***A computational cognitive neuroscience perspective on word meaning in context*[1]**

J. L. McClelland, June 2018

In this section, we consider the nature of human word knowledge and the relationship between such knowledge and the concept of a word embedding. We formulate our conception of human word knowledge in terms of two challenges: inference and integration. By inference we mean the process of determining the contextually appropriate meaning of the word, and by integration, we mean the process of combining the results of this inference process with prior knowledge for subsequent use. We frame our proposed approaches to these challenges within an interactive, distributed processing framework for inference and a complementary learning systems framework for integration. We relate these ideas to the traditional concept of a word embeddings (Embeddings 1.0) and to our proposed updated conception of word embeddings (Embeddings 2.0). There are two key points:

1. Within the proposed framework, word representations are always *constructed* based on multiple sources of information.

2. The knowledge that supports this construction depends both on sustained neural activity and on synaptic weight changes. These synaptic changes include several distinct components: Among these are rapid, temporary changes that quickly fade away (*fast weights* in the parlance of deep learning); longer-lasting changes to synapses among neurons in the medial temporal lobes that largely serve as the substrate of initial storage of new learning; and gradual, prior-knowledge dependent changes to the strengths of connections within and among the various contributing neocortical areas participating in the representation and reconstruction process.

These points contrast with a focus on embeddings *per se* as the representation of word knowledge. It is true that the pattern of neural activity constructed by the inference process plays a role similar to that played by the embedding vector as usually construed. However, this pattern is not directly stored, as it is in a classical embedding framework. Furthermore, and more important, the knowledge underlying the ability to use a word appropriately in context – knowledge we traditionally ascribe to lexical entries within classical theories of language knowledge – is distributed widely throughout the entire neuro-computational system underlying language representation and use, and provides the context within which activity-based representation and synaptic weight changes are effective for supporting new learning.

[1] Draft of a section to appear in an article on the need to extend the concept of word embeddings to capture word meaning and its modulation by context, co-authored with Jason Baldridge, Felix Hill, Maja Rudolph, and Hinrich Schuetze (title and order of authors TBD).

Introduction

Here we present a perspective on human word meaning situated within the fundamental cognitive activity of understanding the world through experience and communication, implemented in a highly distributed and interactive processing system in the human brain. Our work builds on earlier modeling work treating language understanding as a process of mapping spoken or written input to a representation of the situation or event the language input describes (St. John & McClelland, 1990) and exploring the brain representations of learning and semantic memory (McClelland, McNaughton & O'Reilly, 1995; Rogers *et al.,* 2004). These efforts in turn grow out of the *Parallel-Distributed Processing* (PDP) framework for modeling human cognition (Rumelhart, McClelland, and the PDP Research Group, 1986; Rogers & McClelland, 2014). We describe the core tenets of the framework along with some of the relevant findings from the study of human language processing and related aspects of cognition. Incorporating these additional aspects of human language understanding, as characterized in this section, into the efforts of NLP researchers might one day help to enhance and extend the capabilities of existing language processing systems.

We frame our presentation within the context of two challenges facing a processing system (be it a human or a machine) when it encounters a novel word in context, such as the word 'wompamuck' in a sentence such as

> John saw a cute little wompamuck hiding behind a tree.

We call these challenges *inference* and *integration.* By *inference*, we mean the process whereby the system derives a representation of the newly encountered word 'wompamuck' from the single experience of reading this sentence. By *integration*, we mean the process whereby the results of this inference are integrated into the language processing system for subsequent use. We present our perspective on these processes as we understand them to occur in the human brain, treating the human solution as an ideal that existing computational methods do not yet fully achieve. Some elements of the characterization we describe are speculative, but many are grounded in a growing body of convergent computational and empirical investigations.

Inference

We view the inference problem as part of the broader language understanding problem, in which the primary goal in processing linguistic input is to construct an implicit, probabilistic representation of the situation or event being described by the language input (St. John & McClelland, 1990; Rabovsky, Hansen & McClelland, in press). Hereafter we call this a *situation* representation. A range of considerations from cognitive science and psycholinguistics (e.g., Altmann and Kamide, 2009) and more recently for cognitive neuroscience (e.g., Baggio & Hagoort, 2011) have led to a growing body of support for this view. While we use as our example the concrete event of John seeing the wompamuck, we see the approach as ultimately extendable to abstract situations, including the situation described in the example below, which led to the writing of the present article:

> Hinrich saw limitations in existing models of sentence embeddings and organized a workshop to explore extending the concept in new ways.

Figure 1 illustrates the approximate brain distribution of crucial systems involved in both the inference process and the integration process. As shown in the figure, spoken language reaches the auditory

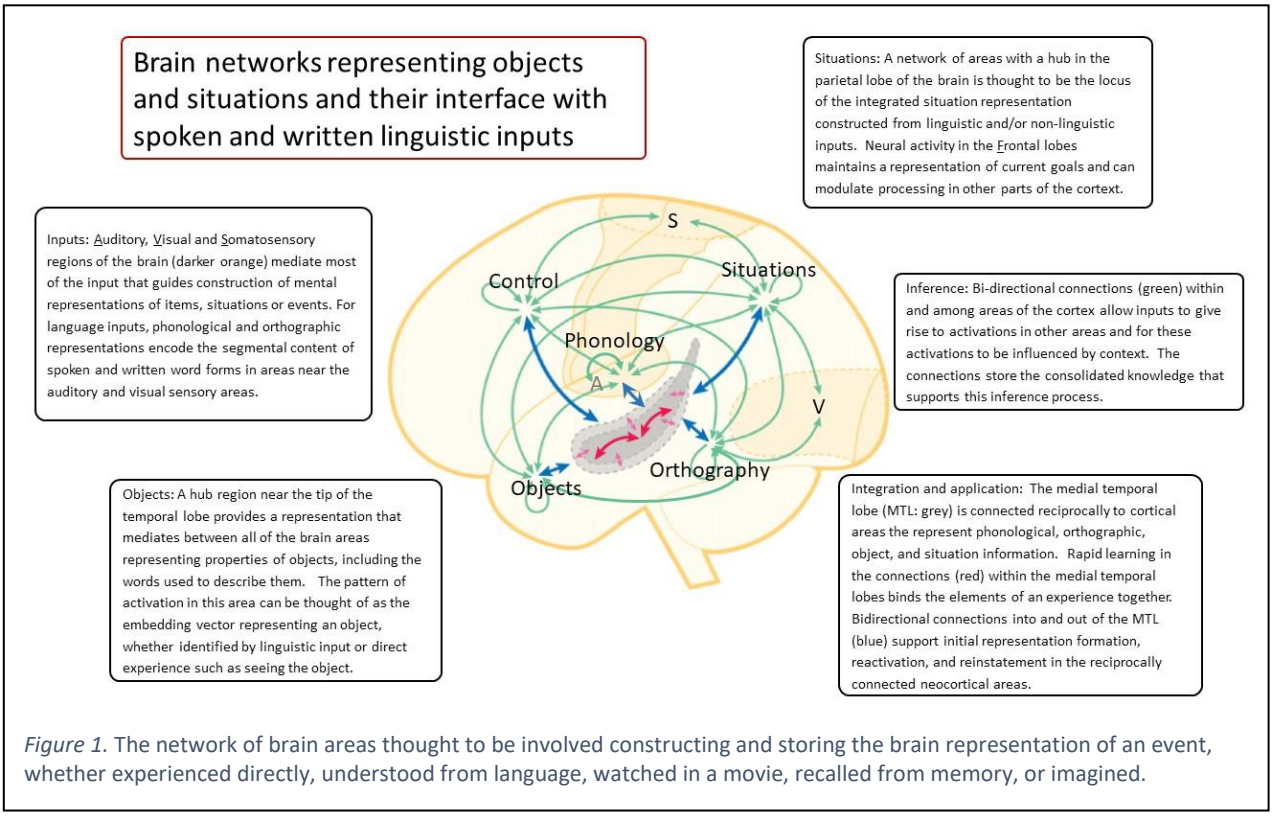

Figure 1. The network of brain areas thought to be involved constructing and storing the brain representation of an event, whether experienced directly, understood from language, watched in a movie, recalled from memory, or imagined.

