Department of Computer Science,
University of Otago

Technical Report OUCS-2016-01

A simulationist model of episode representations in working memory

Authors:

Martin Takac
Department of Computer Science, University of Otago, New Zealand
and
Centre for Cognitive Science, Comenius University, Slovakia

Alistair Knott
Department of Computer Science, University of Otago, New Zealand

Department of Computer Science,
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

http://www.cs.otago.ac.nz/research/techreports.php

A simulationist model of episode representations in working memory

Martin Takac[a,b], Alistair Knott[a]

[a]*Department of Computer Science, University of Otago, New Zealand*
[b]*Centre for Cognitive Science, Comenius University, Slovakia*

Abstract

We present a neural network model of how the brain encodes episodes and individuals in semantic working memory (WM). The model rests on the assumption that concrete episodes, and the individuals that participate in them, are perceived through sensorimotor routines with well-defined sequential structure. In our model, episodes and individuals are stored in semantic WM as *prepared sensorimotor routines*, that can be internally replayed. This assumption allows a neural model of semantic representations with several new features. Firstly, the model provides a novel account of binding: specifically, of how individuals are bound to semantic roles in episodes (such as AGENT and PATIENT). Secondly, it gives a novel account of hierarchical structures in semantic representations: specifically, an account of how the representations of individuals participating in an episode are embedded within a representation of the episode. Thirdly, the model is able to represent large probability distributions over episodes, and over the properties of their participant individuals, rather than just representations of a few selected episodes and individuals. This capability allows a novel account of how top-down expectations influence processing in real time during the experience of new episodes. Finally the model supports a novel account of the interface between WM and long-term memory.

Keywords: semantic working memory, embodied cognition, simulation, neural binding, neural networks

1. Introduction

While traditional models of human working memory (WM) focus on memory for phonological and visuospatial material (Baddeley and Hitch, 1974; Baddeley, 1992), in recent years a consensus has emerged that the human WM system also stores representations that should be termed 'semantic' (Martin *et al.*, 1994; Fletcher and Henson, 2001; Fiebach *et al.*, 2007; Jackendoff, 2010, Shivde and Anderson, 2011). This extended conception of WM partly stems from Baddeley's own revision of his classic model, to include an 'episodic buffer' holding modality-independent semantic representations (Baddeley, 2000). Baddeley argues for this medium on several grounds. One relates to memory storage: Baddeley argues that an episode occurring in the world cannot be directly encoded in hippocampal long-term memory (LTM), because hippcampal learning mechanisms operate over short timescales, on the order of 100ms (Abraham *et al.*, 2002) while episodes in the world take much longer than this to occur. Another argument relates to behaviour: Baddeley argues that an agent must maintain a representation of 'the current situation', identifying events which have recently occurred, and supporting decisions about future actions and events, to explain how patients with impaired long-term memory can perform relatively normally on complex tasks such as playing bridge. A final argument relates to language. In phonological WM experiments, subjects are much better at recalling well-formed sentences than unrelated sequences of words: Baddeley argues this is because the meaning of a sentence, stored separately from its phonology, provides an additional cue to recall. In fact Baddeley proposes that the episodic buffer can hold more than the meaning of a single sentence; he also invokes the episodic buffer to explain why amnesic patients, who are unable to retain information about a text for any length of time, can nonetheless recall the 'gist' of a paragraph of text over a short period.

In Baddeley's account, a key role for the episodic buffer is to interface between several other cognitive modalities. It is a medium linking the low-level sensorimotor (SM) representations produced as episodes are experienced in the world to the more abstract representations of episodes stored in hippocampal LTM; at the same time, it has links to phonological WM, to explain how semantic representations can aid phonological recall. The notion that semantic WM representations are an important interface medium is echoed by many subsequent studies. In Fletcher and Henson's (2001) account, semantic information to be encoded in LTM must first be maintained and/or elaborated in WM, and semantic information retrieved from LTM also becomes active in WM. Other accounts focus on how semantic WM representations interface with language. Fiebach *et al.* (2007) find evidence that interpreting verbal stimuli results in the activation of semantic representations in WM; Shivde and Anderson (2011) find evidence that the semantic representations activated by verbal stimuli are maintained for the duration of semantic WM tasks (but crucially, not beyond this duration). Several studies find evidence that during sentence production, parts of the message to be realised by the sentence are stored in semantic WM prior to its generation (see e.g. Martin and Freedman, 2001; Martin *et al.*, 2010; Slevc, 2011; but see Martin *et al.*, 2014 for some caveats). Finally, there are accounts where semantic WM representations buffer information relevant to action. For instance in Itti and Arbib's (2006) model, an agent builds a representation of a 'minimal subscene' in his current environment, which binds the parameters for an action to be performed or observed. (In this model, the subscene representation also interfaces with language.)

While there is some agreement that semantic WM representations interface with several different modalities, there is still relatively little understanding about the form of semantic WM representations, and about how they are organised. In this paper, we address these questions in a neural network model of semantic WM. A computational model can make a useful contribution to a debate about the format of semantic WM representations, because it can integrate findings about the role of these representations as they interface with several different cognitive processes. If the semantic WM representations that interface with the SM system are the same as those that interface with language, and with LTM, experimental findings about all three interfaces define multiple simultaneous constraints on the format of semantic WM representations. In our model, we use constraints relating to one interface to help answer questions about other interfaces. In the first instance, a constraint on the semantic WM system relating to its interface with SM processing suggests a new solution to a 'binding problem' that features in network models of language processing: how to bind representations of the participants in an episode to the semantic roles they play (e.g. AGENT, PATIENT). The resulting model of semantic representations in turn suggests a new account of hierarchically structured syntactic domains within sentences, which we express in a model of sentence production. It also suggests a new way of representing 'situations' in semantic WM, in the context of a model of inference and decision-making.

In Section 2 we outline some requirements for a model of WM semantic representations, drawing on recent experiments as well as well-known results, and discussing interfaces with language, LTM and the SM system where appropriate. Our aim here is to suggest *linking hypotheses*, that suggest identities between elements of cognitive models of different cognitive systems, that help to state a single unified account of semantic WM. We also introduce a key proposal of our model: that experience of episodes in the world, and of the individuals that participate in them, happens through temporally extended SM routines with well-defined sequential structure, and that semantic representations in WM retain this structure. In our model, individuals and episodes are represented in semantic WM as *prepared SM routines*, that can be internally rehearsed or simulated. In Section 3 we outline a new model of semantic role-binding that exploits this idea, and discuss how the new model supports a new conception of hierarchical domains in syntax, and a new way of representing whole 'situations' using probability distributions over episodes. The model itself is introduced in Section 4, and its training and evaluation are described in 5. In Sections 6 and 7 we describe how the model interfaces with language, and with long-term memory. We conclude with a discussion in Section 8.

2. The structure of the semantic WM system

What types of information does the semantic WM system hold? Baddeley's (2000) model focusses on two types of information: semantic WM is able to hold a single episode representation, encoding the meaning of a recently-heard sentence, or an experienced episode prior to its storage in LTM, but it is also able to store representations larger than a single episode, expressing the gist of a paragraph of text, or the state of play of a complex game. We will add one other type of information to the list: if semantic WM is able to represent a single episode, it must also be able to represent the agents or objects that *participate* in this episode. We will refer to the representations larger than a single episode as **situations**, and the entities that participate in a single episode as **individuals**. In this section we review what is known about the WM storage of individuals, episodes and situations, and discuss how these storage systems relate to one another.

2.1. WM representations of individuals

While some aspects of our WM for individuals are covered in Baddeley's model by the 'visuospatial sketchpad', that holds WM representations of visually presented stimuli, the topic of WM for individuals is in fact a research field in its own right, of great relevance for an account of semantic WM. Research into WM representations of objects and their properties has been conducted within two experimental paradigms, one using behavioural experimental methods and one using brain imaging and recording methods. We will consider these separately.

Behavioural methods. A starting point for many recent behavioural studies is Luck and Vogel's (1997) investigation of subjects' ability to detect a change in a visually presented array of objects. Luck and Vogel made two key observations. Firstly, subjects' ability to detect a change deteriorated rapidly if there were more than four objects in the presented array. This indicates a capacity limit of some kind for the WM resource used to represent the objects. Secondly, subjects' performance did not depend on whether change was constrained to occur in one visual 'feature' of the objects (e.g. colour) or could occur in one of several features (colour, size, orientation and shape). This suggests that the capacity limit is not defined in units of visual features. Luck and Vogel proposed instead that the capacity limit is defined in units of whole objects—specifically, that the WM medium used to store the properties of objects contains four 'slots', which can each store a complete object representation. This proposal is supported by several other experiments; see e.g. Awh *et al.* (2007); Anderson *et al.* (2011). (There is also interesting evidence that each 'slot' can hold a homogeneous *group* of individuals of a given type, as well as a single token individual; see in particular Feigenson, 2008). However, there are also experimental results suggesting that the capacity of WM for object properties is defined in more abstract informational units. Several studies have shown that success in the change-detection task can depend on the complexity of the features that distinguish objects as well as just on the number of objects or groups. Complexity can be defined in terms of the subtlety of colour variations (Wilken and Ma, 2004) or the use of 3D rather than 2D stimuli (Alvarez and Cavanagh, 2004); in either case, if the amount of information that must be retained about each object is increased, the number of objects that can be simultaneously encoded in WM is less than four. A similar reduction is observed if objects are distinguished by properties of their sub-parts (see e.g. Davis and Holmes, 2005). In summary, the medium that holds WM representations of objects and their properties appears to make use of both object-based encodings and feature-based encodings. How it does this remains a fairly open question.

Brain localisation methods. The brain imaging and recording strand of research into WM object representations focusses on analysing the patterns of neural activity and connectivity that occur during the delay period of WM tasks. Aspects of these patterns must encode the information that is retained about objects and their properties. One clearcut finding is that during the delay period of a WM task, there are patterns of activity outside the primary sensory and motor areas, primarily in prefrontal cortex (PFC), that encode aspects of the material to be retained. Single-cell studies of macaque PFC provide particularly clear data about the nature of these patterns.

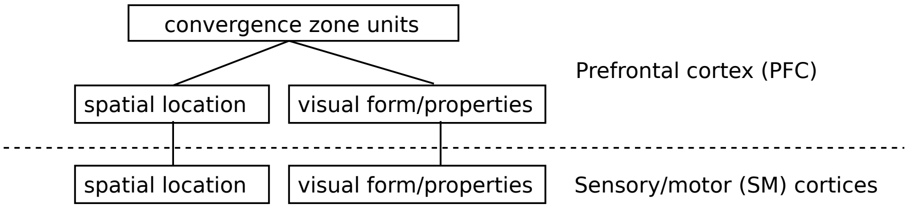


Figure 1: A simple model of PFC mechanisms for holding representations of individuals in WM

Some PFC cells encode individual features of the objects being remembered—for instance, just the location of an object, or just an intrinsic property such as colour or size (see e.g. Wilson *et al.*, 1993; Rao *et al.*, 1997). Others encode task-relevant combinations of features, such as location and colour (Rao *et al.*, 1997; Rainer *et al.*, 1998). Similar results have been found in human PFC (see e.g. Curtis and D'Esposito, 2003; Takahashi *et al.*, 2013). These studies provide good evidence that there are circuits in PFC that can maintain representations of perceptually-encountered objects, even when these objects are not present before the senses. They also suggest that PFC units can hold *associations* between the features of objects in WM. A unit that holds associations between several features must be linked to units that represent each of these features individually: it functions as a 'convergence zone', whose meaning resides in its pattern of links to assemblies in other areas (see Damasio and Damasio, 1994). In some prefrontal circuits such associative links are learned gradually, over extended periods (e.g. Averbeck *et al.*, 2006). But there are also synaptic mechanisms in PFC that operate over short timescales, allowing PFC units to hold short-term associations between stimuli in service of WM tasks (see Stokes, 2015 for a review). The picture of WM representations that emerges so far is sketched in Figure 1. When an object is presented, its properties (e.g. location and form) are represented in assemblies in various sensory areas (predominantly parietal and inferotemporal cortex). These assemblies activate corresponding assemblies in PFC, that again hold particular properties of objects, but can sustain their activity during a delay period when the object is removed. Associations between these PFC assemblies are stored in temporary connections to units in a 'convergence zone', which can also sustain their activity.

One question that arises with this scheme is why there should be intermediate prefrontal units that store the spatial location and intrinsic properties of a perceived object separately. This might be because in some cases the agent is not required to represent all the features of the object in WM. However, as discussed earlier, there is often no penalty in representing more than one of an object's features. We would like to raise another possibility: the location and intrinsic properties of a perceived object are in fact activated at different *times* in the SM cortices, as well as at different locations. The spatial location of an object must be selected and attended to before the object can be categorised in any detail (for classic results, see Treisman and Gelade, 1980, and for recent single-cell evidence see e.g. Moore and Armstrong, 2003; Zhang *et al.*, 2011). In addition, representations of the attended spatial location in primary SM areas fade quite fast (Posner *et al.*, 1984; Ro *et al.*, 2003; Müller and Kleinschmidt, 2007), so there is some utility in storing the location and properties of an object separately in WM buffers, where they can be represented in parallel. In fact the ability of PFC to maintain parallel representations of stimuli that occur sequentially is seen by many as central to its role in WM; and it will be central to our model.

Note that the convergence-zone scheme sketched in Figure 1 can model WM for multiple objects, as well as WM for associations between properties of a single object. It cannot hold multiple object representations in parallel: if we activate the location and properties of two objects simultaneously, there is a well-known 'binding problem', whereby both locations are indiscriminately associated with both sets of properties. However, a convergence-zone unit can learn an association between the features of one object, and retain this for a short period even after it ceases being active. Assume two objects, O_1 and O_2, are attended to in succession. When O_1 is presented, convergence-zone units can be recruited to hold associations between its properties. Then PFC activity can be reset and O_2 can be presented, with a different set of convergence-zone units being recruited to hold associations between O_2's properties. Now activating a property of O_1 will retrieve a pattern in the convergence zone layer that will reactivate the other properties of O_1, and similarly for O_2. In fact, if the pool of convergence-zone units is small enough, and properties are represented with distributed encodings, then crosstalk within the network could explain the 'informational' limits on WM for multiple objects discussed above (Wilken and Ma, 2004; Alvarez and Cavanagh, 2004).

