/1806.08730.

McClelland, J. L., Hill, F., Rudolph, M., Baldridge, J., & Schütze, H. (2019). Extending Machine Language Models toward Human-Level Language Understanding. *arXiv preprint arXiv:1912.05877*. Retrieved from https://arxiv.org/abs/1912.05877.

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

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*(6), 1155.

McKoon, G., Ratcliff, R., & Dell, G. S. (1986). A critical evaluation of the semantic-episodic distinction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 12*(2), 295–306. https://doi.org/10.1037/0278-7393.12.2.295

McNamara, T. P. (2005). Semantic priming: Perspectives from memory and word recognition. Psychology Press. (p. 86)

McNamara, T. P., & Altarriba, J. (1988). Depth of spreading activation revisited: Semantic mediated priming occurs in lexical decisions. *Journal of Memory and Language*, *27*(5), 545–559.

McRae, K. (2004). Semantic memory: Some insights from feature-based connectionist attractor networks. *The Psychology of Learning and Motivation: Advances in Research and Theory*, *45*, 41–86.

McRae, K., Cree, G. S., Seidenberg, M. S., & McNorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559.

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

McRae, K., Khalkhali, S., & Hare, M. (2012). Semantic and associative relations: Examining a tenuous dichotomy. In V. F. Reyna, S. B. Chapman, M. R. Dougherty, & J. Confrey (Eds.), *The Adolescent Brain: Learning, Reasoning, and Decision Making* (pp. 39–66). Washington, DC: APA.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545–567.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227.

Michel, P., Levy, O., & Neubig, G. (2019). Are sixteen heads really better than one?. In Advances in Neural Information Processing Systems (pp. 14014–14024).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems* (pp. 3111–3119).

Miller, G.A. (1995).WordNet: An online lexical database [Special Issue]. *International Journal of Lexicography*, *3*(4).

Mitchell, J., & Lapata, M. (2010). Composition in distributional models of semantics. *Cognitive Science, 34*(8), 1388–1429.

Morgenstern, L., Davis, E., & Ortiz, C. L. (2016). Planning, executing, and evaluating the winograd schema challenge. *AI Magazine*, *37*(1), 50–54.

Morris, R. K. (1994). Lexical and message-level sentence context effects on fixation times in reading. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *20*(1), 92.

Neely, J. H. (1977). Semantic priming and retrieval from lexical memory: Roles of inhibitionless spreading activation and limited-capacity attention. *Journal of Experimental Psychology: General*, *106*(3), 226.

Neely, J. H. (2012). Semantic priming effects in visual word recognition: A selective review of current findings and theories. In *Basic processes in reading* (pp. 272–344). Routledge.

Neisser, U. 1976. *Cognition and Reality*. San Francisco: W.H. Freeman and Co.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 402–407.

Nematzadeh, A., Miscevic, F., & Stevenson, S. (2016). Simple search algorithms on semantic networks learned from language use. *arXiv preprint arXiv:1602.03265*. Retrieved from https://arxiv.org/pdf/1602.03265.pdf.

Niven, T., & Kao, H. Y. (2019). Probing neural network comprehension of natural language arguments. *arXiv preprint arXiv:1907.07355*. Retrieved from https://arxiv.org/pdf/1907.07355.pdf.

Nosofsky, R. M. (1988). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *14*(4), 700.

Nozari, N., Trueswell, J. C., & Thompson-Schill, S. L. (2016). The interplay of local attraction, context and domain-general cognitive control in activation and suppression of semantic distractors during sentence comprehension. *Psychonomic Bulletin & Review*, *23*(6), 1942–1953.

O'Kane, G., Kensinger, E. A., & Corkin, S. (2004). Evidence for semantic learning in profound amnesia: an investigation with patient HM. *Hippocampus*, *14*(4), 417–425.

Olah, C. (2019). Understanding LSTM Networks. *Colah's Blog*. Retrieved from https://colah.github.io/posts/2015-08-Understanding-LSTMs/

Olney, A. M. (2011). Large-scale latent semantic analysis. *Behavior Research Methods*, *43*(2), 414–423.

OpenAI (2019). Dota 2 with Large Scale Deep Reinforcement Learning. Retrieved from https://arxiv.org/abs/1912.06680.

Osgood, C. E. (1952). The nature and measurement of meaning. *Psychological Bulletin*, *49*(3), 197.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. (1957). *The Measurement of Meaning* (No. 47). University of Illinois Press.