cortex (labeled A in the Figure), and from there gives rise to patterns of activation characterizing the sound structure of spoken words. Written language reaches the Visual cortex (V in the Figure) and from there gives rise to patterns of activation specifying the orthographic structure of printed or written words (either the spoken or the written word representation can evoke the other via bi-directional connections between them). These patterns of activation also contribute to the formation of patterns representing the objects described in the linguistic input and of the overall situation or event that the linguistic input describes (Ranganath & Ritchey, 2012). The construction of all of these representations is thought to occur in a parallel, interactive computation, such that each informs the other (Rumelhart, 1977/1994). When the input is spoken language, it has been clear for over 50 years that context and meaning are essential to construct accurate representations of the spoken words in the sentence (Miller et al, 1951), and even to determine which sequences of sounds correspond to words. As an example of the latter, a phonological sequence that might be interpreted out of context as either ‘night rate’ or ‘nitrate’ will be interpreted unambiguously if preceded, e.g. by a context such as ‘the day rate was not as good as the …’ (Cole *et al.,* 1978). As an example of the former, if you hear 'The #eel of the shoe" (where # is a noise burst) you'll hear '#eel' as 'heel'. If 'shoe' is replaced by ‘wagon’ you'll hear 'wheel' (Warren, 1970). This second example illustrates that the disambiguating context need not occur prior to the ambiguous material; it appears the interpretation of ambiguous material remains open at least for a short period of time following the occurrence of the ambiguity. These processes affect the patterns of activation in areas associated with the auditory processing of words, consistent with our interactive framework (Shohoglu et al, 2012).

Since our primary focus is on representation of word meaning in context, we focus in the rest of this section on the object and situation representations. We see the object representation as corresponding

most closely to the word embeddings used by computational linguists, but we begin with a discussion of situation representations, to place our discussion of object representations in the appropriate context.

Situation representations

The situation representation specifies the event or situation conveyed by the language input including the participants and their inter-relationships. In the case of our example sentence, the objects would be a male human, likely English-speaking; a small furry animal; and a tree. The situation representation would integrate these objects and capture their relations as described in the sentence: for our sentence about the wompamuck, the sentence seems to describe a situation in which the human called John is noticing the animal, which is apparently attempting to avoid being seen by positioning itself on the opposite side of the tree from someone or something, possibly the same human John.

Central to our perspective is the proposition that the construction of a situation representation is something that can occur with or without language input (Zwaan & Radvansky, 1998) or through the convergent influence of both sources of information (Altman & Kanade, 2009). Direct experience, seeing a picture, or watching a silent movie can convey a situation without actual linguistic input, and a situation representation can also be reconstructed from memory or even imagined (Zadbood *et al,* 2017). A classic example of a situation understood without language (Heider and Simme, 1942), relies only on a highly schematic animation (Figure 2); when shown as a movie the sequence suggests animate beings with goals and desires, in which the smaller two are seen as fleeing from the larger one, and eventually succeed in escaping by hiding in an enclosure. It should also be clear that a situation representation not only reflects concrete objects and their relationships in space, but also includes animate, sentient beings with perceptions, beliefs, and intentions, taking concrete or abstract actions toward promoting outcomes related to their goals.

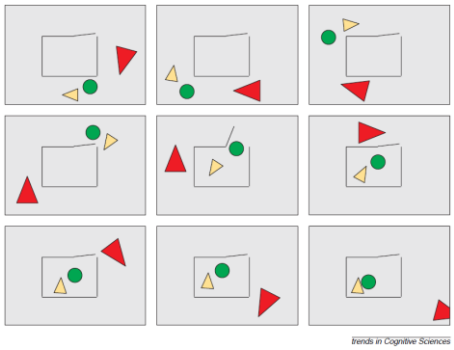

Figure 1. A series of snapshots from an animation from Heider and Simme (1942) that conveys an event in which two animate beings escape from another by hiding inside an enclosed space.

A growing body of evidence supports the proposition that the same process of constructing a representation of a situation and the participants in it occurs whether or not language is involved. A considerable body of cognitive neuroscience research supports the idea that the situation representation (associated with a set of interconnected brain areas centered on the parietal cortex as indicated in Figure 1) exhibit corresponding brain activity during the processing of a temporally extended event sequence, whether this is produced from watching a movie, from hearing a narrative description of the events conveyed by the movie, or from recalling the movie after having seen it. Lower level sensory cortical activations differ, but higher level areas in Parietal cortex show corresponding patterns in all three cases (Zadbood *et al,* 2017; Baldessano *et al.,* 2017). In summary, the process of constructing a mental representation of situations and events is one that can, but need not, be guided by language; language is just one of the sources of information that may contribute to the construction of such a representation.

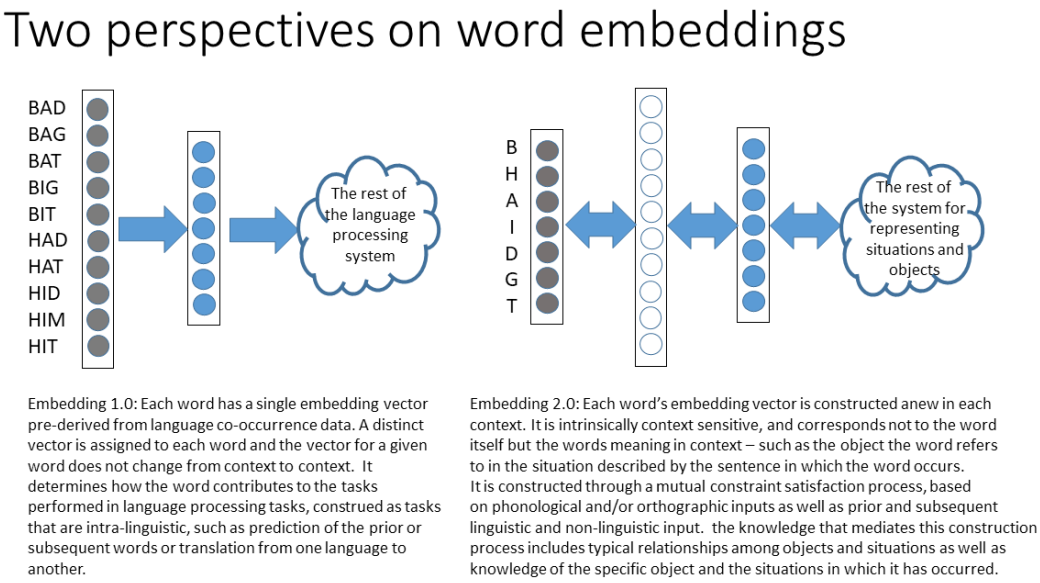

Figure 3. Two perspectives on word embeddings, emphasizing language processing as a part of a system for representing situations and objects.

The latent nature of situation representations. Within the PDP framework, the representations we are describing are thought to be patterns of activation within a neural network, rather than explicit symbolic structures, and thus the PDP framework is highly compatible with contemporary deep learning approaches to language processing. The sense in which these patterns 'specify' the relevant information is potentially latent or indirect. For example, in the PDP model of sentence comprehension of St. John & McClelland (1990), there is a latent, distributed representation that was originally called the *sentence gestalt,* though we would now call it the *situation gestalt* since it is thought to be a representation of the event or situation described be the sentence, not a representation of the sentence itself. The representation is latent in that it does not specify aspects of the situation or event directly. Instead, it provides the input to a query network that can answer questions about aspects of the event or situation when probed. Recently the sentence gestalt model has been extended to capture a large body of findings related to a brain potential called the N400, thought to reflect the update in the latent representation that occurs as each word of an input sentence is processed (Rabovsky *et al,* in press).