The model of prefrontal WM sketched in Figure 1 is somewhat complicated by another set of findings. As well as activity in PFC during the delay period of a WM task, there is also often activity in primary sensory areas, or in motor areas, if there are motor representations of objects to be maintained: see D'Esposito (2007) for a review of these findings. For instance, in macaque, there is delay-period activity in inferotemporal cortex, reflecting intrinsic object properties retained in WM (Nakamura and Kubota, 1995; Takeda *et al.*, 2005), and in intraparietal cortex, reflecting locations to be retained (Andersen and Buneo, 2002). Similar results are again found in humans (see e.g. Curtis *et al.*, 2004; Schluppeck *et al.*, 2006). Computationally, this delay-period activity in primary SM areas is hard to understand, because we expect it to *interfere* with any sensory or motor processing occurring during this time (Deco and Rolls, 2008). However, there is an interesting potential explanation, consistent with a model where PFC assemblies play the primary role in WM storing object properties. It could be that WM representations in PFC support the active *rehearsal* of the SM experiences they encode during the delay period—perhaps at rapid speeds that enable storage in the hippocampus. We can envisage a rehearsal mechanism that temporarily disengages SM areas from interaction with the world, and puts them under the control of PFC assemblies, to allow the re-enactment (possibly speeded) of the SM experiences they store. On this model, the convergence-zone representations of objects held in PFC (see Figure 1) can be *temporarily* activated during the delay period, in a manner that reactivates associated PFC assemblies, and also, indirectly, assemblies in primary SM areas. This hypothesis is supported by several recent studies showing that the neural populations in primary SM areas that encode object properties during a delay period are active transiently, rather than tonically. For instance, Meyers *et al.* (2008) showed that the cells in macaque inferotemporal (IT) cortex that encode the category of a remembered object do so better at particular times during the delay period, that endure on the order of a few hundred milliseconds. A similar result is found in an fMRI study of human extrastriate visual cortex (Sreenivasan *et al.*, 2014). There is some evidence that the reactivation of remembered stimulus properties happens cyclically, at particular phases in the theta cycle; see e.g. Lee *et al.* (2005) for evidence from macaque extrastriate visual cortex and Fuentemilla *et al.* (2010) for MEG evidence in humans. Poch *et al.* (2011) have shown in the human MEG data that this cyclic reactivation is coupled with cyclic activity in the hippocampus, and that the strength of this coupling correlates with performance in a WM task. There is also evidence for dynamic patterns of activity in PFC during WM maintenance. This is shown in the study by Meyers *et al.* (2008) discussed above. Finally, there is evidence that stimuli held in WM are encoded in temporally structured patterns of activity in PFC, not just tonically active patterns. Siegel *et al.* (2009) recorded activity in macaque PFC neurons during a task that required them to remember two objects, presented sequentially. They found that the PFC neurons that encoded the identity of objects during the delay period did so best at two specific phases of a 32Hz cycle. Moreover, information about the first object presented was best decoded at the first of these two phases, while information about the second object was best decoded at the second phase, 'as if. . . object information was multiplexed in the 32Hz oscillation according to object order'. Additionally, the 32Hz oscillation was modulated by a slower 3Hz cycle: spikes carried most information about both objects at a particular phase in this cycle. These findings are certainly compatible with a model in which representations of perceived objects are stored in prefrontal WM, and then replayed, so they can be stored more permanently in the hippocampus. (This idea is further supported by evidence that the brain uses the same codes to represent the SM properties of objects in WM and LTM; see Brady *et al.*, 2011 for a review.)

In summary, there is evidence that objects and their properties are stored in WM in temporary assemblies of PFC neurons that can be transiently reactivated, in a manner that reactivates their associated SM representations (and possibly associated representations in the hippocampus). But there are several questions that remain. Firstly, what prefrontal mechanism is *responsible* for the transient reactivation of these temporary object representations in PFC? Secondly, it should be recalled that alongside transiently active WM assemblies in PFC, there are also PFC assemblies that encode object properties in *sustained* activity over the whole delay period. (Recent studies confirm there are two groups of PFC neurons that encode remembered object properties, one with sustained activity over the delay period and one with variable activity; for instance, this is shown in macaque by Mendoza-Halliday *et al.*, 201, and in rats by Baeg *et al.*, 2003.) If convergence-zone units can store associations between the properties of objects *without* being tonically active, what role is played by the tonically active object representations in PFC? Do they perhaps have a role in reactivating the transient object representations? Finally, there is nothing in the simple model in Figure 1 to explain why in some circumstances the capacity of WM is defined in units of 'whole objects' rather than in informational units. There are some neural network models that obtain object-based effects without any explicit notion of 'slots' for storing object representations (see in particular Wei *et al.*, 2012). However, given that semantic WM must store not only individuals, but the episodes they participate in, there are other reasons to consider a slot-based model, as we will now discuss.

2.2. WM representations of episodes

In Baddeley's model, the semantic WM system is able to hold a representation of a complete episode, involving multiple participants. WM representations of episodes are higher-level objects than the WM representations of individuals that we discussed in the previous section, and consequently harder to study. However, there is good evidence that episodes are a real unit of cognitive organisation: for instance, there is evidence that specialised brain networks become active at the boundaries between experienced episodes (Zacks *et al.*, 2001; 2007; Sridharan *et al.*, 2007). In the current paper we will make four assumptions that help to cast light on how episode-sized chunks of experience are represented.

Firstly, we assume that WM episode representations include, or make reference to, WM representations of individuals. This is a fairly uncontroversial assumption, but it is of real value in developing a model of WM episode representations: it means that any model of WM individuals is also *a component* of a model of WM episodes. This fact places quite strong constraints on both models. For instance, if we know that WM individuals are stored in a circuit encompassing certain prefrontal and primary SM areas, we can infer that this same circuit also participates in the mechanism that stores episodes in WM. Furthermore, the mechanism that associates individuals with the semantic roles they play in an episode (such as AGENT and PATIENT) must operate on the representations of individuals provided by the model of WM individuals. Conversely, the model of WM individuals can assume the existence of an episode-storing mechanism that can represent multiple individuals simultaneously. As already mentioned, one possibility is that the 'multiple slots' proposed in Luck and Vogel's (1997) model of WM individuals could be slots *in a WM episode representation*, that hold the participants in a WM episode.

A second assumption is that the neural machinery that represents an episode in WM should allow this episode to be actively *rehearsed*. That is, it should be able to recreate in relevant ways the patterns of activity that occured when the episode was experienced. This capability is an essential one if WM episodes are to serve as the interface between SM experience and LTM representations in the hippocampus: in Baddeley's model, WM mechanisms must be able to reactivate episode representations at a timescale appropriate for hippocampal learning. Moreover, as discussed in Section 2.1, there is good evidence that WM representations of individual objects can be actively rehearsed. If the WM individuals system is a component of the WM episodes system, as just argued, then we already know that some components of WM episodes can be rehearsed. But there is still a question as to what it means to rehearse 'the experience of a whole episode'.

A third assumption relates to how episodes are experienced. This is an assumption about the SM processes from which WM representations of episodes are built, rather than about the WM system as such. We assume that episodes are experienced through SM routines with well-defined *sequential structure*. It is uncontroversial that experiencing an episode takes time, and there is widespread acknowledgement that neural representations of episodes must encode relationships between temporally discontiguous events (see e.g. Eichenbaum, 2006). The idea that cognitive representations of episodes exploit their sequential organisation already features in some models of episode representation, in particular that of Zacks *et al.* (2007) and Kemmerer (2012). Our proposal is that episodes have a component of *discrete* sequential structure, originating from the SM operations through which they are experienced. Our background assumption is that SM experience is organised into sequentially structured SM *routines*, whose atomic elements are relatively discrete attentional, perceptual or motor operations (see Ballard *et al.*, 1997). We have studied the SM processes involved in experiencing simple episodes involving transitive actions, such as a person grabbing a cup (Webb *et al.*, 2010; Knott, 2012), and causative actions, such as a person opening a door (Lee-Hand and Knott, 2015), and found good evidence for such routines. There is strong evidence that the participants in such episodes are attended to one by one, whether the observer is executing the action him/herself or watching another agent executing the action. Moreover, there is evidence from several experimental paradigms that the observer attends to the AGENT of the episode before the PATIENT, and only activates a representation of the action category when both the agent and patient have been attended to (for a review of this evidence, see Knott, 2012; 2014a). This evidence suggests that the attentional actions through which the AGENT and PATIENT of an experienced actions are identified may be associated with particular *serial positions* in a SM routine—specifically, that the first object attended to is the AGENT, and the second object attended to is the PATIENT. This idea has certainly been proposed in linguistic models of semantic roles. In particular, Dowty (1991) proposes a model in which there are only two semantic roles: 'proto-agent' (the participant that is attentionally most salient) and 'proto-patient' (the participant that is attentionally less salient). In our view, this model is strongly supported by studies of episode perception.

If episodes are experienced in canonically structured SM routines, there are several implications for a model of WM storage of episodes. For one thing, a commonality is highlighted between WM representations of episodes and WM representations of individuals: both representations encode *sequences* of SM stimuli. (Recall from Section 2.1, that WM representations of individual objects maintain parallel representations of location and category, which become active sequentially during object perception.) On this view, a WM episode representation is essentially, a *prepared sequence* of SM operations: given that a WM episode serves to guide the agent's behaviour in generating actions or monitoring incoming episodes, it must be able to reactivate the sequentially encountered components of the episode in real time, in the order they were initially encountered. At the same time, however, the temporally distinct components of an experienced episode should be stored in a format that allows them to be presented simultaneously, or near-simultaneously, to the hippocampus, so the episode can be encoded in LTM.

If an episode is held in WM as a prepared sequence of SM operations, this is useful from a methodological point of view, as it places them within a category of neural representation that has been extensively studied in monkeys. There is very good evidence in macaque that prepared sequences of attentional and/or motor operations are held in assemblies in dorsolateral PFC (see e.g. Barone and Joseph, 1989; Shima *et al.*, 2007). Of particular interest is the finding that the component operations in a prepared sequence are active *in parallel* in these assemblies, even though they are executed sequentially (Averbeck *et al.*, 2002; Averbeck and Lee, 2007). This gives them exactly the properties that are needed to serve as WM episode representations, on the above model: they support sequential replay of experiences, but they also support simultaneous transfer of episode representations to the hippocampus. It also further emphasises the commonalities between WM episodes and WM individuals: in both cases, items that occur sequentially in SM experience are stored in parallel in WM.

A final assumption we will make is that the mechanism that represents episodes in WM interfaces with sentence processing mechanisms. Specifically, we assume it represents (in a hearer) the meaning of a sentence that has just been interpreted, and (in a speaker) the meaning of a sentence which is about to be produced. These assumptions mean that WM episode representations can be studied in sentence processing experiments. The idea that sentence processing experiments provide a means for studying semantic WM representations has surfaced in various forms in the literature: for instance in Potter and Lombardi's (1990) hypothesis that short-term recall of a sentence is mostly based on its meaning, and in the more nuanced claim of Schweppe *et al.* (2011) that such recall engages both phonological and semantic WM processes. Of course sentences also have *syntactic* structure, and there is good evidence that elements of syntactic structure are also stored in WM (see e.g. Potter and Lombardi, 1998; Schweppe and Brunner, 2007). Since some components of syntactic structure are closely linked to semantic structure, it is possible that semantic WM representations also encode some aspects of syntactic structure. We will make two specific proposals, that link models of syntax developed by linguists with an account of semantic WM. Firstly, we will propose that the syntactic relationship between a sentence and its constituent noun phrases (NPs) reflects the relationship between a WM episode representation and its component WM individuals (at least for referential NPs). Secondly, we will propose that the notion of the 'syntactic heads' of NPs and clauses derives from an aspect of the structure of WM episode and individual representations.

2.3. WM representations of situations

In Baddeley's model, semantic WM can hold a representation of a single episode, but it can also represent the 'context' that an episode occurs in: for instance, the current state of play in a game, or the gist of a recently read paragraph of text. A context has both spatial and temporal components. We have already discussed one aspect of the spatial component: as discussed in Section 2.1, there is a WM medium that holds representations of recently-encountered individuals, in a convergence zone that links spatial locations to object properties. This convergence zone can be thought of as holding a representation of the current spatial context. But we still need to consider the temporal component: that is, how WM represents 'the current moment'. We will refer to this representation as a representation of the current *situation*.

A representation of the current situation is even more derived than a representation of the current episode, and thus even harder to investigate empirically. But here again we can make some helpful cross-disciplinary assumptions that provide constraints for a WM model.

Firstly, we assume that the situation representation must support decisions in the motor system about what to do next, and expectations in the perceptual system about what will happen next. Generalising over these cases, we assume that a WM representation of 'the current situation' provides top-down input into the SM processes that determine *what episode is experienced next*, whether the episode involves an action of the agent or of some external individual. In other words, it holds 'prospective' information (guesses or plans) about the future. There is widespread agreement that PFC is important in encoding the 'task' or 'cognitive set' of an agent (see e.g. Warden and Miller 2010 for recent evidence), but the impact of the current cognitive set is also seen in modulation of activity in primary SM areas (see Miller and Cohen, 2001). At issue is how this cognitive set is stored. One interesting finding is that PFC can represent several alternative possible episodes simultaneously. For instance, Averbeck *et al.* (2006) show that if a monkey is choosing between two alternative action sequences, both sequences are represented simultaneously in the activity of dorsolateral PFC cells. Representations of multiple alternatives can also be seen in lower-level SM cortices: for instance, in the parietal and premotor areas that select and control reach actions to target objects, there is evidence that multiple alternative targets are represented simultaneously at an early stage of processing, with the selected target gradually becoming dominant over time (see e.g. Shadlen and Newsome, 2001; Cisek and Kalaska, 2005). A common proposal in recent Bayesian models of neural function is that patterns of neural activity can represent not only specific actions or perceptual judgements, but also probability distributions over *possible* actions or perceptual judgements (see e.g. Kiani and Shadlen, 2009; Pouget *et al.*, 2013). It is attractive to think of situation representations in this way, as distributions over all possible atomic states or events (see e.g. Frank *et al.*, 2009). In our model, in line with this idea, PFC holds a WM representation of 'the current situation', which induces a *probability distribution over possible episodes*. This representation provides top-down guidance to the process of experiencing an episode, in real time, that can be seen in dynamic changes to distributions over possible actions, locations and object properties in lower-level SM cortices. (Note that the convergence zone linking object locations to object properties can be thought of as holding a distribution over possible individuals in the current context, so on this proposal, both the spatial and temporal components of the current context can be modelled as distributions.)