Pacht, J. M., & Rayner, K. (1993). The processing of homophonic homographs during reading: Evidence from eye movement studies. *Journal of Psycholinguistic Research*, *22*(2), 251–271.

Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *45*(3), 255.

Patterson, K., Nestor, P. J., & Rogers, T. T. (2007). Where do you know what you know? The representation of semantic knowledge in the human brain. *Nature Reviews Neuroscience*, *8*(12), 976.

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (EMNLP) (pp. 1532-1543).

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

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Pezzulo, G., & Calvi, G. (2011). Computational explorations of perceptual symbol systems theory. *New Ideas in Psychology*, *29*(3), 275-297.

Pinker, S. (2003). Language as an adaptation to the cognitive niche. *Studies in the Evolution of Language*, *3*, 16-37.

Pirrone, A., Marshall, J. A., & Stafford, T. (2017). A Drift Diffusion Model account of the semantic congruity effect in a classification paradigm. *Journal of Numerical Cognition*, *3*(1), 77-96.

Plaut, D. C., & Booth, J. R. (2000). Individual and developmental differences in semantic priming: empirical and computational support for a single-mechanism account of lexical processing. *Psychological Review*, *107*(4), 786.

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

Polyn, S. M., Norman, K. A., & Kahana, M. J. (2009). A context maintenance and retrieval model of organizational processes in free recall. *Psychological Review, 116*(1), 129.

Posner, M. I., & Keele, S. W. (1968). On the genesis of abstract ideas. *Journal of Experimental Psychology*, 77(3p1), 353.

Posner, M. I., & Snyder, C. R. R. (1975) Attention and cognitive control. In: Solso R (ed.) *Information Processing and Cognition: The Loyola Symposium*, pp. 55–85. Hillsdale, NJ: Erlbaum.

Pulvermüller, F. (2005). Brain mechanisms linking language and action. *Nature Reviews Neuroscience*, *6*(7), 576.

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

Quillian, M. R. (1969). The teachable language comprehender: A simulation program and theory of language. *Communications of the ACM, 12*(8), 459–476.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI Blog*, *1*(8). Retrieved from https://www.techbooky.com/wp-content/uploads/2019/02/Better-Language-Models-and-Their-Implications.pdf.

Rajpurkar, P., Jia, R., & Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. arXiv preprint arXiv: 1806.03822. Retrieved from https://arxiv.org/pdf/1806.03822.pdf.

Ratcliff, R., & McKoon, G. (2008). The diffusion decision model: theory and data for two-choice decision tasks. *Neural Computation*, *20*(4), 873–922.

Rayner, K., Cook, A. E., Juhasz, B. J., & Frazier, L. (2006). Immediate disambiguation of lexically ambiguous words during reading: Evidence from eye movements. *British Journal of Psychology*, *97*(4), 467–482.

Rayner, K., & Frazier, L. (1989). Selection mechanisms in reading lexically ambiguous words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(5), 779.

Recchia, G., & Jones, M. N. (2009). More data trumps smarter algorithms: Comparing pointwise mutual information with latent semantic analysis. *Behavior Research Methods*, *41*(3), 647–656.

Recchia, G., & Nulty, P. (2017). Improving a Fundamental Measure of Lexical Association. In *CogSci*.

Reisinger, J., & Mooney, R. J. (2010). Multi-prototype vector-space models of word meaning. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 109–117). Association for Computational Linguistics.

Rescorla, R. A. (1988). Behavioral studies of Pavlovian conditioning. *Annual Review of Neuroscience*, *11*(1), 329–352.

Rescorla, R. A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Classical Conditioning II: Current Research and Theory*, *2*, 64–99.

Reynolds, J. R., Zacks, J. M., & Braver, T. S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, *31*(4), 613–643.

Richie, R., White, B., Bhatia, S., & Hout, M. C. (2019). The spatial arrangement method of measuring similarity can capture high-dimensional, semantic structures. Retrieved from https://psyarxiv.com/qm27p.

Richie, R., Zou, W., & Bhatia, S. (2019). Predicting high-level human judgment across diverse behavioral domains. *Collabra: Psychology*, *5*(1).

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science, 3*(2), 303–345.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented in lists. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(4), 803.

Roget, P. M. (1911). Roget's Thesaurus of English Words and Phrases (1911 ed.). Retrieved October 28, 2004, from http://www.gutenberg.org/etext/10681

Rosch, E., & Lloyd, B. B. (Eds.). (1978). Cognition and categorization.

Rosch, E., & Mervis, C. B. (1975). Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, *7*(4), 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439.