Object representation and word meaning

We now turn attention to object representations – patterns of activity that we see as corresponding most closely to the word embeddings used in NLP. However, there are many differences between the patterns as they are envisioned within the view we are presenting here and the notion of a word embedding we have characterized as Embeddings 1.0. Our characterization is consistent with the vision for Embeddings 2.0 illustrated in Figure 3. The first difference we describe is closely tied to the proposition that language understanding involves constructing a representation of the situation or event described by language input and of the objects that participate in it. We then turn to other differences that reflect the more general principles of our parallel-distributed processing approach.

Embeddings reflect object semantics, not simply text statistics. First, we emphasize that the pattern corresponding to a noun like *tree* or *wompamuck* is inseparable from the brain representation of the object the word designates. Under the proposition that language understanding is the process of constructing representations of situations and of the entities participating in them, we can think of different words whose meanings appear to be related (such as couch and sofa) not as having similar meanings as words, but as providing cues suggesting objects with similar properties. Although such similarities may be captured at least in part in Embeddings 1.0, where the representations of *couch* and *sofa* are derived only from word co-occurrence statistics, we suggest that ultimately a more satisfactory model will be one in which the patterns for these two items are similar at least in large part because of the similarities in the properties of the objects referred to by the words. On our view, it is these underlying object similarities that explain why such words have similar patterns of co-occurrence with other words in sentences; because of their similar properties, they participate in similar ways in a similar range of events, and the sentences used to describe these events are otherwise similar.

One important source of evidence for the view that the brain's representations of the meanings of words is inseparable from its representations of the objects they refer to this lies in the neuropsychological condition known as *semantic dementia.* This condition, which arises from a neurodegenerative disease process that gives rise to a progressive loss of neurons in a region near the label *Objects* in Figure 1, results in a gradual, and ultimately profound, loss of information about both words and objects. Affected individuals lose the ability not only to associate an object with its name but also to associate an object with its function or to associate a black and white line-drawing of an object with its typical color (Bozeat *et al.*, 2000). The deficit affects associative knowledge as well: patients with this condition gradually lose the ability to pair objects with others they co-occur with, independently of whether or not the stimuli are presented as words or as pictures. For example, semantic dementia patients fail in the 'pyramids and palm trees' test, in which the participant must pair, for example, a pyramid with a palm tree rather than a pine tree; there is a high degree of correspondence in the extent of the deficit across patients when the objects are shown in pictures or designated by printed words, suggesting that neither words nor pictures are accessing representations of the objects, preventing a determination of which ones are related. In addition, the disorder affects knowledge of the spellings and sounds of words as well as knowledge of objects and their properties (Patterson et al, 2006).[2]

Inferring word meanings in context

Embeddings are constructed, not stored as such. According to ideas that provided part of the motivating context for the initial development of the Parallel Distributed Processing framework, our approach emphasizes the idea that words do not *have* meanings – instead, they provide clues or cues to meaning (Rumelhart, 1979[JM1]). That is, words serve as one source of the information that constrains aspects of meaning, along with other sources, including linguistic and non-linguistic context. This idea is reflected in the conception of Embeddings 2.0 presented in Figure 3, where the representation of a

[2] The regions in Figure 1 are illustrated in the left hemisphere. Language input and output tend to be heavily lateralized to the left, but the object and situation representations are balanced across the two hemispheres of the brain. Behavioral and modeling studies suggest that connections between the spoken and written language areas and what we are calling the object area are predominantly left-lateralized as well (Lambon Ralph *et al.*, 2001; Shapiro *et al*., 2013).

word's meaning is constructed anew each time a word occurs, based on the spoken or written input as well as other sources of information. Here we elaborate two specific points closely related to the view that embeddings are constructed in the moment rather than retrieved as fixed patterns. The first point is that *word representations are derived through the convergent influences of a range of sources of information rather than retrieved from a fixed list of embedding vectors*. These influences include (i) phonological and orthographic influences; (ii) general knowledge of typical event and situation scenarios; (iii) co-occurring non-linguistic information, and (iv) previously-provided information about an object and the other objects with which it co-occurs within a particular situation. We emphasize that all of these influences operate concurrently and interactively, influencing the construction of the representation of the object associated with a given word and in turn being influenced by this representation. The second point is that the *patterns corresponding to objects of the range of types designated by a particular word lie as points in a continuous space*, so that no two patterns associated with a particular word need ever be exactly the same (Elman, 1990).

Sources of influence on object representations

Phonological and morphological influences. Instead of looking up the embedding vector that corresponds to a word in a list, we propose that there is a learned mapping from input phonology or spelling to meaning, which would allow for similarity based generalization and quasi-compositionality in the mapping from spelling or sound to word meaning. There are also quasi-compositional learned mappings between spelling and sound representations as well (Seidenberg & McClelland, 1989) though these are less relevant to the current focus on meaning in context.

The connections from the spoken or written input leading to the layer in the network where the word's meaning is represented could be thought of as performing a role similar to the way we ordinarily think of embeddings, but with the proviso that the connections support similarity based generalization and quasi-compositional construction of novel word meanings as well as providing a source of input that constraints but does not uniquely specify the patterns corresponding to a familiar word's meaning.

Although sound-meaning and spelling-meaning mappings can be quite arbitrary, it is clear that spelling and sound can provide clues to meaning as well – the sequence of characters between spaces is not completely arbitrary, but rather reflects a range of effects ranging from idiosyncratic to almost fully systematic. At the arbitrary end, though *huge*, *gigantic*, and *enormous* appear to have no phonological or orthographic structure in common, '*humongous*' and '*ginormous*' can be understood whether or not they have been previously encountered as such, because of their orthographic and phonological similarities to these existing words. Jabberwocky and Joycean neologisms capitalize on this aspect of form-meaning relationships. Indeed, we suggest that the word 'wompamuck', though in part arbitrary, contributes to the suggestion of a small foraging mammal due to its orthographic and phonological similarity to chipmunk, and we note that novel drug names and other product names are often chosen specifically to be suggestive even when they are not existing words as such. Turning to the more systematic end of the continuum: While some form-meaning relationships are highly systematic, such as the relationship between the suffix *-ing* and the progressive aspect of verb meaning in English or between the suffix *-ed* and the English past tense, but many aspects of morphology are far less systematic (Marchand, 1969). Such quasi-regular relationships are difficult to capture with explicit systems of rules but can be exploited and used in neural network models to capture human patterns of word meaning processing (Plaut & Gonnerman, 2000).

General knowledge of typical event scenarios. A second source of constraint on the representation of the object indicated by a word arises from the linguistic context in which the word occurs. According to our object-and-situation based perspective, influences from other words and non-linguistic context jointly influence the evolving situation representation which in turn constrains the representation we assign to the object referenced by a word. Let us consider a variant of our example sentence in this context:

John saw a cute little animal hiding behind a tree.

Here we have used a familiar word – animal – instead of the novel word, wompamuck. In a different sentence (John saw a heard of large animals grazing on a hillside) we would get a very different impression of the animals involved. Experimental studies have verified that the meaning listeners take away from the use of a word depends on the context in which the word occurs (Barkley et al, 1974). After hearing 'The camper petted the animal' memory for the sentence could be cued by the phrase 'something friendly' but not 'something ferocious', while after hearing 'The camper escaped from the animal', 'something ferocious' but not 'something friendly' was the effective cue. These and related studies support the view that the representation we derive of the object referred to by a word depends on the overall situation described by the sentence. Here, a word like animal provides one source of clues, but our general knowledge of different types of animals and their behaviors in different situations constrains how we represent an object in context.

A similar process is involved in cases where the word can refer to two completely unrelated kinds of things, but where the situation described in a sentence provides a clear indication of which type is intended. Classic examples frequently used in the psychological literature include the words *bank, bat, bug,* and *ball.* According to our approach, a context such as 'The boy hit the ball with the bat' engenders the construction in the mind of the listener of an event in which the ball denotes a spherical object rather than a fancy dance and the bat denotes a longish swingable object rather than a flying mammal. Indeed the context is likely to be more specifically constraining for many listeners, who will tend to think of a baseball or softball and a baseball or softball bat.