A second assumption is that the WM representation of the current situation must also support *update* operations. There are various ways in which an agent's cognitive set can be changed. When unexpected circumstances arise, agents may have to change their plans quite dramatically (see e.g. Braver and Cohen, 2000), but updates can also occur more gradually and/or predictably during the performance of a particular task, to register a task-relevant stimulus (Warden and Miller, 2007; Mante *et al.*, 2013) or to mark a particular component of a complex task as completed (Cooper and Shallice, 2000). In either case, the update is a function of the episode that has just been experienced. There is good evidence that update operations are implemented in *recurrent* circuits in PFC (Mante, 2013), that extend into subcortical areas in the striatum (see e.g. Dominey and Bossaoud, 1997). The dopaminergic neurons in these striatal areas appear to have a role in enabling updates to task set (see D'Ardenne *et al.*, 2012 for recent evidence, and O'Reilly and Frank, 2006 for a computational model). Computationally, the update operation is a function that takes a representation of 'the current situation', plus a representation of the episode that has just been experienced, and returns an updated situation. What this means is that the 'current situation' can reflect not just the most recent episode, but a *sequence* of recent episodes: that is, it holds 'retrospective' information about recently-experienced episodes as well as 'prospective' information about upcoming episodes. We assume this is why it is able to hold a representation of the 'gist' of a multi-sentence paragraph of text.

A third assumption is the representation of the 'current situation' that is maintained in WM communicates with an analogous representation in LTM. There is good evidence for representations of spatial and temporal context in LTM, particularly in parahippocampal areas (see e.g. Epstein *et al.*, 2007; Turk-Browne *et al.*, 2012): these are important for individuating token episodes stored in memory (Tulving, 1983). Episodic LTM shows strong sequential organisation, and is often modelled using a recurrent circuit, in which representations of individual episodes can update a representation of spatial and/or temporal context (see e.g. Lisman, 1999; Howard *et al.* and Kahana, 2002). There is a growing consensus that representations of context in WM are synchronised with those in hippocampal LTM, particularly those representations that support the encoding of sequential stimuli (for a review, see Burgess and Hitch, 2005).

A final assumption is that the WM situation representations that influence motor behaviour and perceptual expectations and communicate with episodic LTM are also involved in language processing. A representation of 'the current situation' is of great importance in language processing. Most obviously, such a representation exerts a strong influence on sentence intepretation, in helping to choose between alternative possible meanings (see e.g. Pickering and Traxler, 1998; Traxler, 2014). If the current situation induces a probability distribution over possible episodes, it can directly inform this disambiguation process, in much the same way it helps decide between alternative sensory or motor operations during direct experience of the world. In fact, even setting contextual influences to one side, the basic process of *parsing* a sentence requires the enumeration of a very large number of alternative possible structures and associated interpretations. The best-performing computational models of sentence parsing rely heavily on a 'chart' data structure that represents a large probability distribution over possible interpretations (see e.g. Collins, 1996); it is likely the brain requires a data structure of this kind too (see e.g. Charniak, 2011). If a WM situation representation induces a probability distribution over possible episodes, it could also perhaps serve as the chart that represents alternative possible parses of an incoming sentence. On this model, the WM situation representation would have a role both in enumerating alternative possible sentence structures/interpretations, and in weighting these alternatives in the light of the current context (see Traxler, 2014 for evidence connecting these two roles).

Another important role for the 'current context' representation in language is to hold a set of salient discourse entities. Indefinite noun phrases (e.g. *a dog*) introduce entities into the discourse, while definite noun phrases (e.g. *the dog*) and pronouns (e.g. *it*) presuppose entities with particular properties, that have recently been introduced or become salient (Kamp, 1981; Heim, 1982). The WM medium that stores visuospatial representations of recently-encountered objects (see Section 2.1) also has a role in representing recently-encountered individuals. There is some evidence that this medium is also involved in representing salient discourse referents: for instance Wallentin *et al.* (2008) show that subjects processing a sentence introducing discourse referents create spatial representations in parietal cortex, which can be reactivated by pronouns in a subsequent sentence. However, there is also evidence that discourse referents and real objects are represented in distinct neural media (see e.g. Hickok and Bellugi, 2010), so if visuospatial WM has any direct role in language processing, it is likely to be supplemented with other more language-specific representations. Nonetheless, there is some evidence that nonlinguistic representations of the current spatial context also play a role in a linguistic account of reference.

2.4. Summary

In the model outlined above, the semantic WM system comprises five media. One represents a single object, or a single homogeneous group of objects, specifying its location and intrinsic properties, or a distribution over likely locations and properties. We will call this medium the **WM individual** medium. (We intend the term 'individual' to cover both a single object and a homogeneous group of objects of a single type.) A second medium is a convergence zone whose units hold short-term associations between the properties and locations of recently-encountered individuals. We will call this medium the **candidate WM individuals** medium, since it represents the possible individuals that can be attended to at the current time. This medium encodes an aspect of the current 'spatial context'. A third medium holds a representation of a single whole episode. This could be an episode that has just been experienced through SM interaction with the world, or an episode featuring an action that the agent is preparing to execute. We will call this medium the **WM episode** medium. We argue that the 'slots' that allow storage of three or four separate individuals in WM are components of the WM episode medium, whose primary purpose is to represent the multiple participants in an episode. The fourth medium holds a probability distribution over possible episodes, which guides online processing of new episodes. We will call this medium the **candidate WM episodes** medium. The distribution in this medium is generated by the representation in the final medium, the **current situation**. The current situation is updated after each episode that is experienced. In our model, WM individuals and WM episodes are represented in PFC as *prepared sequences* of SM operations: one of their roles is to support the active replay of the sequentially structured SM processes through which individuals and episodes are perceived. At the same time, they hold tonically active, parallel representations of their component elements: a format that supports the transmission of individual/episode representations to longer-term storage in the hippocampus. In our account, all five WM media interface with LTM representations. WM representations of individuals use the same 'codes' as LTM representations of individuals, and interface with LTM representations of spatial environments; WM representations of single episodes can be replayed to hippocampal LTM, and WM representations of whole situations are mapped to sparser, more localist representations of context in hippocampal/parahippocampal LTM. We also propose that replay operations provide a way for WM representations to interface with language. Specifically, we propose that producing a clause involves replaying a WM episode, and producing a noun phrase involves replaying a WM individual.

In the rest of the paper, we will present a neural network model of representations in these five media, and how they interact with each other, and with LTM and language. We introduce the main new ideas in the model in Section 3, and present the model in detail in Section 4; in Section 5 we describe how the network is trained and tested, and evaluate its performance.

3. Technical innovations in the network

There are two key technical challenges in implementing a WM model with the properties just outlined. We will describe these in turn, along with our proposed solutions.

3.1. Binding WM individuals to semantic roles

One challenge is that the WM episode representation must make reference to multiple distinct individals, and associate each of these with a particular semantic role. If there is a single medium within which individuals are represented, it is difficult to represent multiple individuals without blending their properties. If the individuals have to be associated with different semantic roles, it is also hard to assign specific roles to specific individuals. Several solutions have been proposed to this problem. We will focus on the two best-known solutions. In the model of van der Velde and de Kamps (2006), a collection of special-purpose binding units allow the reification of binding relationships in short-term synaptic weights: each individual participating in an episode can be linked to its own binding unit, which is in turn linked to a particular semantic role. In the model of Stewart and Eliasmith (2012), multiple bindings are made possible by the use of high-dimensional vector representations that permit the results of binding operations to be represented in the same vector space that represents the elements being bound. However, if episodes are experienced in canonically structured SM sequences, and are stored in WM as prepared SM sequences, as we propose, there are some alternative ways of solving the binding problem that are worth considering. In this scenario, the key point is that representations of the individuals participating in an episode are guaranteed to be activated *one by one* during experience—and in a particular order, whereby the agent is activated first, and then the patient.[1] The process that builds a WM episode representation can take advantage of this ordering constraint.

We envisage that during experience of an episode involving two participants, the agent activates two WM individual representations in the succession in the same medium: first one representing the agent, and then one representing the patient. Our first key proposal is that these two representations are *copied to different slots* in the medium representing the current WM episode: the first representation is copied to a slot representing the agent of the episode; the second is copied to a slot representing the patient. In the WM episode representation, therefore, the agent and patient are coded 'by place': agents and patients occupy separate areas. Using place-coding to bind individuals to semantic roles is normally dismissed as something of a straw man: in a simple place-coding scheme, the representation of a man as an agent has nothing in common with the representation of a man as a patient, and this introduces many problems: for a discussion see Chang (2002). But if slots in a WM episode just hold *copies* of representations in the (single) medium representing WM individuals, many of these problems go away: the place-coded slots in a WM episode representation just hold *pointers* to representations of WM individuals, which are defined one by one during experience, and activated one by one during simulation or replay.[2] While the slots representing the agent and patient in the WM episode hold pointers *to the same medium*, there is no confusion because the pointers are initialised at different times during experience of an episode, and dereferenced (i.e. followed) at different times during the replay of a WM episode.

Note that this model of role-binding makes reasonably high demands on storage space: the slots representing agent and patient in the WM episode medium must each hold a copy of the WM individuals medium. However, the copy need not be complete: in our model, for instance, as we will discuss in Section 4, we do not copy information about the location of individuals, only information about their type and properties, which limits the combinatorial possibilities. Moreover, the number of copies is quite small: we only need as many copies as there are possible participant roles within a single episode, which if we use verb subcategorisation as a guide, appears to be four (see e.g. Traxler, 2011).

Note also that the above proposal only addresses one specific binding problem in semantic representations, namely how semantic roles in episodes are bound to representations of individuals. The binding schemes of Stewart and Eliasmith and of van der Velde and de Kamps are much more

[1] Or the proto-agent and then the proto-patient, to use the terminology of Dowty (1991). We will continue to use the terms 'agent' and 'patient', for brevity's sake.

[2] The concept of 'pointers' is also used by Eliasmith 2013 to refer to the vector-based representations of Stewart and Eliasmith, in the sense that they are pointers to SM experiences. Our WM representations are also pointers in this sense; however the 'copy' operations we envisage create pointers that have a particular technical role within our model.

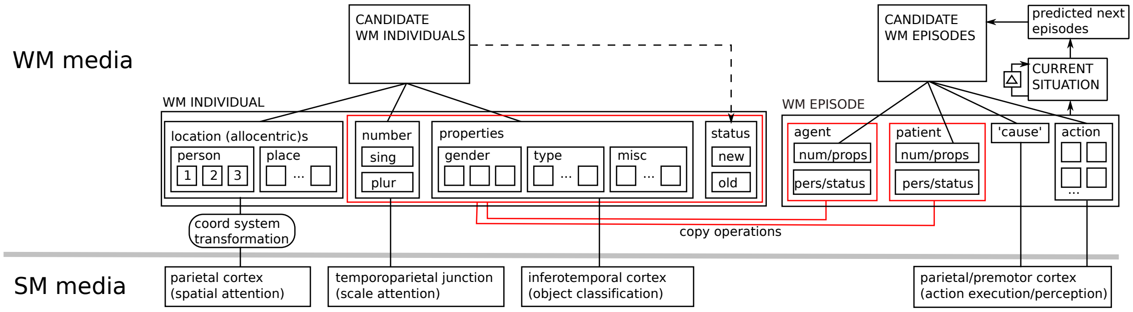


Figure 2: Architecture of the model of WM individuals and WM episodes. The connection annotated with the 'delta' (Δ) symbol holds a stored copy of the situation at the previous time point.

general than this: they also cover nested clauses (e.g. in relative clauses and complement clauses), and they allow representation of arbitrarily deeply nested structures. In fact our model makes proposals about nested clauses too, but we will defer discussion of these until Section 8. In the meantime, our model does permit a minimal model of hierarchically nested structures, in that the representations that are bound to semantic roles (WM individuals) themselves have some internal structure. In the account we develop in Section 4, the representations of the agent and the patient that are activated one by one when an episode-denoting sequence plan is rehearsed are also sequence plans, that can be rehearsed in their own right, so that individual steps within the episode-denoting sequence can have their own sequential structure. In fact, we propose that this structure of sequences-within-sequences supports an interesting model of the syntactic relationship between NPs and their host clauses, as we will discuss in Section 6.

3.2. Representations of probability distributions over episodes

The other challenge for our model of WM is to represent a large distribution of *possible episodes* using a neural network. The binding schemes developed by van der Velde and de Kamps (2006) and Stewart and Eliasmith (2012) allow the representation of a small number of episodes simultaneously, but have not been shown to support the simultaneous representation of large numbers of episodes. Using place-coded representations of agent and patient in the WM episode medium offers an important advantage here, as it allows very simple *localist* representations of whole episodes: we can define localist units that encode particular combinations of agent, patient and action type, and we are free to envisage a large collection of these localist units, that represent many different possible episodes or episode types. A pattern of activation over these localist units can straightforwardly encode a probability distribution over possible episodes, and these distributions have many applications, as we will discuss in Section 4. Note that while these localist episode representations encode structures that *point* to WM individuals, rather than to WM individuals themselves, these pointers can be dereferenced in sensible ways during SM experience, to allow SM processing to be modulated by expectations about WM episodes, as we will demonstrate in Section 5.

4. Architecture of the model

Our model is illustrated in Figure 2. The WM media are above the thick grey line; SM media are below it. WM media representing individuals are on the left, and WM media representing episodes are on the right.