Rotaru, A. S., Vigliocco, G., & Frank, S. L. (2018). Modeling the Structure and Dynamics of Semantic Processing. *Cognitive Science*, *42*(8), 2890–2917.

Rubinstein, D., Levi, E., Schwartz, R., & Rappoport, A. (2015). How well do distributional models capture different types of semantic knowledge?. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers,* 726–730.

Rumelhart, D. E. (1991). Understanding understanding. *Memories, thoughts and emotions: Essays in honor of George Mandler*, 257, 275.

Rumelhart, D. E., Hinton, G. E., & McClelland, J. L. (1986). A general framework for parallel distributed processing. *Parallel Distributed Processing: Explorations in the Microstructure of cognition*, *1*(45-76), 26.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1988). Learning representations by back-propagating errors. *Cognitive Modeling, 5*(3), 1.

Rumelhart, D. E., & Todd, P. M. (1993). Learning and connectionist representations. *Attention and Performance XIV: Synergies in Experimental Psychology, Artificial Intelligence, and Cognitive Neuroscience*, 3–30.

Sahlgren, M. (2008). The distributional hypothesis. *Italian Journal of Disability Studies*, *20*, 33–53.

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.

Saluja, A., Dyer, C., & Ruvini, J. D. (2018). Paraphrase-Supervised Models of Compositionality. *arXiv preprint arXiv:1801.10293*.

Schank, R. C., & Abelson, R. P. (1977). Scripts. *Plans, Goals and Understanding*.

Schapiro, A. C., Rogers, T. T., Cordova, N. I., Turk-Browne, N. B., & Botvinick, M. M. (2013). Neural representations of events arise from temporal community structure. *Nature Neuroscience*, *16*(4), 486.

Schneider, T. R., Debener, S., Oostenveld, R., & Engel, A. K. (2008). Enhanced EEG gamma-band activity reflects multisensory semantic matching in visual-to-auditory object priming. *Neuroimage*, *42*(3), 1244–1254.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417–424.

Shallice, T. (1988). Specialisation within the semantic system. *Cognitive Neuropsychology*, *5*(1), 133–142.

Shen, J. H., Hofer, M., Felbo, B., & Levy, R. (2018). Comparing Models of Associative Meaning: An Empirical Investigation of Reference in Simple Language Games. *arXiv preprint arXiv:1810.03717*.

Siew, C. S., Wulff, D. U., Beckage, N. M., & Kenett, Y. N. (2018). Cognitive Network Science: A review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*.

Silberer, C., & Lapata, M. (2012). Grounded models of semantic representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* (pp. 1423–1433). Association for Computational Linguistics.

Silberer, C., & Lapata, M. (2014, June). Learning grounded meaning representations with autoencoders. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 721–732).

Sivic, J., & Zisserman, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the Ninth IEEE International Conference on Computer Vision*.

Sloutsky, V. M., Yim, H., Yao, X., & Dennis, S. (2017). An associative account of the development of word learning. *Cognitive Psychology*, *97*, 1–30.

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(3), 214.

Socher, R., Huval, B., Manning, C. D., & Ng, A. Y. (2012). Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning* (pp. 1201–1211). Association for Computational Linguistics.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., & Potts, C. (2013). Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical methods in Natural Language Processing* (pp. 1631–1642).

Spranger, M., Pauw, S., Loetzsch, M., & Steels, L. (2012). Open-ended procedural semantics. In L. Steels & M. Hild (Eds.), *Language grounding in robots* (pp. 153–172). Berlin, Heidelberg, Germany: Springer.

Stanton, R. D., Nosofsky, R. M., & Zaki, S. R. (2002). Comparisons between exemplar similarity and mixed prototype models using a linearly separable category structure. *Memory & Cognition*, *30*(6), 934–944.

Stella, M., Beckage, N. M., & Brede, M. (2017). Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, *7*, 46730.

Stella, M., Beckage, N. M., Brede, M., & De Domenico, M. (2018). Multiplex model of mental lexicon reveals explosive learning in humans. *Scientific Reports*, *8*(1), 2259.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*(1), 41–78.

Sutton, R. and Barto, A. (1998). Reinforcement learning: An introduction. Cambridge, MA, MIT Press.

Swinney, D. A. (1979). Lexical access during sentence comprehension: (Re) consideration of context effects. *Journal of Verbal Learning and Verbal Behavior*, *18*(6), 645–659.