A large body of cognitive science research beginning in the 1960's has investigated the processing of word meaning in context. Initially, it was argued that the presentation of a word initially activated all of a word's possible meanings independent of the context in which the word occurred, and that it was only a subsequent process that selected the contextually appropriate meaning from the alternative meanings of a word (Swinney, 1979). Subsequently, however, a considerable body of evidence build up supporting the view that when a word was encountered, both the relative frequency of alternative uses of a word (e.g. to refer to insects or secret recording devices) together with the constraints described in the unfolding sentence jointly determined how strongly each alternative meaning is activated in context (Simpson, 1984) as we should expect within our constraint-satisfaction approach (McClelland, 1987).

Co-occuring non-linguistic information. As we indicated above, it is now widely accepted in cognitive science and cognitive neuroscience that linguistic and non-linguistic input jointly constrain the situation representations we construct. Seminal research in the 1990's (Tanenhaus et al, 1995) supports the view that the situation representation we construct is jointly and immediately constrained by both visual and linguistic input, and visual input from a display is used to accelerate the process of identifying the referent of a spoken word. Tanenhaus et al demonstrated that, if the scene a participant is viewing contains only one object whose name begins with the syllable 'can' (such as a piece of candy), the

participant will start looking at this object almost immediately after the onset of the syllable; if there are two objects denoted by words beginning with the syllable 'can' (e.g., a candle as well as a piece of candy) eye movements to the target word are delayed until the spoken input uniquely identifies it. In another relevant study (Kamide, Altmann & Haywood, 2003), participants viewed scenes including a man, a young girl, a motorcycle, and a carousel, among other objects. Shortly after the onset of 'ride' in the unfolding sentence 'The child will ride' participants look toward the carousel and not the motorcycle, while the opposite is true when the sentence begins 'The man will ride'. Clearly then, both language and non-linguistic information influence the formation of a situation representation, and within that, the representations of the objects referred to in the sentence; here of the child as a girl, rather than a boy, since only a girl is shown in the picture; and of the object the child will ride, prior to any mention of the object.

Of course, we gain a great deal of information about previously unfamiliar objects by encountering them in real-world situations, in which language is only one part of the input we receive. A natural way of informing someone about a previously unfamiliar object, including the object's name, is to present a picture or a display containing the object, along with the name in spoken or written form. This approach is frequently employed in psycholinguistic research on word learning. In some cases, objects are presented in isolation along with their name, but in other cases, participants may be shown two or more objects, and asked to 'look at the dax' or 'the numbat'. If the other object in the display is a familiar object with a familiar name (for example, a cup or a dog) the participant will focus attention on the unfamiliar object (Markman & Wachtel, 1988). Thus, participants are able to infer the referent of a new word if there is only one plausible referent, even without other indicators such as the direction of pointing or gazing by a speaker. This and related inference processes are clearly relevant to word learning, and must be a part of an intelligent solution to the problem of assigning a meaning to a word.

Previously provided information about an object and the other objects it co-occurs with in a situation. Finally, we call attention to the fact that the meaning we assign to a word is highly sensitive to information presented earlier in a discourse. A striking example of this is provided by the study of Nieuwland and Van Berkum (2006). These authors took advantage of a brain potential called the N400, which occurs when a listener is processing a word that is incongruent with the preceding context. As one example, if a participant hears 'The peanut was', and then hears 'in love' a large N400 will occur, but no N400 occurs if instead the participant hears the word 'salted'. To demonstrate the effect of prior discourse information, the authors preceded the presentation of a sentence beginning with 'The peanut was' with the following little story:

> A woman saw a dancing peanut who had a big smile on his face. The peanut was singing about a girl he had just met. And judging from the song, the peanut was totally crazy about her. The woman thought it was really cute to see the peanut singing and dancing like that.

In this situation, if the story continues 'The peanut was *in love'* no N400 occurs; instead, the N400 now occurs if the story continues 'The peanut was *salted'.* This and several other studies by a number of researchers demonstrate that the meaning that we attach to a particular word (in this case *peanut*) depends, not only on the orthographic and phonological cues, general knowledge about objects and contexts and co-occurring non-linguistic information, but also on the properties of the object we have previously identified with a particular word within a given scenario or discourse context. The peanut

we are thinking about when the word occurs at the end of this little discourse is not just an ordinary peanut.

No two instances of the meaning of a word may ever be exactly the same. We now turn to the second key point we wish to emphasize about the constructive nature of the process of forming a representation of the object referred to by a spoken or written word: The resulting pattern of activation is not a fixed pattern corresponding to the word; instead, it corresponds to the instantiated meaning of the word in context. We already considered how the word animal designates a particular instance of an animal whose properties depend on the context. This is captured by the effect of the context on the specific instantiation of the pattern of activation for the animal described. In our conception here, we envision a vast space of such patterns, corresponding to different instances of animals in different contexts. With some words, of course, there will be large gaps in the space, with the patterns in each of two regions representing completely distinct types of objects. For example, the vector for the word *bat* in the sentence 'the boy hit the ball with the bat' would be completely different from the one for the word *bat* in 'The boy saw a small flying animal that turned out to be a bat'. For others, like 'animal', there may be a sense of a shared core across all of the variants, with many possible clusters of nearby points. In this conception, arbitrary homonyms (like the flying bat and the baseball bat), related but distinct meanings (such as the noun and verb meanings of run or the different types of containers use, say, to hold coffee or to hold apples), and subtle shadings of meaning (such as, perhaps, the different kinds of love we think if from the sentences 'John loves Mary, Mary loves John, the mother loves her baby, the pope loves sinners, everyone loves ice cream)' lie at different points along a continuum of similarity (McClelland, 1992). This view naturally captures the family resemblance among the many different meanings of a word that was famously noted by Wittgenstein (1953). Those who write dictionary entries can be seen as carving this continuous space into entries corresponding to points that lie at centroids of fairly densely populated regions of the space where the differences in meaning and context are characterizable.

Interim summary and comment. We have argued that words produce activations of patterns capturing the properties of the objects they correspond to within situations described in language. These patterns are constrained (i) by the orthographic and phonological properties of words; (ii) by their co-occurrence with other words in typical event scenarios; (iii) by accompanying non-linguistic input; and (iv) by previously presented information about the object the word refers to and other objects it co-occurs with in a particular situation. Our starting example sentence about the wompamuck provides the first two of these kinds of information; it seems straightforward to envision further constraining the representation of the object referred to by the word wompamuck by either or both of the other two sources of information.

Before concluding this section, we offer four additional comments.

First, the inference process we see as operating upon encountering a previously unseen word like wompamuck is not in any way distinct from the process that occurs upon encountering a word one has seen one or many times before. Because the meaning of existing words is so context dependent, we are always using all of the different sources of influence we have described in constraining the representation of the item to which a presented word refers, whether the word has previously been encountered or not. Thus, assigning a meaning to a new word in context is on a continuum with assigning a specific meaning to any word in any context.

Second, although we have focused on representations of concrete objects, we believe a similar approach can be extended to other words. For example, an action verb such as *cut* or *make* can be used to designate a wide range of different actions, and all of the same considerations we have described above would apply to the formation of a representation of the action corresponding to a verb, just as it applies to the formation of a representation of the object referred to by a noun. Verbs in conjunction with prepositions also specify relations among objects in a situation; as with nouns and concrete action verbs, we would suggest that all of the considerations we described above also still apply. It is, of course, also important to acknowledge that nouns are not only used to refer to concrete objects. Though Lakoff and Johnson (2008) and others have argued that abstract nouns are understood on analogy with concrete nouns (so that we use similar patterns of language to refer to relationships among parts of an argument and parts of a building), it is not clear that this is sufficient. It may be worth noting the neuropsychological phenomenon called 'phonological dyslexia', which typically occurs after extensive damage to the language areas and surrounding frontal and parietal areas in the left hemisphere. Such patients have lost the ability to read non-words aloud correctly, suggesting a loss of the ability to map directly from spelling to sound. These patients typically read concrete nouns more accurately than verbs or abstract nouns, and have extreme difficulty with function words. It is tempting to suppose that our representations of such words depend on representations other than concrete object representations. One proposal is that such representations are located in regions of the brain near the language areas in the frontal and parietal lobes of the brain (Warrington & Shallice, 1984). However, alternative accounts have been proposed. One approach holds that all words draw on the same representational system, but those that are less concrete are more prone to disruption by damage (Plaut and Shallice, 1993).[JM2] Another recent view proposes that abstract words tend to have more variable word meanings, and that cognitive control is required to manage the appropriate instantiation of word meanings in context (Hoffman, McClelland & Lambon-Ralph, in press).