The **WM individual** medium holds a representation of a single selected individual. It stores the sequence of SM operations through which a single object, or a homogeneous group of objects, is established. There are three operations in the sequence. The first operation activates a spatial **location**. While locations are initially perceived in an egocentric frame of reference, we assume the location stored in WM is centred on the observer's local environment, so it is invariant across the observer's attentional shifts: in the prevailing model, this is a location in the hippocampal 'cognitive map' (Schmidt *et al.*, 2007). The activated location can be an arbitrary place in the observer's current environment, but it can also be the location of the observer himself, or of his interlocutor, which can both be defined in the environment-centred frame of reference. The selected location therefore sets a **person** field, to either 1 (self), 2 (interlocutor) or 3 (external individual). In each case, a **place** field is also set, indicating the location in a map of the current environment. The second operation selects a **number**, which can be singular or plural. Our account of this operation rests on the model of Walles *et al.* (2014; 2008), in which attention can be allocated either to the global form or the local form of a selected stimulus (in the sense of Navon, 1977). In the former case, the object classifier is configured to identify a single object at the attended location; in the latter case, it is configured to classify a homogeneous group of objects, and return the type these objects collectively share. The final operation identifies the **properties** of the attended object or group, which can include an open-class object type, but also other information, including semantic information that is picked up in grammatical gender, and miscellaneous properties that set the object/group apart from others of its type.

When the fields of a WM individual are fully defined, a rehearsal operation is enabled, that replays the sequence of SM operations through which the individual was established. This involves transient activation firstly of a spatial location (in parietal cortex), then of a spatial scale (in the temporoparietal junction, see e.g. Robertson *et al.*, 1988; Fink *et al.*, 1996), then of a type and associated properties (in inferotemporal cortex). Recall from Section 2.1 that there is strong evidence for this kind of transient reactivation of SM representations during a delay period. While the replay process activates SM representations sequentially, the fields of the WM individual stay active tonically, in parallel, in line with evidence about the prefrontal assemblies that store prepared SM sequences in monkeys, as discussed in Section 2.2.

The layers representing a WM individual provide input to another layer, the **candidate WM individuals (cWM-ind)** layer, which stores associations between the location, number and properties of attended individuals over a short interval, and thus comes to represent a collection of individuals that have recently been attended to. This is our model of the convergence zone that was sketched in Figure 1. A partially specified WM individual can function as a query to the cWM-ind layer: if we specify a location, we may be able to retrieve an associated number and set of properties (and vice versa). If an individual is retrieved from the cWM-ind layer, it is classed as 'old'; if not, it is classed as 'new'. These attributes are recorded in the **status** field of the WM individual, which is not part of the prepared sequence. We envisage both linguistic and nonlinguistic roles for the cWM-ind layer. Linguistically it can represent the set of salient referents in an ongoing discourse. Nonlinguistically it can hold expectations about the location and identity of objects in the current scene.

The WM media representing episodes are structurally similar to those representing individuals. The **WM episode** medium holds a representation of a single selected episode, stored as a planned sequence of operations. The first operation activates a representation of the **agent** of the episode. The second operation activates a representation of the **patient** of the episode (if there is one). The remaining operations activate a representation of the action that occurs. This can be causative or noncausative; in the former case, a dedicated network for controlling causative actions is activated before the action proper is represented (see Lee-Hand and Knott, 2015 for details of this proposal, which are not relevant to the current model). Again these planning representations are content-addressed pointers to operations in other media: they are active in parallel in the planning medium, but when the WM episode is executed or rehearsed, the representations they point to become active one a time. A key idea is that the 'agent' and 'patient' media contain pointers *to WM individuals* rather than directly to SM signals. These pointers are created when the episode is experienced. The first WM individual activated during experience of an episode is copied to the 'agent' medium of the WM episode, and later, the second WM individual to be activated is copied to the 'patient' medium. These copy operations are shown in red in Figure 2. (In fact, as the figure shows, we do not copy all the fields of a WM individual into slots in the WM episode, only information about number and properties. In this way, the WM system for individuals specialises in encoding locations of individuals, while the WM system for episodes abstracts away from information about location.) When all the fields in a WM episode have been filled, the episode can be rehearsed, just like a WM individual. In this process, the WM individuals representing the agent and patient become active in sequence, one at a time, creating temporally separate opportunities to rehearse their own associated sub-sequences.

As noted in Sections 1 and 2, our WM model is intended to contribute to a model of syntax as well as to an account of SM processing. The linguistic roles played by the model will be discussed in Section 6, but for the moment it is useful to preview that the WM individual and the WM episode media both carry information that can be conveyed syntactically by **heads** in a syntactic structure. Referring to Figure 2, a WM individual carries information about the **person**, **number**, **gender** and **type** of the individual in question: all this information is conveyed by heads in a noun phrase, as we will discuss in Section 6. A WM episode carries copies of all this information for both the agent and patient, as well as information about the open-class type of the action: this information is conveyed by heads in a clause, as we will again discuss in Section 6.

The fields in a WM episode also provide input to a layer that represents a probability distribution over possible episodes, the **candidate WM episodes (cWM-ep)** layer. This layer is a self-organising map or **SOM** (Kohonen, 1982): when exposed to training episodes, it learns to represent episodes as localist units, organised so that similar episodes are close together in the map. Each localist unit can encode a particular combination of representations in the agent, patient and action media, and thus can represent a complete episode by itself. As already noted in Section 3.2, this is only possible because of our use of place-coded representations of agent and patient in WM episodes. The 'agent' and 'patient' fields of a WM episode index their fillers *by content* rather than just by a formal role label, so these place-coded representations carry information about both roles *and their fillers*.

A final component of the network is a medium representing the **current situation** connected to a layer trained to predict next episodes. The current situation is implemented in a type of recurrent SOM, that takes input from the current WM episode, and also from a representation of its own activity after the previous episode was experienced.[3] This network learns to represent 'the current situation' as a function of the episode sequences most commonly encountered during training, so that similar episode sequences activate similar situations. The **predicted next episodes** medium is trained to map the currently active situation onto a representation of the episode that actually occurs next in the cWM-ep layer, through supervised learning. Because the next episodes layer is isomorphic to the cWM-ep medium, after training, a representation of the current situation predicts a distribution of likely next episodes in the cWM-ep layer.

Note that as the cWM-ep layer and the current situation layer are SOMs, they use localist units to represent episodes and situations respectively. This has some important implications. For one thing, it means that the model cannot represent every possible episode, or every possible situation, within these media: there are too many of these in any realistic scenario. But that is not the purpose of these SOMs: their role is to store the set of episodes and situations that are *actually encountered*, and they are configured to do this as efficiently as possible. In particular, as we will show in Section 5, they are able to learn *generalisations* over encountered episodes and episode sequences where capacity is limited. This ability is very important for a model of WM, as it allows the agent to generate sensible expectations in the face of unseen episodes. At the same time, there is a medium in the model that can accurately store every possible episode, including unseen ones: this is the WM episode medium, as we will again demonstrate in Section 5.

A second positive implication of using SOMs is that the network can represent multiple possible episodes, and multiple possible situations, simultaneously. In particular, as we will show in Section 5, the network can learn to represent a *probability distribution* over likely next episodes in the cWM-ep SOM. This is extremely useful, in many ways. In nonlinguistic processing, this can be used to select the most likely or most desired episode, or more subtly, to generate distributions over expected properties and locations in the WM individuals medium as experience of an episode is under way. In linguistic processing, the distribution can play a useful role in sentence interpretation, providing information that can help decide between alternative interpretations of incoming ambiguous sentences, as proposed in Section 2.3.

[3] For efficiency, this is stored using the *weights* of the unit representing the previous situation.

We now give some technical details about the network's architecture. The WM individual layer (69 units) consists of localist sets of feature units for person (1, 2, 3), number (Sg, Pl), gender (Male, Female, Neutral) and status (new, old). Each set of units can either encode a single property unambiguously, or a probability distribution over properties. The type area also contains feature sets coding animacy, particular object type (human, dog, cat, bird, cup, ball, chair) and several properties introducing similarity relations among types (is-human, can-fly, can-be-grabbed), see Table A.1. Location of objects (situated on a 100×100 grid) is coded by a population of 6×6 neurons (the 'place' area on Figure 2) with Gaussian receptive fields evenly covering the grid. The 'misc' area represents a single property, colour, whose value is defined stochastically according to object type. Colour is coded by a population of 11 neurons with Gaussian receptive fields in 3D RGB space, responding maximally to 11 basic colours (see Figure 4a). Such population coding is neurally plausible and there is a straightforward mathematical way of computing the likelihoods of different stimuli given the activities of neurons in the population (Jazayeri and Movshon, 2006). First-order SM representations (below the thick grey line in Figure 2) are isomorphic to the WM individual and WM episode areas they are linked to. Likewise, agent and patient layers of a WM episode are isomorphic to the relevant parts of a WM individual: the num/props areas are isomorphic to the number and properties areas of the WM individual, and the pers/status areas are isomorphic to person and status areas of the WM individual. 'Cause' is a single neuron that is either on or off. The 'action' area consists of 22 localist units for actions and 11 units for their distributed featural codes (see Table A.2). The cWM-ind layer is a variable-sized convergence zone of units fully connected with the WM individual layer: when a novel candidate individual is encountered, a new unit in the cWM-ind layer is recruited and the current values of WM individual units are copied into its connection weights (one-shot learning). The cWM-ep layer is a SOM with 400 units. Each unit has a 90-dimensional vector of incoming weights from the WM episode medium and a scalar weight reflecting the relative frequency of 'hits' for this unit, i.e. the proportion of times this unit was the most active unit. These frequency weights serve as priors for computing the Bayesian probability that the current input corresponds to an episode represented by a particular unit (see Appendix A.2 for details). The current situation medium is implemented as an MSOM (Strickert and Hammer, 2005)—a type of SOM enhanced with recurrent connections. Units representing similar episodes happening in similar contexts are close to each other on the map. The activity of the whole map can be interpreted as a probability distribution over situations (episodes in contexts). The MSOM has 400 units. These form the input to the predicted next episodes medium implemented as a single layer of 400 perceptrons with linear activation functions and the softmax function applied to their outputs. The output layer is isomorphic to the cWM-ep medium and represents a distribution of possible next episodes.

5. Training and testing of the network

5.1. Training

Our system is exposed to a continuous stream of 20000 episodes[4] generated stochastically from transcription rules (see Table A.3). Once an episode is generated, token individuals are generated for each type participating in the episode. The system can encounter novel individuals (with probability 0.01), but also re-encounter some of the recently seen individuals of the same type (if available). Each individual has a type, a number, a colour and a location. Colour is stochastically chosen from Gaussian distributions centred on 11 basic colours; location is selected randomly from positions on a 100×100 grid, represented using coarse coding in the system's 6×6 location medium. Each episode is presented to the WM system as a sequence of input items. Episodes are of three types: **transitive** (agent→patient→trans-action), **intransitive** (agent→intrans-action) and **causative** (agent→patient→cause-signal→unergative-action). In each case the agent and patient signals have a sequential structure of their own, namely location→number→type/properties. Each of these latter sequences is sent to the WM individuals medium, activating the different components

[4] For measuring purposes, we divided this stream into 40 epochs of 500 training episodes each.

of a WM individual representation one by one. When complete, the WM individual is first passed as a query to the cWM-ind layer, to find out whether the individual it represents has recently been encountered. For each candidate unit currently active in the cWM-ind layer, we compute the likelihood that it corresponds to the current stimulus in the WM individual (Jazayeri and Movshon, 2006). This reduces to the average pairwise KL divergence (Kullback and Leibler, 1951) between the respective areas of the WM individual and the candidate unit weights (see Appendix A.4). If a likely-enough candidate is returned, it is updated if necessary and the WM individual's status is set to 'old'; otherwise a new entry in the layer is created and the WM individual's status is set to 'new' (candidate units that have not been updated for N episodes are removed). The WM individual is then copied (along with its status) to the appropriate layer in the WM episode medium: either the 'agent' layer or the 'patient' layer. When a complete episode has been presented to the system, the layers in the WM episode are passed as input to the cWM-ep SOM. This SOM learns in the standard way (Kohonen, 1982). The current situation MSOM is also trained in a standard way (Strickert and Hammer, 2005). The weights of connections to the predicted next episodes layer are adapted using a standard Delta rule (Rosenblatt, 1962). In order to learn to predict the next episode, the system must be trained at the moment when the Next-ep's output layer contains the prediction from the previous episode, but the cWM-ep already represents the actual next episode. This can be achieved by the following sequence of operations:

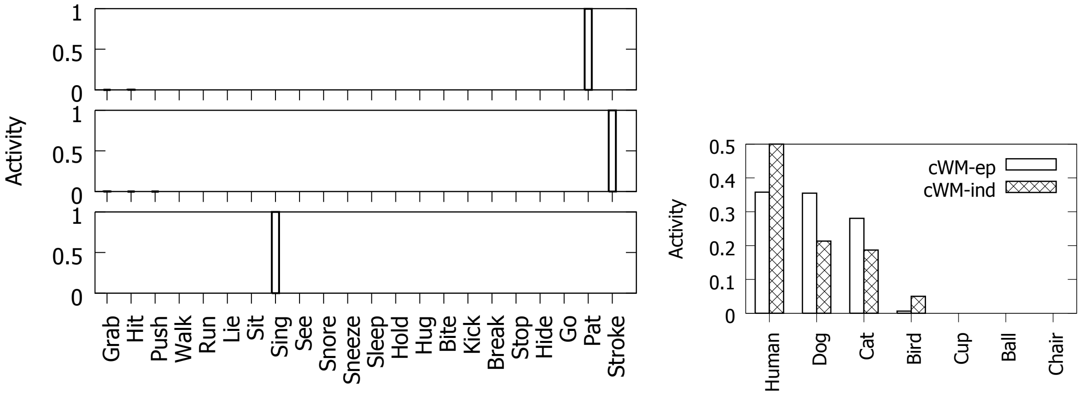
1. **Update the current situation** by propagating the current content of the WM episode through links to the Current situation.
2. **Predict the next episode** by propagating through links from the Current situation to the Next-ep medium.
3. **Perceive the next episode** by filling the WM episode and propagating through links to the cWM-ep medium.
4. **Train the Next-ep system** by the error signal defined by difference between the activity pattern in cWM-ep (the target) and the content of the Next-ep.
5. Continue from Step 1.

Note that while learning in the cWM-ind layer happens in a 'one-shot' manner, it only happens gradually in the cWM-ep SOM, current situation MSOM, and the predicted next episodes layer. For more technical details, including the values of learning parameters see Takac and Knott (2015b).