Tabossi, P., Colombo, L., & Job, R. (1987). Accessing lexical ambiguity: Effects of context and dominance. *Psychological Research*, *49*(2-3), 161–167.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. arXiv preprint arXiv:1905.05950. Retrieved from https://arxiv.org/pdf/1905.05950.pdf.

Thompson-Schill, S. L., Kurtz, K. J., & Gabrieli, J. D. E. (1998). Effects of semantic and associative relatedness on automatic priming. *Journal of Memory and Language*, *38*, 440–458.

Tulving, E. (1972). Episodic and semantic memory. *Organization of Memory*, *1*, 381–403.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, *37*, 141–188.

Tversky, A. (1977). Features of similarity. *Psychological Review*, *84*(4), 327.

Upadhyay, S., Chang, K. W., Taddy, M., Kalai, A., & Zou, J. (2017). Beyond bilingual: Multi-sense word embeddings using multilingual context. *arXiv preprint arXiv:1706.08160*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Vigliocco, G., Kousta, S. T., Della Rosa, P. A., Vinson, D. P., Tettamanti, M., Devlin, J. T., & Cappa, S. F. (2013). The neural representation of abstract words: the role of emotion. *Cerebral Cortex*, *24*(7), 1767–1777.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(2), 219–247.

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

Vitevitch, M. S., Chan, K. Y., & Goldstein, R. (2014). Insights into failed lexical retrieval from network science. *Cognitive Psychology*, *68*, 1–32.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. *arXiv preprint arXiv:1804.07461*. Retrieved from https://arxiv.org/abs/1804.07461.

Wang, S. I., Liang, P., & Manning, C. D. (2016). Learning language games through interaction. *arXiv preprint arXiv:1606.02447*.

Warstadt, A., Singh, A., & Bowman, S. R. (2018). Neural network acceptability judgments. *arXiv preprint arXiv:1805.12471*. Retrieved from https://arxiv.org/abs/1805.12471.

Watts, D. J., & Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature, 393*(6684), 440.

Westbury, C. (2016) Pay no attention to that man behind the curtain. *The Mental Lexicon, 11*(3), 350–374

Widdows, D. (2008). Semantic Vector Products: Some Initial Investigations. In *Proceedings of the Second AAAI Symposium on Quantum Interaction.* Retrieved from https://research.google/pubs/pub33477/.

Willems, R. M., Labruna, L., D'Esposito, M., Ivry, R., & Casasanto, D. (2011). A functional role for the motor system in language understanding: evidence from theta-burst transcranial magnetic stimulation. *Psychological Science*, *22*(7), 849–854.

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

Wulff, D. U., Hills, T., & Mata, R. (2018). Structural differences in the semantic networks of younger and older adults. Retrieved from https://psyarxiv.com/s73dp/.

Xu, Y., & Kemp, C. (2010). Inference and communication in the game of Password. In Advances in neural information processing systems (pp. 2514–2522).

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological review*, *114*(2), 245.

Yee, E., Chrysikou, E. G., Hoffman, E., & Thompson-Schill, S. L. (2013). Manual experience shapes object representations. *Psychological Science*, *24*(6), 909–919.

Yee, E., Huffstetler, S., & Thompson-Schill, S. L. (2011). Function follows form: Activation of shape and function features during object identification. *Journal of Experimental Psychology: General*, *140*(3), 348.

Yee, E., Jones, M. N., & McRae, K. (2018). Semantic memory. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, 3, 1–38.

Yee, E., Lahiri, A., & Kotzor, S. (2017). Fluid semantics: Semantic knowledge is experience-based and dynamic. *The Speech Processing Lexicon: Neurocognitive and Behavioural Approaches*, *22*, 236.

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, *23*(4), 1015–1027.

Yessenalina, A., & Cardie, C. (2011). Compositional matrix-space models for sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 172–182). Association for Computational Linguistics.

Zacks, J. M., Kurby, C. A., Eisenberg, M. L., & Haroutunian, N. (2011). Prediction error associated with the perceptual segmentation of naturalistic events. *Journal of Cognitive Neuroscience, 23* (12), 4057–4066.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence?. *arXiv preprint arXiv:1905.07830*. Retrieved from https://arxiv.org/pdf/1905.07830.pdf.

Zemla, J. C., & Austerweil, J. L. (2018). Estimating semantic networks of groups and individuals from fluency data. *Computational Brain & Behavior, 1*(1), 36–58.

Zhu, X., Sobhani, P., & Guo, H. (2016). Dag-structured long short-term memory for semantic compositionality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 917-926).