Third, the relation between word and object is not as simple as the story we have been telling seems to suggest. Language often suggests the presence of objects not explicitly mentioned. A sentence in which someone is shot implicates both a gun and a bullet, even though neither is explicitly mentioned, and psycholinguistic research has demonstrated that listeners infer these objects, endorsing sentences that explicitly mention them as having occurred in passages that they have read even though these objects were not actually explicitly mentioned (Barclay *et al.,* 1974). Other words, such as adjectives, seem to function primarily to constrain representations of objects rather than to have independent meaning in their own right.

Finally, it is important to note that our knowledge of the object, action, or relationship designated by a word is not simply the pattern for it in the object area. Rather, our knowledge includes what we know about the situations it enters into. This applies to concrete objects as well as other things. People who are familiar with the game of baseball know many things about a baseball bat: it is used in the game for hitting the ball; can break if a ball is struck too close to the handle; is made of very hard wood; and comes in many different sizes suitable for use by players of different sizes and hitting styles. Thus, though our focus here is on embeddings, in our view these only serve as the entry point into a large body of articulate knowledge that we can express if probed for the objects role in different situations. This is an important part knowledge part of the knowledge that we have, and it cannot be left out of any full consideration of the knowledge a word draws upon when we encounter it in context.

Integration

Once an inference has been made about the properties of the object referred to by a new word, a person is then in a position to use this word when it occurs on a subsequent occasion. Here, we consider the nature of the process by which this outcome can occur. We begin by describing the striking evidence from the effects of damage to the medial temporal lobes on learning and memory. We then turn to describing a theory that builds on this evidence within the conception of the distributed network of brain systems described above. We will discuss the initial learning of the new information and how it can be effectively used soon after initial learning, leading up to a discussion of its gradual full integration into the system of representation and processing.

A key observation that lies at the heart of our account is the dramatic decrement in the ability to learn about new words and situations that occurs as a consequence of extensive damage to the Medial temporal lobes (MTL) in the brain (the region depicted in gray in Figure 1). The famous patient Henry Molaison (HM) provided the first, striking evidence on this point, as a consequence of an operation he underwent in 1953 (Scoville & Milner, 1957). To ameliorate his intractable epilepsy, HM's left and right medial temporal lobes were removed in a surgical procedure that profoundly affected his ability to learn and remember. After the operation, he had to be told why he was in the hospital and who the people around him were, over and over again. A person could speak with him for half an hour, then leave the room for but a minute or two, and HM would receive the person upon their return as if he had never met them before. He appeared to retain nothing about the experience, and, of course, had no memory for the person's name. Yet, HM did learn new things, albeit very slowly. He came to know that he was not good at remembering, and would greet people he did not recognize with the disclaimer that although they seemed unfamiliar, he was aware he might have met them before. In 1968, he identified John F. Kennedy as having been president and having been assassinated, even though these events happened long after his surgery. He came to understand new words that had come into use since his operation (e.g., 'bit' as a unit of information). However, he had a profound inability to learn new vocabulary in a laboratory learning setting, even with extensive repetition of words along with written definitions of their meanings.

HM's profound deficit in the ability to learn about new situations, people, objects, and their names was not accompanied by a corresponding loss of previous knowledge of words and their meanings, and he retained the ability to engage in meaningful conversation for the duration of a conversational episode. It was perfectly possible to explain to HM that he had had an operation and that he was in the hospital recovering from it, and he would ask cogent follow-up questions, and he could carry on a conversation about ordinary topics at considerable length. Brenda Milner, then a young memory researcher at McGill University, tested HM extensively. Her testing revealed he had both a verbal and a performance IQ in the normal range. He was able to recall facts about his own earlier life, and he was knowledgeable about major historical and political events, although he did not retain information he was only exposed to within a period of several months prior to his surgery, including memory of the many conversations and decisions leading up to his operation. This loss of memory extended back for more than a year, though the exact boundary was difficult to delineate since it was difficult to be sure in many cases exactly when he had encountered specific pieces of information. Also, as previously noted, although HM's conversations could sometimes return to topics previously discussed without awareness of this on his part, he seemed cogent and aware of the current context of a conversation. In one instance, he was asked to study a sequence of three digits which he was told to try to memorize. He described how he

considered the relationship among the digits and was able to retain them over a period of a couple of minutes while he was not distracted from this activity. However, after only a brief distraction to another topic, he failed completely to recall that he had been attempting to retain a set of digits, nor could he recall what they were.

Milner (1972) proposed an explanation for these facts by supposing that the human brain relies on three distinct forms of memory. The first, described as *primary memory* by the 19th century psychologist William James, consists of the current contents of thought. The contents of primary memory appears to decay over an interval of about 20 seconds after an individual takes up an alternative line of thought. This form of memory appears not to depend on the MTL, since it appears to remain intact in HM and other patients with similar patterns of brain damage. The second form of memory, which we will call MTL-dependent memory, is required for the rapid initial formation of memories for situations and their contents, including the objects that occur in them and the words used to describe these objects. The third, which we will call consolidated memory, is sufficient for the comprehension of language, and includes all of the knowledge of objects, their names, and the typical situations in which they occur, that underlie the inference processes we described above. As Milner (1972) described, this form of memory is acquired gradually, so that retention of knowledge acquired in a period of time before the removal of the MTL will be impaired. The extent of the time dependence has been the focus of extensive research; it can vary from days to decades depending on the species, age of participants and many other factors (Winocur, 1990; Cohen & Squire, 1981; MacKinnon & Squire, 1989).

An explicit mechanistic characterization of the second and third forms of memory has been developed within the PDP framework, linking findings from human studies with findings from studies in non-human animals (McClelland, McNaughton & O'Reilly, 1995; Kumaran, Hassabis & McClelland, 2016). This characterization draws on seminal neuro-computational ideas proposed by David Marr (1970; 1971) that were then carried forward into the PDP era by McNaughton (McNaughton & Morris, 1987), Rolls (Treves & Rolls, 1994), and others.

Consolidated memory in connections within the neocortex

The account begins with the idea that consolidated memory is structured knowledge, stored in the connections among neurons outside of the medial temporal lobes, primarily in the neocortex. According to the theory, it is these connections that allow patterns of activation in each region of the neocortex to constrain the construction of patterns of activity in other neocortical areas, subserving the process of constructing situation and object representations as described above, as well as underlying perceptual learning, motor skill learning, and the acquisition of expertise in a wide range of cognitive domains. These connections are thought to be established gradually through the accumulation of changes to connections within the neocortical system. As an example suited to our focus on word learning, the connections mediating the process of reading words aloud (i.e., translating from orthography to phonology) are thought of as being established by the accumulation of adjustments occurring each time a word is read to the connections between neurons within and between the visual and phonological word form areas. In this way, the correct pronunciation of a word is not learned in a single, or even a few trials, but rather, the connections for producing it are established gradually, over an extended period, during which the connections required for pronouncing other words are also gradually acquired. This learning process is model in early PDP models of reading (Sejnowski & Rosenberg, 1987; Seidenberg & McClelland, 1989) and language understanding (St. John & McClelland,

1990) as well as models that learn the relationships between objects, their properties and their names (Rumelhart & Todd, 1993; McClelland & Rogers, 2003). Such a process also underlies the procedures whereby deep neural network models of language processing learn to perform tasks ranging from word prediction (Zaremba *et al,* 2104) to language-to-language translation (Wu *et al,* 2016) to grounded use of language within a virtual environment (Hermann, Hill *et al,* 2017).