5.2. Testing the sequence-based binding scheme

To demonstrate the new binding scheme, we must show how the WM representations created during experience of an episode allow it to be *replayed.* To test this, after each episode is presented, the WM episode medium is used as input to a replay process, in which the layers in this medium activate the representations they point to one by one. Whenever a representation is activated in the WM individuals medium, an analogous replay routine is executed in this medium. If the binding scheme is effective, we should recover the same sequence of first-order SM signals that were presented to the network during experience of the episode. In our tests, the sequence was perfectly reconstructed for 99.7–100% of episodes across all training epochs; this demonstrates that our proposed binding mechanism is effective.

It is important that the network is able to encode and replay episodes that it has not seen before. This is a problem for some models of episode representation; for instance, models that encode episodes as sequences using simple recurrent networks (SRNs, Elman, 1990) have difficulty representing episodes involving unseen sequences. (This is the case, for instance, for the early model of McClelland *et al.*, 1989, and for more recent models such as Sutskever *et al.*, 2014). In our model, the WM episode medium should be able to encode unseen episodes just as well as seen ones, since all WM episodes are constructed using the same general procedure. To confirm this, we tested the trained network's ability to encode and replay 100 episodes that were not encountered during training. 99.5–100% of these unseen episodes were perfectly reconstructed, indicating that the WM episode network can effectively represent unseen episodes, and allow them to be replayed.



(a)

(b)

Figure 3: (a) Action types predicted in the cWM-ep layer for 3 episode fragments. From top to bottom: man→dog→?, man→cat→?, bird→?. (b) Prior expectations on the agent type generated top-down by the cWM-ep and cWM-ind layers.

5.3. Testing the network's prediction/generalisation abilities

The network can make several kinds of prediction; we will focus on three progressively more complex predictions. To begin with, the cWM-ep SOM can make predictions about the episodes that are likely to occur, which are refined as an episode is experienced. Its predictions about actions are easiest to demonstrate, since it represents actions directly. To test the accuracy of these predictions, we introduced some regularities into the episodes that were presented to the system. Birds always sang (bird→sing); also when people interacted with dogs and cats, they always patted dogs and stroked cats (person→dog→pat, person→cat→stroke). We presented a bird as agent, or a person as agent and a dog or cat as patient, to the trained cWM-ep SOM, to generate a distribution over expected episodes. We then used the pattern of activities in the whole SOM to reconstruct a distribution of expected actions (see Appendix A.3). Figure 3a shows that these distributions are correctly weighted towards the actions encountered during training.

The cWM-ep SOM can also make predictions about the agents and patients of episodes. These are more complex, because its predictions must be relayed to the WM individuals system, which refines them based on its own knowledge. We first consider the system's predictions about the agent of an episode. To test these, we exploited the regularity that all training episodes had animate agents. We then generated a prior distribution over episodes in the cWM-ep SOM (based on the relative hit frequency for each SOM unit remembered in its scalar weight). We reconstructed a distribution over types/properties in the agent part of the WM episode, again based on the pattern of activities in the whole SOM. Then we copied this distribution to the WM individual layer, where it provided input to the cWM-ind network. Because this input represents an expectation, each unit in the cWM-ind layer is activated proportionally to how well it matches the expectation (based on the KL divergence between the unit's weights and the expected WM individual, see Appendix A.4). The predicted distribution of types/properties in the WM individual medium is then generated top-down as a linear combination of types/properties stored in the weights of all cWM-ind units, mixed proportionally to their activities. In this way, the resulting distribution reflects the system's knowledge of recently-encountered individuals. Figure 3b shows the system's predictions about the type of the agent, both within the WM episode system and in the WM individual system, where they are biased by knowledge of the individuals that have actually been encountered in the scene. Both systems predict that inanimate agents are not possible. However, in the context where predictions were made, there were many more humans than animals; the WM individuals system thus biases its expectations about the agent towards humans.

The WM episodes and WM individuals systems also interact in generating useful predictions about the locations and properties of individuals encountered during episode perception. To test these, we exploited theseregularities in the training episodes: in all episodes involving humans interacting with dogs, the dogs were black if the agent was a man, and white if the agent was a woman; additionally, humans always appeared in the top-left quadrant of the spatial array, and animals in the top-right quadrant. We then generated an underspecified representation in the WM episode: in the agent part, we activated a representation of a person (either man or woman), and in the patient part we activated the type 'dog' (unspecified for colour); the rest of the WM episode units stayed inactive. We used this representation to generate a distribution in the cWM-ep SOM, and took the 10 most active units from this distribution to reconstruct predicted distributions of patient features. These were in turn copied to the WM individual, where they were refined by the cWM-ind network in light of its own knowledge, as before. Figure 4a shows the activity in the colour-coding features of the resulting WM individual expectation. The system correctly predicts a colour centred on black in RGB space for man→dog episodes, and on white for woman→dog episodes. Importantly, unlike expectations in WM episode medium, the cWM-ind layer is also able to generate expectations about the location of the dog: these are illustrated in Figure 4b. There is a general bias towards the quadrant containing animals, since dogs always appear in this quadrant. But there are also specific biases towards the location of the black or white dogs that the system has recently encountered, that are based on its expectations about the colour of the patient dog.

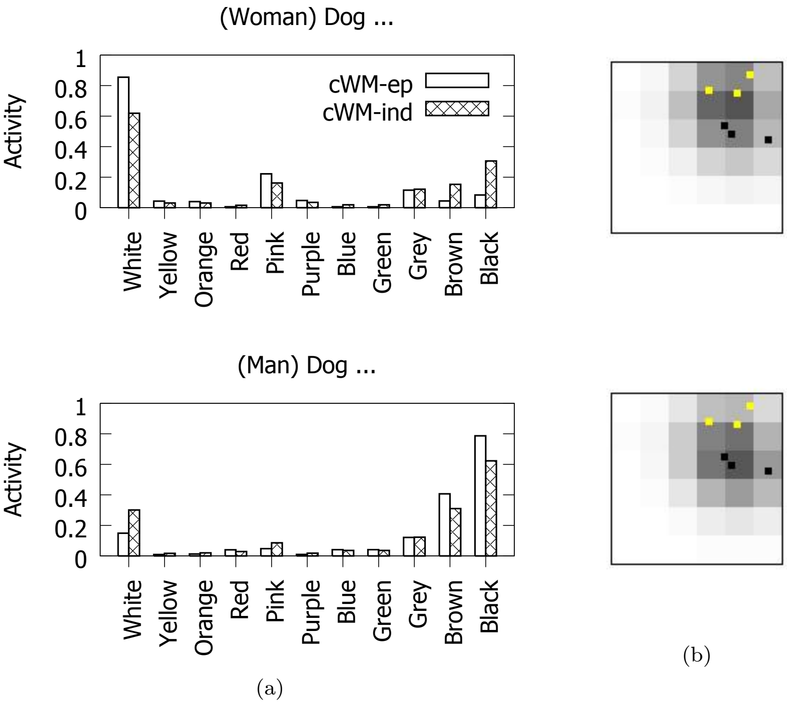


Figure 4: (a) Expectations about the colour of the patient generated by the cWM-ep and cWM-ind layers for episode fragments woman→dog (top) and man→dog (bottom). (b) Expectations about location of the patient generated by the cWM-ind layer for these episodes. Darker areas mean stronger expectations. Black (yellow) dots represent actual locations of currently present black (white) dogs.

Finally, note that generalisations in the cWM-ind layer also allow it to make sensible predictions about unseen episodes. For instance, if the system has experienced episodes where people interact with dogs and cups as targets, but not with cats or balls, it should nonetheless make predictions about the likelihood of episodes in which people interact with cats or balls, based on the similarity relations between these types of target. Cats are similar to dogs, and cups are similar to balls in our coarse-coded object representations, so the system should predict that people's actions on cats will more closely resemble their actions on dogs, while their actions on balls will more closely resemble their actions on cups. To test this, we retrained the network using episodes generated by a version of transcription rules in Table A.3 in which people always patted dogs and grabbed cups and no episodes involved a cat or a ball as a patient. Then we presented the cWM-ep SOM with a person as an agent and a cat or a ball as a patient, and generated a distribution over expected episodes. We used all active units in this distribution to reconstruct a distribution over expected actions. As shown in Figure 5, the action for the unseen cat target is biased towards ‘pat’ while the action for the unseen ball target is biased towards ‘grab’.

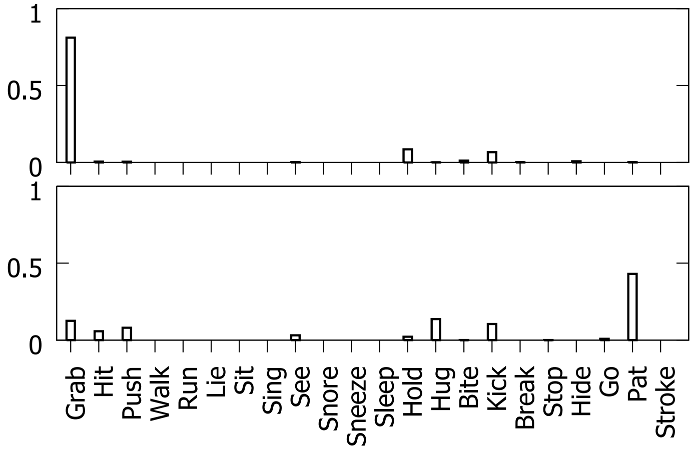


Figure 5: Action types predicted in the cWM-ep layer for episode fragments person→ball→? (top) and person→cat→? (bottom).

5.4. Testing the situation update network

In this section we test the situation update network at the top right of Figure 2. Recall from Section 2 that the distribution of candidate episodes can be interpreted as a reflection of ‘the current situation’. It is important that this distribution is updated whenever an episode occurs, because the occurrence of an episode can have a large impact on the episodes which the agent should now expect to happen, or on the actions he himself should now execute.

To test this ability, we presented our system with a sequence of training episodes, encoded as in the previous experiments, but with additional constraints on transitions between episodes: when a person hit a dog and then the (same) person patted the (same) dog, the dog always bit the person; however, when a person patted a dog without hitting it previously, any random episode would follow. Then we tested the trained network by presenting it with an episode (person→dog→pat) in two conditions—either preceded by the episode of the person hitting the dog (A), or a different episode (B). In each condition we propagated the information through the Current situation and Next episode prediction media to obtain a distribution of possible episodes. This distribution was then propagated as a prior top-down expectation to the candidate WM episode system. From there we reconstructed an expected distribution of agents, patients and actions in the WM episode medium in a standard way. Figure 6 shows the predicted agent and patient types in both conditions: while in the condition A (patting after hitting the dog, top row in the figure) there is a clear prediction of dog agent and human patient, while in the condition B (bottom row) there is a general prediction of animate agents and all possible patients.

Regarding action (Figure 7), the system correctly predicts biting in the condition A. ‘Bite’ is the strongest candidate in the condition B too, but the distribution is flat and involves other actions too.

6. Applications of the WM model in a model of syntax

As discussed in Sections 1 and 2, our model of WM is also intended to model certain aspects of language processing. In those sections we discussed how the WM model can supply aspects of

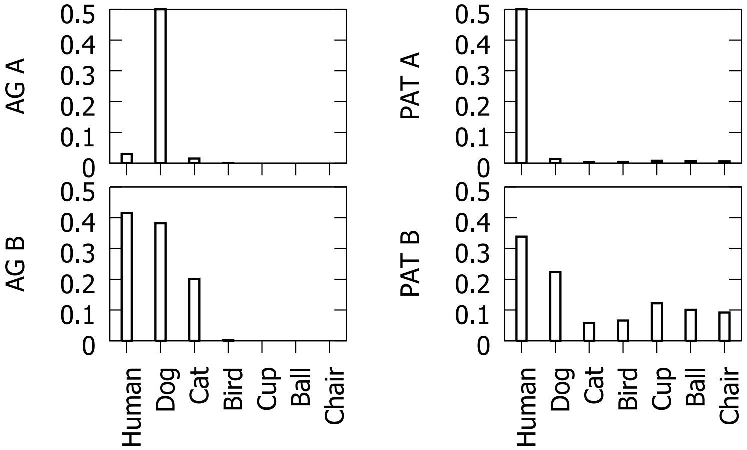


Figure 6: Prior expectation about the agent and patient of an episode following the episode (person→dog→pat) in the condition A (the man hitting the dog previously, top row) and B (a different previous episode, bottom row).

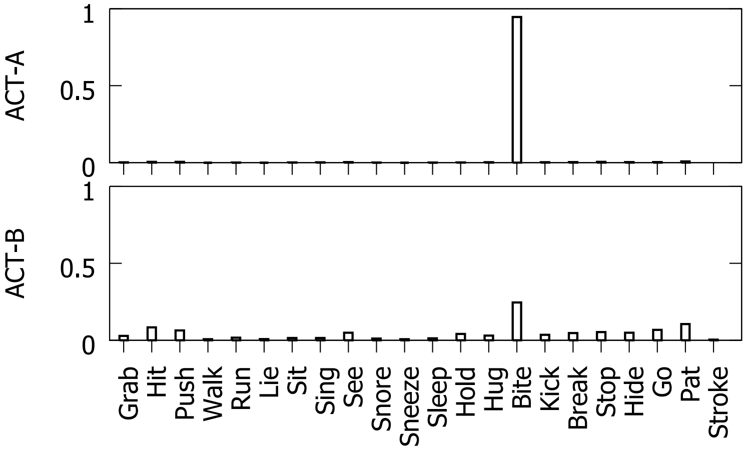


Figure 7: Prior expectation about the action of an episode following the episode (person→dog→pat) in the condition A (the man hitting the dog previously, top row) and B (a different previous episode, bottom row).

the 'current discourse context' representation that plays various roles in an account of parsing, interpretation and referring expression semantics. In this section, we argue that the WM model can also play an interesting role in an account of syntactic structures: in particular in an account of **syntactic heads**.