The kind of learning system we have just described learns by a process that is best characterized as prior-knowledge-dependent learning (McClelland, 2013). During initial exposure to a new domain, learning is intrinsically slow in deep neural networks, because the propagation of learning signals that allows learning at one layer of a network depends on knowledge in the weights in other layers of the network (Saxe, McClelland & Ganguli, 2013a). Once knowledge in the weights has been built up, learning will proceed more rapidly, allowing deep neural network models to account for stage-like developmental transitions (Rogers & McClelland, 2004; Saxe, McClelland & Ganguli, 2013b). Even after knowledge has been built up, the acquisition of arbitrary new associative information will also generally be very slow, although information that is highly consistent with what is already known can be learned rapidly (McClelland, 2013). Because words are typically highly arbitrary in their relation to objects, initial learning that links a word form with an object name will proceed slowly within such systems. Notably, however, when novel words or phrases designating objects *are* largely consistent with existing knowledge, amnesic patients can learn these associations very rapidly (Duff, Hengst, Trenel & Cohen, 2006). For example, if an amnesic patient views a randomly constructed pattern of shapes that suggests a recumbent person wearing a Mexican hat, the patient and an interlocutor may settle in the course of discussion on a convention of describing this pattern as 'sombrero man'. In this way an amnesic can acquire new phrases to describe a dozen or so randomly constructed patterns at a normal rate, in the course of a couple of ½ hour sessions.

Role of the MTL in learning and Memory: Initial acquisition and use of information about a new object

Given the characterization of consolidated memory and the nature of the system for representation, processing and learning in which it is embedded that we described in the previous section, we are now in a position to discuss the nature and role of the MTL dependent memory system itself. In essence, this system is seen as forming a distinct representation of the distributed pattern of activity in the neocortex that arises in processing each episode we experience, which is then stored through large changes to connections within the medial temporal lobes, supporting subsequent use of the contents of this episode. For example, when a person reads or hears the sentence about the wompamuck, the distributed pattern for the entire processing episode would give rise to a distinct representation of it within the hippocampus. Should the same person hear the word wompamuck subsequently in another sentence, bi-directional connections between the hippocampus and the neocortex as well as the intra-hippocampal connections would support the re-instatement of this representation in relevant neocortical areas. Ordinarily, in memory research, this process is thought of as supporting explicit memory for the episode as well as the characteristics of the object designated by the word. Within the context of language understanding, however, we might envision a new sentence such as 'Bill saw a wompamuck and it …' as leading to the prediction of object and situation characteristics as well as subsequent possible linguistic input consistent with the representation formed as a result of processing the initial episode, and results of studies using the N400 response bear this out (Menenti, Petersson, Scheeringa & Hagoort, 2009). These characteristics would be quite different from those we would

predict if the previous sentence were 'John saw a fierce wompamuck chasing a gazelle' or 'John dropped his wompamuck and it broke into 1,000 tiny pieces'.

Knowledge dependence of MTL input. It is important to note that the pattern of activity that is made available to the MTL depends heavily on knowledge already encoded in the connections within the neocortex. This knowledge structures the patterns of neocortical activity; since the knowledge is acquired gradually, the representations in turn change over developmental time. In addition, according to the complementary learning systems theory, the connections that mediate between the neocortex and the MTL (shown in blue in Figure 1) are also thought to be subject to gradual learning and likewise prior-knowledge-dependent. Thus, learning within the neocortex and between the neocortex and the MTL plays an important role in allowing connection weight changes within the MTL itself to support new learning and subsequent use of the contents of newly-experienced episodes.

Integration of new information into the neocortex: interleaved learning

So far, we have characterized the initial, MTL-dependent, memory formation process that is essentially absent after removal of the medial temporal lobes. We now consider how this knowledge may ultimately be integrated into neocortical networks, so that it is no longer MTL dependent.

Based on earlier observations, we can already see that even after removal of the MTL, gradual learning within the neocortical system may still occur. HM's gradual acquisition of an understanding of his own condition and of John F. Kennedy's presidency and assassination (Milner, Corkin & Teuber, 1968) indicates that gradual learning is still possible in the neocortex. The fairly rapid learning of new language conventions that are largely consistent with pre-existing knowledge appears also to be possible without the MTL. In both cases, according to the complementary learning systems theory, this learning occurs directly in the connections supporting the structured knowledge system in the neocortex.

Why is cortical learning usually slow? A crucial question arises at this point: If knowledge is ultimately stored in the connections among neurons in the neocortex, why isn't the brain set up in such a way as to allow the integration of new information into the neocortex immediately? We have already seen a part of the reason for this: If (as the CLS proposes) the neocortical learning system is a deep neural network, new learning of arbitrary associations mediated by many layers of neurons and connections will necessarily be slow. There is a second, equally important reason as well: In a multi-layer network that has learned a structured body of knowledge, making large adjustments to accommodate information even partially inconsistent with prior knowledge can result in interference with the information already known (McClosky & Cohen, 1989). This was illustrated in McClelland et al (1995) by considering the case of a penguin, which was partially inconsistent with a structured body of pre-existing knowledge a network had acquired based on examples of typical birds and fish. A penguin shares some properties with other animals, but it is only partially consistent with prior knowledge, in that some of its properties are consistent with those of other birds, while others of its properties are more consistent with those of fish. Through repeated presentation of information about a penguin, it is possible to force a neocortex-like neural network to learn these characteristics of the penguin quickly, but by doing so the result is what McClosky and Cohen (1989) called 'catastrophic interference' with the information previously known about other birds and fish. To state the matter generally, learning in deep networks generally requires *interleaved learning*, involving many repetitions of experience with each item, interleaved with presentations of other items. Only information highly consistent with what is already known can be incorporated into neocortical networks quickly.

In sum, the key idea arising from complementary learning systems theory is that the MTL and cortical learning systems provide a solution to the catastrophic interference problem. Working together, they allow for rapid initial storage and subsequent use of new learning, while avoiding catastrophic interference, complemented by a slower learning process that occurs in the neocortex such that the new information can be gradually integrated into the structured neocortical knowledge system, interleaved with ongoing exposure to other experiences.

Replay and the dialog between MTL and neocortex. One way new information may be gradually integrated into neocortical networks is through repeated exposure based accumulated effects of repeated relevant experiences. This is likely to be the basis on which HM came to appreciate his own condition and learned about the presidency and assassination of JFK (Kennedy's huge popularity would have led to extensive exposure to both of these situations, even though the assassination itself was a single episode, since it was the focus of media attention over a very extended period in the early 1960's). In addition, however, the process of integrating new information into neocortical networks is also thought to occur as a result of off-line reactivation and reply of patterns of activation initially stored in the MTL.

A large body of neurobiological research in rodents supports the view that spontaneous replay of short snippets of previously experienced episodes occurs within the MTL during sleep, and such activity also occurs while animals are resting between episodes in which they are actively engaged in exploring an environment to find food rewards (See Kumaran, Hassabis & McClelland, 2016, for a review). Replay episodes arising within the MTL appear to be coupled with replay episodes in the neocortex as well. Based on these findings, it has been proposed that the hippocampus and neocortex engage in a dialog of sorts during sleep, such that replay events originating either in the hippocampus or the neocortex can prime replay of related information in cortex. Within the complementary learning systems theory, these replay events can be understood as promoting a selective form of interleaved learning, such that those items that are prone to interfere with each other will be replayed more than others, and such selective replay can increase the efficiency of interleaved learning (McClelland, Lampinen, and McNaughton, in preparation). There is also a large body of human research on the role of sleep in the formation of representations that influence on-line processing and support use of new learning to support inferences bridging multiple items of learned information. Several studies have shown that these influences of new learning may not be observed if assessed within an hour or so of the initial learning experience or after several hours being awake, but may be observed after several hours if those hours included a period of sleep.

These results just reviewed have often been taken as suggesting that integration into neocortical networks can be completed overnight. However, the fact that in humans, MTL lesions result in loss of memories that may have been formed years or even decades prior to the lesion (MacKinnon & Squire, 1989), seems more consistent with the view that, in humans at least, a first period of sleep may strengthen the intra-MTL connections and stabilize the memory (see Kumaran & McClelland, 2012, for further discussion). In any case, once integration of new information into the connections within the neocortex has occurred, it will affect subsequent inference and processing independent of the involvement of the MTL.

The neural substrate of primary memory

Thus far we have discussed the mechanistic basis of MTL-dependent and consolidated memory, but not the mechanistic substrate of primary memory. We briefly contrast two ways of thinking about primary memory, noting that these need not be mutually exclusive. Indeed, it seems likely both play important roles in the short-lived, MTL-independent consequences of very recent experience.

Primary memory as ongoing activity. The first idea, discussed by William James over 120 years ago, is the proposal that primary memory consists of ongoing activity that continues after an input has come and gone. Recurrent neural networks provide a mechanism that allows neural activity at a particular time to reflect input that occurred at earlier times, and this idea has been employed in several models that attempt to capture how input at one point in time can influence processing of an input occurring at a later time; a number of models developed in the late 1980's and early 1990's explored these ideas. An important theme in much of this work was that gradual learning of the type we have attributed to the neocortex may play an important role in promoting the development of connections that establish and maintain information for subsequent use (Cleeremans & McClelland, 1989; Munakata, McClelland, Johnson & Siegler, 1994). Thus, primary memory may depend on consolidated memory, just as MTL-dependent memory does.

Recent work with deep neural networks has vastly extended these ideas beyond what was possible with the mechanisms of learning and processing in use in modeling cognition using neural networks in the 1980's and early 1990's. In the late 1990's, an extension of recurrent neural networks based on Long Short Term Memory (Hochreiter & Schmidthuber, 1997), and over the last 10 years, this mechanism has become central to deep neural network models of language processing, including contemporary language prediction models (Zaremba *et al*, 2014) and the Google neural machine translation system (Wu *et al*, 2016). These newer models share with their predecessors the assumption that the storage and retention of information in a sustained pattern of activation over units in a neural network is learning dependent. A further extension of these ideas, implemented in the Differentiable Neural Computer (Graves, Wayne, et al, 2016) and other models that focus specifically on memory (Santoro et al, 2016) demonstrates how gradual learning processes may establish the knowledge that determines what information we store in maintained activity patterns, and how and when we use and then forget information presented within a learning episode. These ideas can also be thought of as contributing to learned policies for determining when we store, retrieve, and forget information through connection weight changes within the MTL.

Primary memory as large changes to connections that rapidly decay. The second possible mechanism for primary memory also has an extensive history. This is the idea that changes to connection weights may have a larger, short-lasting component as well as a a smaller, longer-lasting, residual component. This idea is supported by studies of biological synaptic plasticity: When synaptic connections are activated by action potentials arising from a sending neuron, a chemical substance called a neurotransmitter is released. If the receiving neuron is sufficiently depolarized within a short time window after transmitter release, a cascade of events occurs on the receiving side of the synapse, making it more responsive to subsequent input (McNaughton, Douglas & Goddard, 1978; Barrionuevo & Brown, 1983). This change in responsiveness has a short-lived component, decaying back to baseline over seconds or minutes, as well as a smaller, longer lasting component, that may persist for hours or days (McNaughton & Morris, 1987). For present purposes, the main point is that the short-lived

components provide a substrate that may provide at least one contributor to short-term retention of information, independent of the medial temporal lobe memory system. Variants of this idea were used in a number of early machine-learning (Hinton, & Plaut, 1987) and cognitive neuroscience-focused models of learning and memory (Hinton & Sejnowski, 1986; McClelland & Rumelhart, 1985). Related ideas have recently been employed in contemporary deep-learning models (Ba *et al.,* 2016; Sprechmann *et al,* 2018).

Interactive, Distributed Processing, Complementary Learning Systems, and Brain Based Knowledge of Word Meanings

In summary, we repeat the key points with which we began this section:

1. Within the interactive, distributed processing framework for inferring word meaning that we have described, word representations are always *constructed* based on multiple sources of information.

2. Within the complementary learning systems framework for understanding integration of the results of inference for later use, the knowledge that supports this construction depends on sustained neural activity within the neocortex, initial MTL dependent learning based on synaptic plasticity in that part of the brain, and on gradually-acquired changes to the strengths of connections within and among the various contributing neocortical areas participating in the representation and reconstruction process.

As we noted at the outset of this article, these points contrast with a focus on Embeddings *per se* as the representation of word knowledge. It is true that the pattern of activation we have identified as the representation of the object designated by a word plays a role similar to that played by the embedding vector as envisioned under Embeddings 1.0. However, this pattern is not directly stored, as it is in a classical embedding framework. Furthermore, and more important, the knowledge underlying the ability to use a word appropriately in context – knowledge we traditionally ascribe to lexical entries within classical theories of language knowledge – is distributed widely throughout the entire neuro-computational system underlying language representation and use. Once that system has been established over the first years of life learning language in context, a child will then have the capacity to acquire new words rapidly; and with continual learning over a lifetime, more and more prior experience will inform exactly what it is that is learned. The embedding-like representation will play an essential role in mediating the learning and use of this information, but the knowledge that supports it will be distributed over the connections throughout the extensive distributed network of brain areas as depicted in Figure 1.

Bibliography

Altmann, G. T., & Kamide, Y. (2009). Discourse-mediation of the mapping between language and the visual world: Eye movements and mental representation. *Cognition*, *111*(1), 55-71.

Ba, J., Hinton, G. E., Mnih, V., Leibo, J. Z., & Ionescu, C. (2016). Using fast weights to attend to the recent past. In *Advances in Neural Information Processing Systems* (pp. 4331-4339).

Baggio, G., & Hagoort, P. (2011). The balance between memory and unification in semantics: a dynamic account of the N400. *Language and Cognitive Processes*, *26*(9), 1338-1367.

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709-721.

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior*, *13*(4), 471-481.

Barrionuevo, G., & Brown, T. H. (1983). Associative long-term potentiation in hippocampal slices. *Proceedings of the National Academy of Sciences*, *80*(23), 7347-7351.

Bozeat, S., Ralph, M. A. L., Patterson, K., Garrard, P., & Hodges, J. R. (2000). Non-verbal semantic impairment in semantic dementia. *Neuropsychologia*, *38*(9), 1207-1215.

Cleeremans, A., & McClelland, J. L. (1991). Learning the structure of event sequences. *Journal of Experimental Psychology: General*, *120*(3), 235.

Cohen, N. J., & Squire, L. R. (1981). Retrograde amnesia and remote memory impairment. *Neuropsychologia*, *19*(3), 337-356.

Cole, R. A., Jakimik, J., & Cooper, W. E. (1978). Perceptibility of phonetic features in fluent speech. *The Journal of the Acoustical Society of America*, *64*(1), 44-56.

Duff, M. C., Hengst, J., Tranel, D., & Cohen, N. J. (2006). Development of shared information in communication despite hippocampal amnesia. *Nature neuroscience*, *9*(1), 140.

Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwińska, A., ... & Badia, A. P. (2016). Hybrid computing using a neural network with dynamic external memory. *Nature*, *538*(7626), 471.

Hinton, G. E., & Plaut, D. C. (1987). Using fast weights to deblur old memories. In *Proceedings of the ninth annual conference of the Cognitive Science Society* (pp. 177-186).

Hinton, Geoffrey E., and Terrence J. Sejnowski. Learning and relearning in Boltzmann machines. In D. E. Rumelhart & J. L. McClelland (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, Vol 1*, 282-317.

Heider, F. and Simmel, M. (1944). An experimental study of apparent behavior. *Am. J. Psychol.* 57, 243–249.

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*.