In models of syntax, sentences have hierarchical structure: they are nested structures of **phrases**, rather than flat lists of words. Phrases define local syntactic domains within a sentence. While some components of a phrase have relatively fixed positions within it, others have a domain that extends over the whole phrase: they can influence elements elsewhere in the phrase, and cross-linguistically, they can appear at different positions within the phrase in different languages. These elements whose syntactic influence extends over a whole phrase are called heads. The concept of a syntactic head is introduced in different ways in different syntactic frameworks. We will adopt a Chomsykan syntactic framework, which we will term 'Minimalism' (1995)[5] which presents a particularly clear model of the aspects of syntactic structure that are found in all languages. (If we are looking for aspects of syntax that reflect the semantic WM system, we can expect to find them cross-linguistically, rather than just in some languages.) We outline the Minimalist account of phrases and heads in Section 6.1; in Section 6.2 we propose that aspects of this account reflect structures in semantic WM, as it is conceived in our model.

6.1. The concept of a syntactic head, in the Minimalist framework

In Minimalism, a sentence has two syntactic structures: a **logical form** (**LF**) and a **phonetic form** (**PF**). The LF of a sentence represents its semantic structure, roughly speaking; accordingly, LF structures are relatively invariant across languages. The PF of a sentence is derived from its LF structure. Crucially, there are several alternative ways of doing this, and different languages have different conventions about how it is done: this means that PF structures are language-dependent. In the Chomskyan model, LF structures encode *innate* aspects of syntactic knowledge, that infants do not have to learn; infants only have to learn the language-specific conventions about how to map LF structures to PF structures. This innate knowledge is traditionally taken to be *language-specific* knowledge, encoded in a dedicated module of the brain. But another possibility, more consistent with modern neuroscience, is that LF structures convey information about general-purpose cognitive mechanisms (Hauser *et al.*, 2002) or about the architecture of specific cognitive systems that interface with language, such as the SM or WM systems, in accordance with 'embodied' accounts of language (see e.g. Feldman and Narayanan, 2004; Barsalou, 2008). Our general hypothesis is that LF structures convey information about the architecture of the semantic WM system—and indirectly, about the sequential structure of the SM processes that interface with this system.

With the above preliminaries, we will now introduce the Minimalist conception of heads, for two types of phrase: clauses and noun phrases. The LF structure of the transitive clause *a dog chases the big cats* is shown in Figure 8a, and the LF structure of its object noun phrase *the big cats* is shown in Figure 8b. The square boxes indicate the core structural elements of each phrase. Each box is an **X-bar schema** or **XP**, which is the basic recursive building block for syntactic structures (in Minimalism and several other syntactic frameworks, most prominently Pollard and Sag, 1994). Each word in the phrase appears at the head of its own X-bar schema: thus a verb (V) heads or 'projects' a VP (see Figure 8a) and a noun (N) heads/projects an NP (see Figure 8b). Heads are shown in red in the figures.

In the Minimalist representation of a transitive clause, the VP is dominated by two higher XPs, that introduce the verb's arguments: AgrSP introduces the subject, and AgrOP introduces the object. These elements are introduced at **specifier** positions, which are shown in blue in the figures.[6] The heads of AgrSP and AgrOP are not words, but 'agreement features', that carry the kind of information signalled by agreement inflections on verbs. For instance, the head of AgrSP

[5] We use the term 'Minimalism' somewhat loosely: our adopted model also includes elements from the theory preceding Minimalism, which is also succinctly summarised in Chomsky (1995).

[6] The subject and object also appear at positions within the VP, but we will not discuss those here.

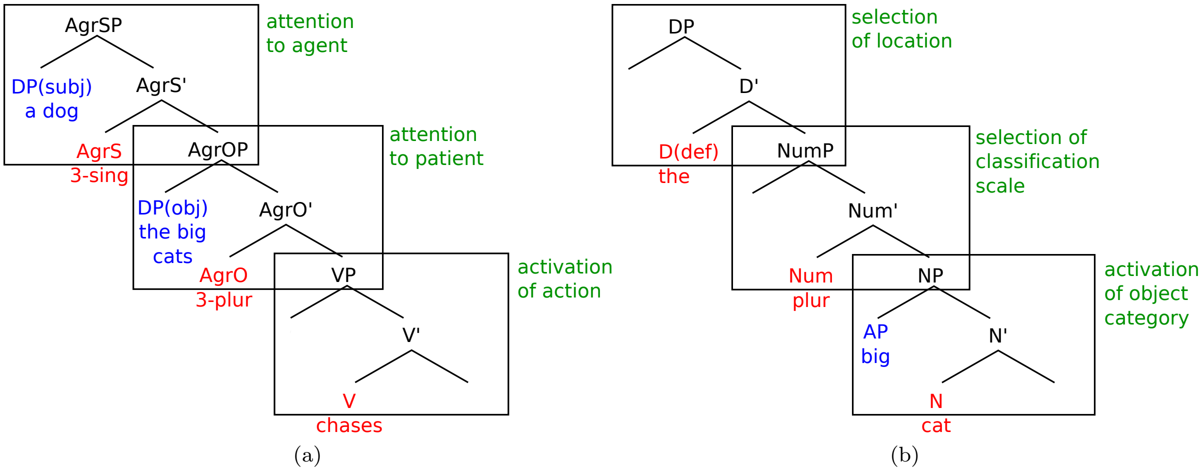


Figure 8: (a) LF structure of a transitive clause. (b) LF structure of a determiner phrase.

carries the information signalled by the agreement inflection *-s* in the verb *chases*. Agreement features can relate to PERSON, NUMBER and various types of GENDER; they convey coarse-grained information about the verb's arguments.

The key thing about the positions occupied by heads is that information can *travel* between these positions. This is represented in different ways in different syntactic theories: in Minimalism, heads are required to *move* from one head position to other head positions. For instance, in the clause structure in Figure 8a, the inflected verb *chases* originates at the head of VP, but must move to the head of AgrOP and then the head of AgrSP. This movement mechanism models how the verb is able to carry information about its subject and object, even if it is distant from these constituents in the clause. Head movement operations are also used to explain differences in surface word ordering conventions in different languages. In some languages, like Māori and French, the verb is pronounced early, while in others, like Japanese and English, it is pronounced late: in Minimalism, these differences are attributed to different conventions about how LF structures map to PF structures.

A similar notion of head movement is used in an account of the structure of noun phrases. Since the work of Abney (1996), the noun projection (NP) is taken to be introduced by a projection of the determiner (DP). The head of this projection (D) introduces a referential element—an anonymous 'x'—and the head of NP supplies a predicate to apply to this x. A key semantic contribution of the D head is to indicate whether the x it introduces is *new* in the discourse or not: an **indefinite** determiner (e.g. *a*) indicates that it is, while a **definite** determiner (e.g. *the*) indicates that it is not.[7] In most Minimalist models, there is an intermediate XP between the DP and NP, **NumP**, whose head introduces a NUMBER agreement feature, as shown in Figure 8b (see e.g. Ritter, 1991; Zamparelli, 2000). (The GENDER and PERSON features do not head their own XPs: GENDER is assumed to be conveyed by the N head, and PERSON by the D head.) The head movement operation explains many phenomena in the syntax of nominals: for instance, how nouns and determiners can carry NUMBER information, or how in some languages nouns can appear at 'high' positions, locally with determiners (see e.g. Grosu, 1988; Taraldsen, 1990). To take a simple example, consider the differences in the ordering of nouns and adjectives in English and French: in English we say *the big cats*, while in French we would say *the cats big*. If we assume that the adjective *big* occupies the specifier of NP, this ordering difference can be explained by positing

[7] 'Quantifying' determiners (e.g. *all*, *most*) also introduce referents: we will discuss these determiners briefly in Section 7.3.

that N is pronounced at its low position in English, but at a higher head position in French.

Whichever syntactic framework is used, the idea that information can 'move' between head positions in a right-branching structure of XPs is common currency for syntacticians. Note that this movement is only permitted within certain limits. For instance, the heads in a DP structure cannot freely move out of the DP to head positions in the clause. *Some* head information from *some* DPs is transmitted to the clause: for instance, in French, the PERSON, NUMBER and GENDER features of the subject appear on the verb, while in Hungarian, verbs must sometimes also agree with the object. A key task for a neural model of language is to identify the neural mechanisms responsible for agreement phenomena: that is, to provide a model of how information is transmitted between head positions in syntactic structures.

6.2. A SM interpretation of LF structures and syntactic heads

In many neural models of syntax, the syntactic structure of a sentence is a declarative representation, in which different parts of the structure are represented by different assemblies of neurons (see e.g. Reilly, 1992; Mayberry and Miikkulainen, 2008; Kalchbrenner *et al.*, 2014). In these models, implementing head movement involves transmitting information spatially, from one part of the assembly to another. This is a difficult operation for neural networks. However, there is another possible interpretation of syntactic structures, which is more consistent with the model of SM processes and WM representations presented in this paper. On this view, an LF structure represents a *dynamic* SM process—specifically, a sequentially organised SM routine. Recall from Section 4 that a WM episode is stored as a prepared SM routine, involving three operations: an action of attention to the agent, and action of attention to the patient, and the activation of a (possibly causative) motor action. These three operations can be neatly mapped onto the LF structure of a transitive clause, as shown by the green annotations in Figure 8a. A WM individual is also stored as a prepared SM routine, again involving three operations: first, selection of a spatial location (which can reactivate an existing WM individual or create a 'new' one), next activation of a classification scale (which determines whether a singular or plural stimulus will be categorised), and finally activation of an open-class object category. These operations map neatly onto the LF structure of a nominal expression, as shown by the green annotations in Figure 8b. The highest XP (DP) selects a referential element x, and identifies whether this is new or old in the current context: this is exactly what is done by the operation of selecting a salient spatial location, and determining whether or not it matches one of the candidate individuals in WM. The next XP (NumP) identifies the referent as being singular or plural: this is exactly what is done by the operation of selecting a classification scale. The last XP (NP) identifies an open-class object category; this is what is done by the operation of object classification.

These mappings between LF structures and SM routines are the basis for a strongly embodied model of language syntax (see Knott, 2012; 2014b for details). In this embodied model, the LF structure of a phrase denoting a concrete individual or episode is interpreted as a *description of the SM routine* through which this individual or episode was experienced. Experiencing the individual or episode involves executing a sequence of SM operations, and the individual or episode is stored in semantic WM as a prepared sequence of SM operations (as described in the current paper). Generating a phrase that denotes the individual or episode involves *replaying* the stored SM routine, in a special cognitive mode called **language mode**, where SM signals can activate output phonological items, through associations learned by exposure to a given language. The right-branching structure of XPs in the LF structure of a phrase is a reflection of the sequential structure of this replayed SM routine. A neural network model of sentence generation based on this proposal is presented in Takac *et al.* (2012; 2015a).

Within this SM interpretation of LF structure, there is a very natural account of head movement. As discussed in the current paper, the prefrontal assembly that stores a prepared sequence of SM operations in WM holds representations of each of the prepared operations *in parallel.* When the assembly is used to replay or simulate the stored SM sequence, there will be tonically active representations of *all* of the prepared operations in prefrontal cortex *throughout the replay process*, alongside the sequence of transiently active representations. If syntactic heads are phonological items that are read from the prefrontal areas holding these tonically active representations,

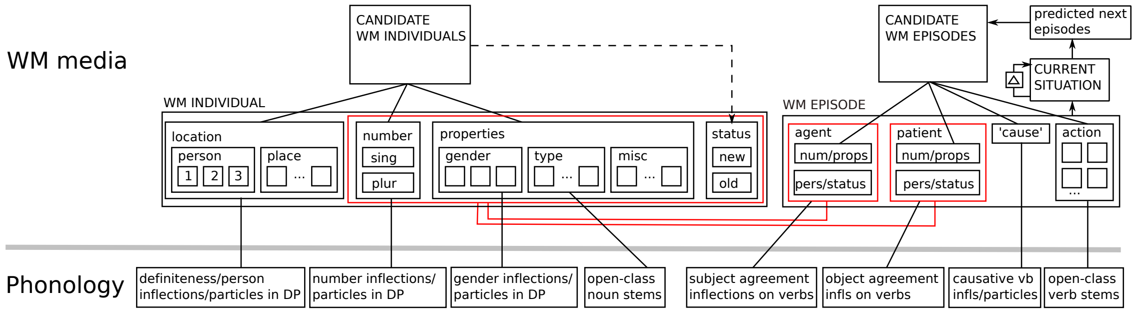


Figure 9: Interfaces between the semantic WM system and surface phonology

as shown in Figure 9, we can directly explain their extended syntactic locality: they can be pronounced at any point during the replay process. On this account, head movement does not reflect transmission of information 'in space', from one part of a neural structure to another, but rather the persistence of information *in time*. Specifically: the right-branching structure of XPs in LF represents a temporally extended process—the process of rehearsing a stored SM routine—and the movement of material between head positions reflects the presence of neural signals that are *sustained in time* during this rehearsal process. This account of head movement is quite straightforward to implement in a neural network. The network presented in Takac *et al.* (2012; 2015a) can learn languages with different constituent orderings, by learning to pronounce heads 'early' or 'late'; it can also learn a variety of non-local syntactic dependencies that manifest the extended syntactic domain of heads, such as agreement inflections on verbs and pronominal clitics on verbs, all with over 98% accuracy, even when generating structures unseen during training.

The WM model presented in the current paper extends this plan-based conception of heads in two ways. Firstly, it provides an account of nested syntactic structures, in which a phrase containing one group of heads appears at a point within a larger phrase with its own group of heads. In the current WM model, a transitive episode is stored in WM as a planned SM routine, featuring an action of attention to the agent, an action of attention to the patient, and a motor action. These tonically active planned actions are conveyed by the heads in a transitive clause: the elements shown in red in the clause structure in Figure 8a. But when the stored SM routine is rehearsed, there are particular points when *WM individuals* are transiently activated: the WM individual representing the agent is activated as the first replayed operation, and the WM individual representing the patient is activated as the second replayed operation. These activations of WM individuals happen at specific points during replay of the episode, as indicated by the blue elements in Figure 8a. But when a WM individual is activated, this presents an opportunity for a *secondary* replay operation, in which the SM routine involved in apprehending the associated individual is rehearsed. During this secondary replay process, the planned operations associated with *a particular WM individual* are tonically active. These active elements correspond to the heads *of a given DP* within the clause, (see e.g. the red elements in the object DP structure shown in Figure 8b). There are also transient SM signals active at particular points in the secondary replay process. These correspond to the fixed-position specifiers within the DP, for instance adjectives (as shown in blue in Figure 8b). In short, the concept of nested replay operations in our WM model is the basis for an account of the local domain of heads within a DP: they can move within a DP, but not beyond it. While constraints about head locality are essentially stipulated in a stand-alone model of syntax, the current account *explains* them: they are derived from a model of semantic WM for SM processes.