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

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

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*(1), 133-156.

Kumaran, D., Hassabis, D., & McClelland, J. L. (2016). What learning systems do intelligent agents need? Complementary learning systems theory updated. *Trends in cognitive sciences*, *20*(7), 512-534.

Lakoff, G., & Johnson, M. (2008). *Metaphors we live by*. University of Chicago press.

Lambon Ralph, M. A., McClelland, J. L., Patterson, K., Galton, C. J., & Hodges, J. R. (2001). No right to speak? The relationship between object naming and semantic impairment: Neuropsychological evidence and a computational model. *Cognitive Neuroscience. 13:3*, 341-356.

Mackinnon, D. F., & Squire, L. R. (1989). Autobiographical memory and amnesia. *Psychobiology*, *17*(3), 247-256.

Marchand, H. (1969). *The categories and types of present-day English word-formation: A synchronic-diachronic approach*. Beck.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive psychology*, *20*(2), 121-157.

Marr, D. (1970). A Theory for Cerebral Neocortex. *Proceedings of the Royal Society of London Series B*, *176*, 161-234.

Marr, D. (1971). Simple memory: a theory for archicortex. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences*, *262*(841), 23.

McClelland, J. L. (1987). The case for interactionism in language processing. In M. Coltheart (Ed.), *Attention & performance XII: The psychology of reading* (pp. 1-36). London: Erlbaum.

McClelland, J. L. (1992). Can connectionist models discover the structure of natural language? In Morelli, R., Brown, W. M., Anselmi, D., Haberlandt, K., Lloyd, D. (Eds.) *Minds, Brains & Computers,* pp. 168-189. Ablex Publishing: Norwood, NJ.

McClelland, J. L. (2013). Incorporating rapid neocortical learning of new schema-consistent information into complementary learning systems theory. *Journal of Experimental Psychology: General*, *142*(4), 1190.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419-457.

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

McClelland, J. L., & Rumelhart, D. E. (1985). Distributed memory and the representation of general and specific information. *Journal of Experimental Psychology: General*, *114*(2), 159.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation* (Vol. 24, pp. 109-165). Academic Press.

McNaughton, B. L., Douglas, R. M., & Goddard, G. V. (1978). Synaptic enhancement in fascia dentata: cooperativity among coactive afferents. *Brain research*, *157*(2), 277-293.

McNaughton, B. L., & Morris, R. G. (1987). Hippocampal synaptic enhancement and information storage within a distributed memory system. *Trends in Neurosciences*, *10*(10), 408-415.

Menenti, L., Petersson, K. M., Scheeringa, R., & Hagoort, P. (2009). When elephants fly: differential sensitivity of right and left inferior frontal gyri to discourse and world knowledge. *Journal of Cognitive Neuroscience*, *21*(12), 2358-2368.

Miller, G. A., Heise, G. A., & Lichten, W. (1951). The intelligibility of speech as a function of the context of the test materials. *Journal of Experimental Psychology*, *41*(5), 329.

Milner, B. (1972). Disorders of learning and memory after temporal lobe lesions in man. *Neurosurgery*, *19*(CN_suppl_1), 421-446.

Milner, B., Corkin, S., & Teuber, H. L. (1968). Further analysis of the hippocampal amnesic syndrome: 14-year follow-up study of HM. *Neuropsychologia*, *6*(3), 215-234.

Munakata, Y., McClelland, J. L., Johnson, M. H., & Siegler, R. S. (1997). Rethinking infant knowledge: Toward an adaptive process account of successes and failures in object permanence tasks. *Psychological Review*, *104*(4), 686.

Nieuwland, M. S., & Van Berkum, J. J. A. (2006). When peanuts fall in love: N400 evidence for the power of discourse. Journal of *Cognitive Neuroscience, 18*, 1098–1111.

Patterson, K., Ralph, M. A. L., Jefferies, E., Woollams, A., Jones, R., Hodges, J. R., & Rogers, T. T. (2006). "Presemantic" cognition in semantic dementia: Six deficits in search of an explanation. *Journal of Cognitive Neuroscience*, *18*(2), 169-183.

Plaut, D. C., & Gonnerman, L. M. (2000). Are non-semantic morphological effects incompatible with a distributed connectionist approach to lexical processing? *Language and Cognitive Processes*, *15*(4-5), 445-485.

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

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (*in press*). Change in an implicit probabilistic representation captures meaning processing in the brain. *Nature Human Behavior.*

Ranganath, C., & Ritchey, M. (2012). Two cortical systems for memory-guided behaviour. *Nature Reviews Neuroscience*, *13*(10), 713.

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

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

Rogers, T. T. & McClelland, J. L. (2014). Parallel Distributed Processing at 25: Further Explorations in the Microstructure of Cognition. *Cognitive Science, 6*, pp. 1024-1077. DOI: 10.1111/cogs.12148.

Rumelhart, D. E. (1977). Toward an interactive model of reading. In S. Dornic (Ed.), *Attention and Performance VI*. (pp. 573–603). Hillsdale, NJ: LEA. Reprinted as: Rumelhart, D. E. (1994). Toward an interactive model of reading. In R. B. Ruddell, M. R. Ruddell, & H. Singer (Eds.), *Theoretical models and processes of reading* (pp. 864-894).

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

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.

Santoro, A., Bartunov, S., Botvinick, M., Wierstra, D., & Lillicrap, T. (2016). One-shot learning with memory-augmented neural networks. *arXiv preprint arXiv:1605.06065*.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013a). Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv preprint arXiv:1312.6120*.

Saxe, A. M., McClelland, J. L., & Ganguli, S. (2013b). Learning hierarchical categories in deep neural networks. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 35, No. 35).

Schapiro, A. C., McClelland, J. L., Welbourne, S. R., Rogers, T. T., & Lambon Ralph, M. A. (2013). Why bilateral damage is worse than unilateral damage to the brain. *Journal of Cognitive Neuroscience, 25*(12), 2107-2123. doi:10.1162/jocn_a_00441.

Scoville, W. B., & Milner, B. (1957). Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, *20*(1), 11.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523.

Sejnowski, T. J., & Rosenberg, C. R. (1987). Parallel networks that learn to pronounce English text. *Complex systems*, *1*(1), 145-168.

Simpson, G. B. (1984). Lexical ambiguity and its role in models of word recognition. *Psychological Bulletin*, *96*(2), 316.

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.

Sohoglu, E., Peelle, J. E., Carlyon, R. P., & Davis, M. H. (2012). Predictive top-down integration of prior knowledge during speech perception. *Journal of Neuroscience, 32*(25), 8443–8453.

Sprechmann, P., Jayakumar, S. M., Rae, J. W., Pritzel, A., Badia, A. P., Uria, B., ... & Blundell, C. (2018). Memory-based parameter adaptation. *arXiv preprint arXiv:1802.10542*.

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

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632-1634.

Treves, A., & Rolls, E. T. (1994). Computational analysis of the role of the hippocampus in memory. *Hippocampus*, *4*(3), 374-391.

Warren, R. M. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*(3917), 392-393.

Warrington, E. K., & Shallice, T. (1984). Category specific semantic impairments. *Brain*, *107*(3), 829-853.

Winocur, G. (1990). Anterograde and retrograde amnesia in rats with dorsal hippocampal or dorsomedial thalamic lesions. *Behavioral Brain Research, 38*, 145-154.

Wittgenstein, L. (1953). *Philosophical Investigations.* Blackwell.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., ... & Klingner, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zadbood, A., Chen, J., Leong, Y. C., Norman, K. A., & Hasson, U. (2017). How we transmit memories to other brains: constructing shared neural representations via communication. *Cerebral Cortex*, *27*(10), 4988-5000.

Zaremba, W., Sutskever, I., & Vinyals, O. (2014). Recurrent neural network regularization. *arXiv preprint arXiv:1409.2329*.

Zwaan, R. A., & Radvansky, G. A. (1998). Situation models in language comprehension and memory. *Psychological Bulletin, 123*(2), 162–185.