The second contribution of the current WM model is in accounting for how information about heads can be transmitted *between* DPs and their host clauses. Recall from Section 6.1 that some head information from DPs is transmitted to the clause: for instance, the PERSON, NUMBER and GENDER of the subject and object can sometimes surface in verb inflections or clitics. Not all head information can be transmitted this way: in particular, information about the open-class noun head cannot move outside its local DP in any language. In our model, this transmission reflects a genuine transmission of information within WM structures: namely, the copying operations through which place-coded representations of the agent and patient are created in WM episodes (again annotated by red lines in Figure 9). Recall that during experience of an episode, the WM individual medium first holds a representation of the agent, and afterwards, a representation of the patient: however, each of these representations is copied to distinct areas within the WM episode medium—and these latter representations are tonically active within a replayed WM episode. These copy operations thus provide a mechanism which allows heads in the clause to convey information about the agent and patient—which would allow the verb to carry inflections or clitics signalling the agent or patient. An important point is that the areas in the WM episode that hold copies of the fields of a WM individual *have their own interfaces to phonology* (as shown explicitly in Figure 9). This means that the information that can be expressed phonologically about the agent and patient from an active WM episode might not be the same as the information that can be conveyed from an active WM individual. In particular, we can posit that information about person, number and gender can be conveyed phonologically from both media, while information about open-class object type can only be conveyed from WM individuals. On this hypothesis, the pattern of transmission of head features from DPs to clauses is explained by the copy operation that creates place-coded representations of the agent and patient in a WM episode, plus ideas about the capacity of the interfaces between WM individuals/WM episodes and surface phonology. (Note that the copy operation that creates place-coded representations of the participants in a WM episode plays an essential role in this account. This is another piece of evidence in favour of a place-coded model of WM episode participants, separate from the representational advantages of place-coding discussed in Section 3.)

7. Applications of the WM model in an account of LTM

As discussed in Sections 1 and 2, semantic WM representations interface not only with SM experience and with language, but also with LTM. Although this interface has not yet been implemented in our model, the WM model was designed with certain ideas about the interface in mind; in this section, we will outline these ideas.

7.1. Interfaces to episodic and semantic LTM

A key idea is that that there are divisions in the WM model that echo the distinction in LTM between **semantic memory** and **episodic memory**. Semantic memory is memory for facts about objects, while episodic memory is memory for events occurring at particular times and places (Squire, 1987). The prototypical facts about objects are facts about their properties, so one central component of semantic memory is memory for the properties of objects. Within this type of memory, facts about the location of objects are a special case, held in specialised circuits; see e.g. Moscovitch *et al.* (1995). Semantic memory also supports the representation of generalisations across groups of objects, to allow statements about the properties collectively possessed by several objects, or the typical properties of objects of a certain type (Csibra and Gergely, 2009; Leslie and Gelman, 2012). Thus the facts that my dog Fido is brown, or that both my cats are grey, or that dogs typically have tails, are assumed to be stored in semantic LTM, while the fact that my dog chased my cat this morning is stored in episodic LTM.

Our WM model comprises a system for representing individuals (and token objects or groups) and a system for representing episodes. Our first proposal is that the WM individuals system interfaces primarily with semantic memory (which separately holds long-term memories of object properties and object locations), while the WM episodes system interfaces primarily with episodic memory. These interfaces are sketched in Figure 10.

On the semantic memory side, the key medium is a set of **LTM individuals**: sparse representations of particular objects, probably stored in hippocampal or parahippocampal regions (see e.g. Quiroga *et al.*, 2005; Eichenbaum *et al.*, 2007; Diana *et al.*, 2007). Each of these is a convergence zone that is linked to representations of its properties in the WM individual medium, so that activating a LTM individual activates a set of associated properties, and activating a set of properties

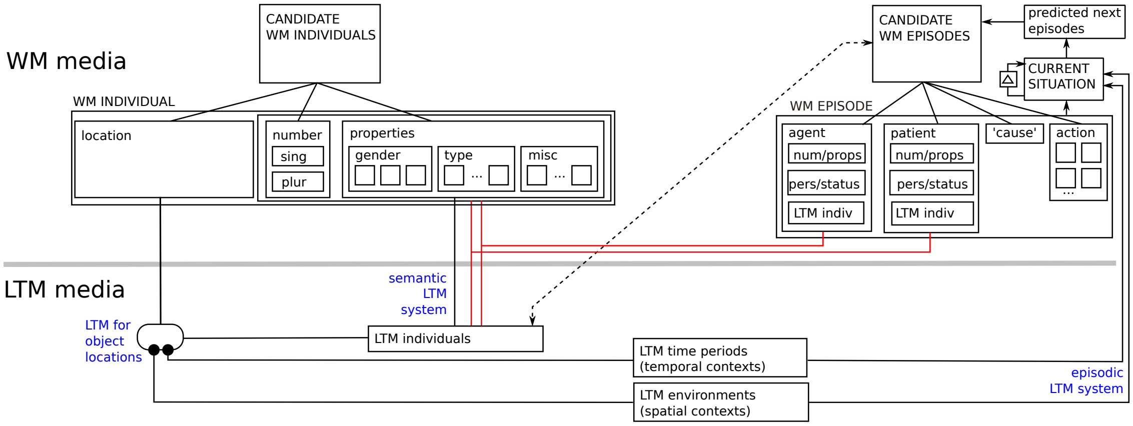


Figure 10: Interfaces between the semantic WM system and LTM systems

can activate an LTM individual (if there is one whose properties match well enough). An associated system stores links between LTM individuals and allocentric locations in the hippocampal cognitive map (see e.g. Manns and Eichenbaum, 2009). Importantly, these links are modulated by representations of both spatial and temporal context (see Ranganath, 2010). Spatial context is important, because locations in the cognitive map are only defined relative to a currently active spatial context representation (Muller and Kubie, 1987). Temporal context is important because objects can have different locations at different times.[8] In fact, the associations between LTM individuals and properties should also be modulated by temporal context, since properties can change over time, but for simplicity's sake this link is left out of Figure 10.

LTM individuals have their own interface to surface phonology, though again this is not shown in the figure. This interface represents the system of **proper nouns**, which is distinct from the system linking common nouns to representations of object types. The distinction is shown most clearly by lesion studies (see e.g. Semenza, 2006). The neural basis for the proper noun circuit is not yet well understood, though there are indications the uncinate fasciculus and left temporal pole play an important role (Papagno, 2011; Semenza, 2011). When a WM individual and an associated LTM individual are both active, the agent has a choice about how to express this linguistically. One method involves rehearsing the WM individual, and producing a full noun phrase, as discussed in Section 6.2. Another method is to generate a proper noun directly from the active LTM individual (if an associated proper noun is known).

We turn now to how WM episodes interface with LTM: a central concern in the current paper. Our first proposal is that the WM episode medium *holds copies of LTM individuals* as well as of WM individuals, as shown by the red lines in Figure 10. (A replayed WM episode therefore activates WM individuals *coupled with* associated LTM individuals.) This can be motivated both conceptually and empirically. Conceptually, it allows the WM episodes network to represent expected or planned episodes that involve specific individuals, rather than just individuals with certain properties. Agents often have such expectations, so there must be a provision for them in the model. Empirically, it sits well with Poch *et al.*'s (2011) finding that the periodic reactivation of object properties in sensory areas during WM maintenance is coupled with activity in the hippocampus, as discussed in Section 2.1.

More far-reachingly, this proposal allows the circuit that generates expectations about forthcoming episodes to be extended to hold memories of specific past episodes—that is, episodic memories. The key new elements we envisage in the episodic memory circuit are LTM media that hold representations of token times in the past, and of token spatial contexts. As with LTM

[8] There are several ways these associations might be stored—for instance, the object location memory system could be a SOM that takes inputs from locations, LTM individuals, and also temporal and spatial contexts.

individuals, we assume these are relatively sparse, localist representations of particular times in the agent's past, and particular places he has experienced. New temporal contexts are constantly being created, while new spatial contexts are only created when the agent is in a place he fails to recognise.[9] As illustrated in Figure 10, we envisage the current WM situation SOM taking input from these representations, as well as from the current WM episode and from a copy of its previous state. The WM situation SOM will learn to generalise over token times and places, where appropriate, just as it generalises over token individuals, and so should still be able to generate sensible expectations about forthcoming episodes at the present moment. However, it should also be able to store memories of situations the agent encountered in the past. These can be retrieved by presenting partial inputs to the current WM situation SOM: for instance, a WM episode representation by itself, or a place repesentation by itself. These function as *cues*, to re-activate situation representations encoded in the past, as we will briefly discuss in Section 8.1.2. When a situation is retrieved from the past, the situation SOM allows retrieval of 'the episode that happened next', via exactly the same mechanism that predicts a distribution of likely episodes during online experience. This offers a possible account of the well-known role of constructive processes during LTM retrieval (see e.g. Schacter *et al.*, 1998): the distribution over likely next episodes will be informed by the currently active token time and place representations, but also by a more generic representation of the retrieved situation, of the kind that would inform expectations about what would happen next if there were token memory. It also supports an interesting potential model of counterfactuals: if the the most active unit in the cWM-ep SOM represents the episode that actually occurred next, then the other active units can be taken to represent episodes that *could* have occurred next.

In this extended model, the cWM-ep and current situation SOMs play a role in episodic LTM as well as in WM, in line with Burgess and Hitch's (2005) proposal that representations of 'context' provide a link between LTM and WM systems. In fact, even in the basic WM model we presented in this paper, the cWM-ep and current situation SOMs quite clearly hold long-term memories: the representations they learn are acquired gradually, and implemented in long-term synaptic weights. Their role in the WM system relates to the units which are currently *active* at any given time within these media. In this sense, our model conforms to the proposal that items 'in WM' are simply elements within a LTM storage medium that are currently active—a proposal that has been voiced many times (see e.g. Cowan, 1999; D'Esposito, 2007). At the same time, our model also contains several media that are specialised WM areas (the WM individual and cWM-ind media and the WM episode medium), in line with the proposal that there are dedicated neural areas for WM representations, as in Baddeley's conception of WM. Our main point is that these two proposals need not be seen as incompatible: in our model, *some* WM representations are re-activated LTM representations, while others are patterns of activity in dedicated WM areas.

In the light of this model, we will associate the current situation SOM with the hippocampus. As discussed in Section 1, a key role of the WM episodes medium is to buffer experienced episodes as they occur in real time, so they can be transferred fast to the hippocampus, within the timeframe of hippocampal learning mechanisms. We now associate this fast storage process with the process of communicating a completed WM episode representation to the current situation SOM. Since this operation also serves to update the current situation representation, this proposal also gels with the known role of the hippocampus in encoding temporally discontiguous events (see e.g. Wallenstein *et al.*, 1998). (Note that we still consider the representation that is *active* in the current situation SOM as part of WM, in accordance with the conception of Cowan, 1999 and D'Esposito, 2007.)

7.2. Episode-based properties of LTM individuals

In many models of semantic memory, the properties of an object can be facts about episodes it has participated in, or typically participates in. These properties include so-called 'functional'

[9] We assume there will also be representations of *types* of time and place; however, these do not feature in the current sketch.

properties of classes of individuals: for instance, the property of knives that they cut things, or of animals that they move and breathe (see e.g. Tyler and Moss, 2001). But they can also include facts about the participation of token objects in episodes, if this participation is memorable in some way. For instance, if John goes on a date with Mary, this can be an interesting fact about John, as well as just an event to be recorded in episodic memory. We will refer to both kinds of property as **episode-based properties**. While there has been much debate about functional properties in psychological models of object classes (see Yee *et al.*, 2013 for a review), the question of how such properties incorporate reference to episode representations has been far less studied in neuroscience. There must be something that distinguishes an episode-level property from an actual epsiode, because they are stative facts, rather than episodes—and yet episode-level properties must ultimately be identified by experiencing actual episodes. Our model of WM suggests an interesting model of episode-level properties in semantic LTM, which we will briefly outline.

While the representation of episode-level properties is seldom considered by neuroscientists, it has been the focus of considerable scrutiny in linguistics, where such properties play a central role in models of relative clauses and quantification (see e.g. Heim and Kratzer, 1998). Linguists represent episode-level properties using a semantic operation called **lambda abstraction** that turns an episode into a property. For instance, the fact $go\_out\_with(John, Mary)$ can be turned through lambda abstraction into the property $\lambda x[go\_out\_with(x, Mary)]$, which is of the same semantic type as a simple property like 'happy' ($\lambda x[happy(x)]$): either of these can be directly predicated of the individual $John$ to create a stative proposition.

There must be some analogue of lambda abstraction in a neural model of semantic representations. In our model there is a very natural analogue. A full WM episode includes representations of all its participant individuals, which are copied from the WM/LTM individuals media. To create a WM episode that abstracts over one participant, we can erase one of these copies, and replace it with a representation that conventionally denotes a lambda-variable. For instance, to abstract away from the individual 'John' in the WM episode representing 'John goes out with Mary', we can erase the pattern in the 'agent' field of the WM episode, and activate a new pattern in this field denoting 'lambda-variable'. We can then create a unit in the candidate WM episodes SOM that encodes this abstracted episode, in the normal way. However, this SOM unit represents a *property*, rather than an episode. Finally—and this is the important step—we can associate this property with the LTM individual whose representation was abstracted over: in this case, the one representing John. This can be done simply by creating an association between this LTM individual and the SOM unit encoding the abstracted episode. This kind of direct association is very similar to the direct associations that link LTM individuals to 'regular' properties in the WM individual medium. It is illustrated in Figure 10 by the dashed diagonal line between LTM individuals and the candidate WM episodes buffer. The new type of link allows the semantic memory system to record episode-level properties of LTM individuals, as well as simple properties. Now, when we activate a LTM individual, we can activate not only a distribution over a set of simple properties in the WM individuals medium, but also a distribution over a set of episode-based properties in the candidate WM episodes buffer. (Note that this proposed account of episode-based properties rests critically on the capacity of the candidate WM episodes buffer to hold a *distribution* of episodes.) We will refer to this model of episode-based properties in Section 7.3, which discusses quantified sentences, and in Section 8.1.3, which discusses relative clauses.

7.3. Quantified sentences

Our model of the syntactic relationship between DPs and clauses rests on a model of the relationship between WM individuals and WM episodes. However, the model introduced in Section 6.2 only considers *referential* DPs, such as *a dog* or *the cats.* It is important to show that the model also accounts for **quantifying** DPs, such as *most students* or *no students*, in sentences such as *Prof. Smith likes <u>most students</u>*, or *<u>No students</u> like Prof. Smith.* These DPs do not directly report on attentional processes that identify an object or group: they have a more complex semantic denotation. If we want to make a general claim linking DP structures to WM individuals, we need to set out a model of how quantified propositions are cognitively represented, and we need to give WM individuals a role in this process that squares with the role that quantified DPs play in quantified sentences. We cannot consider this in any detail in the current paper, but we will sketch an account along the required lines.

In the standard linguistic model of quantification (see e.g. Barwise and Cooper, 1981), a quantified sentence introduces two *sets* of individuals, and reports the cardinality of their intersection. For instance, *Prof. Smith likes most students* introduces a set of students, and a set of individuals who Prof. Smith likes, and asserts that the intersection of these sets contains some large proportion of the set of students. The quantifying DP *most students* introduces one of these sets through its noun *students*, and supplies the assertion about cardinality through its determiner *most*. The other set (of individuals liked by Prof. Smith) is introduced by a clause representing an abstracted episode of the kind discussed in Section 7.2: $\lambda x[likes(Smith, x)]$. At PF, the quantifying DP appears within this clause, at the position of the abstracted element: *Prof. Smith likes most students*. However, at LF it is assumed to appear at a high structural position outside the clause, and dominating it (again see Heim and Kratzer, 1998). This high position can be seen as reflecting the semantic structure of a quantified proposition, where the quantifier is always the highest operator: in the current case, $\underline{most(x)}[student(x), likes(Smith, x)]$. Our general hypothesis that DPs convey material from a (replayed) WM individual suggests that WM individuals have a special role in representing quantified propositions, and that something about this role is captured by the raised position of a quantified DP in the LF of a quantified sentence. We argue that this is indeed the case.

Our proposal begins from the assumption that quantified sentences do not report SM experiences directly, but rather report operations that query an agent's *LTM* for SM experiences—specifically, an agent's *semantic* LTM system. Our main proposal is that *queries to the semantic LTM system are presented within the WM individuals medium*, just as queries to the episodic LTM system are presented within the WM episodes medium (as discussed in Section 7.1).

Recall that semantic LTM is LTM for the types and properties of individuals. For instance, we suggest that the open-class object type 'student', placed in the 'type' field of a WM individual, can be used to retrieve a *set of LTM individuals* that are associated (through semantic memory) with this type: that is a set of token students. These individuals may have been encountered at different times or places—but since semantic memory is implemented as enduring associations between LTM individuals and types/properties in the WM individual medium, types can function as queries, to retrieve sets of LTM individuals encountered in a range of different contexts. A second, separate suggestion is that the size (or rather, numerosity) of the activated set of LTM individuals can also be stored in the WM individual medium, in its 'location' field. There is good evidence that the numerosity of a perceived group is stored in the spatial attention system, and in prefrontal spatial WM; see in particular Nieder and Miller (2004). Our proposal is that in operations querying semantic LTM, the prefrontal numerosity system, which is part of the location field of a WM individual, can represent the number (or rather numerosity) of LTM individuals retrieved by a query to semantic memory.

Through semantic memory, the LTM individuals in the set activated by a type query will each be associated with an idiosyncratic distribution of properties in the WM media that hold the properties of perceived individuals. Through these associative connections, we can activate an *ensemble* of properties in these WM media. Some of these might be activated by single LTM individuals, while others might be activated by more than one individual—or, in the right situations, by arbitrarily large subsets of the active LTM individuals. We propose there is a mechanism in the WM system that *selects* a single property in this WM medium: perhaps one of those in the activated ensemble, or perhaps a property the agent is interested in for independent reasons. The selected property can be used as a second query to semantic memory, that reduces the set of active LTM individuals to those that possess this property. (Possibly the empty set, if none of them have the property.) Importantly, this second query must be expressed *within the same WM individual* that was used to express the first query and hold its results: the property selected for the second query can be selected from amongst the results returned by the first query, so a single WM individual functions as an evolving 'workspace' for the two queries. A key aspect of its evolution relates to the location representation encoding the number of active LTM individuals. After the second query, only those LTM individuals that have both queried properties will remain active, so the represented numerosity might diminish: either to zero, or to some proportion of its original size. We propose that the location field can record both absolute numerosities (including zero) *and relative drops in numerosity*—and that quantifying determiners like *most* and *no* are read from these recorded values in the location field of a WM individual.

Finally, note that some properties associated with LTM individuals can be encoded in the WM episodes medium: namely the episode-level properties discussed in Section 7.2. Say that the episode-level property 'Prof. Smith likes x' was recorded for several of the students retrieved by the first query 'student'. This property might be identified as interesting, and selected as the second query, reducing the set of active LTM individuals to those students liked by Prof. Smith, and recording the relative drop in numerosity within the 'location' field of the controlling WM individual. This WM individual can now function as the semantic representation from which the quantified DP *most students* is read.

It remains to explain how this DP ends up at a position *inside* the clause expressing the episode-level property in a surface sentence. Our proposal here begins with the observation that if the episode-level property in the candidate WM episodes SOM is recreated in the WM episode medium, we have a structure which can potentially be rehearsed to produce a sentence. There will be a 'gap' in the recreated structure, representing a lambda variable rather than a participant. However, the WM individual used to formulate the queries is a natural filler for this gap. If this WM individual is copied into the gap position, using the normal mechanisms for copying WM individuals into participant slots, we have a WM episode which, if rehearsed in the 'language mode' discussed in Section 6.2, will produce a quantified sentence with the quantified DP in an argument position. Crucially, while this DP is produced at a specific transient moment during replay of the WM episode, the WM individual it encodes conveys high-level information about the complete query process, encompassing both queries to semantic LTM and their results. This account of quantified sentences explains both why the quantifying DP has a position above the clause at LF, and why it appears within the clause at PF.

8. Discussion

In this paper we have presented a model of semantic WM representations. Our emphasis has been on semantic WM as an interface medium, that links separately to primary SM media, to surface language representations, and to LTM. These separate interfaces are shown graphically in Figures 2, 9 and 10. The model is supported in a variety of ways. The model of WM individuals is supported by empirical experiments identifying a mixture of transient and persistent representations of objects in WM, and suggesting that these representations have the capacity to actively recreate the sequences of SM stimuli they encode (see Section 2.1). The model of WM episodes, which uses place-coding to represent episode participants, is supported by experimental evidence for a 'slot-based' storage medium for object representations in WM, and evidence that episodes are experienced through sequentially structured SM routines, along with evidence about the nature of the planning representations in PFC that store prepared sequences of SM operations (see Section 2.2). The place-coding model of WM episodes is also supported by its computational properties: it provides an attractive model of how representations of individuals are bound to semantic roles such as AGENT and PATIENT (see Section 3.1), which in turn supports an attractive account of how distributions over episodes are represented (see Section 3.2). The model of probability distributions over episodes then provides the basis for a WM representation of 'the current situation', or in experimental terminology the current 'cognitive set', which biases the SM experience of episodes towards expected or desired outcomes, and which supports the performance of relatively complex tasks by amnesic patients (see Section 2.3). The WM model is also independently supported by a different set of data, relating to the role it can play in an account of linguistic representations. It contributes directly to an account of how syntactic structures can be interpreted in neural terms: in particular, it gives a very straightforward account of the domains of syntactic head features, and of how head features are transmitted between DPs and their host clauses (see Section 6). It also supplies a natural representation of the 'current discourse context', along with operations for updating this context, and for introducing new discourse referents and reactivating existing referents (see Sections 2.3 and 6). Finally, the WM model makes some proposals about how semantic WM representations interface with LTM, which was a key concern for Baddeley (2000) in his original arguments for an episodic buffer. We can envisage separate WM interfaces for semantic and episodic LTM systems (see Section 7.1). These two interfaces jointly allow an interesting model of episode-level properties and of quantified propositions (see Sections 7.2 and 7.3): these both contribute to a model of linguistic representations as well as to a model of LTM.

The model we propose makes several novel experimental predictions. One set of predictions relate to our place-coded model of WM episode representations. We posit there are several distinct media for holding WM representations of individuals as these participate in episodes, which are specialised for different semantic roles—for instance for agent and patient participants of an episode. On this basis we predict there will be more activity in brain areas involved in WM for individuals when subjects retain a transitive episode in WM, compared to an intransitive episode.[10] More specifically, since in our model WM individual representations are copied to the agent and patient media *at different times* during experience of an episode, we predict that there will be changes in the functional connectivity of the agent and patient WM media during the timecourse of episode perception. (These changes might not be identifiable at the coarse temporal resolution of fMRI, but may be revealsed in EEG or MEG analyses.) Another set of predictions relate to our claim that the structure of semantic WM representations is reflected in certain syntactic phenomena: specifically, that the extended syntactic domain of heads results from the tonic activity within WM individual and WM episode representations. This claim leads to the prediction that patients with impaired ability to hold individuals and episodes in WM will also show impairments relating to the use of syntactic heads: for instance in the processing of agreement morphology on verbs and nouns.

Of course the model presented here is very preliminary, and leaves many open questions. There are two obvious questions to be addressed, relating to the model's storage requirements, and to whether it can give an account of nested propositional structures. We will conclude by addressing these questions.

8.1. Semantic representations containing nested propositions

Our model already accounts for some types of hierarchical structure in syntactic representations. Through its account of the interface between WM individuals and WM episodes, it accounts for how DPs, with their own internal structure, can appear at positions within clauses. It also accounts for the hierarchical relationships between the chain of right-branching XPs in a single clause: higher XPs denote SM operations that occur earlier in SM routines, and thus create the context within which the SM operations denoted by lower XPs are interpreted. However, the model must also be able to represent semantic structures containing multiple propositions, or multiple episodes. In grammatical frameworks, these structures are described with *recursive* syntactic rules, allowing clauses to be embedded inside other clauses. We will consider three examples. The first is a **complement clause**, introduced by a modal verb: for instance *John says/believes/hopes/fears [that the sky is blue].* The second is a **subordinate clause**, introduced by a subordinating conjunction: for instance, *[When/if John attacks], run away.* The third is a **relative clause**, introduced within a DP: for instance *The dog [that chased me] barked.* In each case, we will argue that our model can represent aspects of these structures that other network models of nested semantic representations cannot.

In each case, our account turns on a single key assumption, which is already built into the model: namely that activating a semantic representation in WM triggers the execution of a temporally extended *sequence* of signals, in which different signals are active at different times. In our model, activating a WM representation of an individual involves *simulating a SM experience*: a

[10] In fact there is already some evidence for this: for instance Shetreet *et al.* (2011) found activity in areas of cingulate cortex and precuneus increased as a function of the number of thematic roles to be retained. These areas are implicated in WM for objects, as detailed in the paper.

process that is extended in time, during which first-order SM representations of location, number and type are active at different times. Activating a WM representation of an episode involves simulating a higher-level SM routine, during which the WM individuals medium holds different individuals at different transient moments (while at other moments, first-order representations of motor actions become active). There is a natural extension of this model to representations involving multiple episodes: we propose that activating such a representation involves a temporally extended routine, in which *different episode representations occupy the WM episode medium at different times*. In this model, the WM medium that holds episode representations only need represent one episode *at a time*. This avoids some of the technical problems faced by existing models of nested propositional structures, in that it eliminates the possibility of cross-talk between episodes. At the same time, the temporal separation of episodes has specific advantages, for each distinct type of nested episode, which we will discuss below.

8.1.1. Complement clauses

We have already presented a model of sentences containing complement clauses as sequences of episode representations, as part of a larger network model of vocabulary development (Caza and Knott, 2012; Knott, 2014a). Our model is an implementation of Tomasello's (2003) social-pragmatic theory of word learning. In this theory, before infants can learn word meanings efficiently, they must first do some meta-level learning about the social institution of communication: they must learn that certain physical actions (e.g. talking) are special, in that they convey meaning; and they must learn to represent these actions in a special way, that identifies the conveyed meaning. The special semantic representations are effectively the semantics of clauses with sentential complements: for instance, *Mother says (to me, or some other interlocutor) that P*, where P is a whole clause. In our model of word learning, infants learn that physical actions of talking are special because they predict good opportunities to learn word meanings. (When the infant perceives a speaker talking, and establishes joint attention with this speaker, the relationship between incoming words and incoming SM concepts is temporarily less noisy than normal.) The infant uses this meta-level learning to focus her regular word-learning processes on talk actions, in line with evidence reported by Tomasello (2003). In our model, infants operate in two distinct cognitive modes, with different patterns of connectivity: in 'experience mode', semantic representations in WM are activated through the SM system, and in 'language mode' they are activated by phonological representations. (In relation to the current paper, these two modes are activated by selectively enabling the interfaces shown in Figures 2 and 9.) Engaging language mode is under operant control: infants learn to engage language mode as a conditioned response to identifying a physical 'talk' action. Crucially, after this learning, when an infant monitors a talk action, she evokes two WM representations *in succession*: first, a representation of the physical action of talking, 'Mother executes a speaking action (directed to interlocutor X)', and then, after language mode is engaged, a representation of the *content* of the speech action just identified, activated by its phonological words. If these words form a sentence that denotes an episode, the infant's representation of the complete communicative action will comprise *a sequence of two WM episodes*, separated by an intervening operation that changes the cognitive mode. This representation conforms to the general proposal advanced in the current paper: semantic WM representations represent sequentially structured SM routines. In this case the routine involves activation of two successive episodes in the WM episode medium.

This model of clausal complements has an advantage over many other models of nested clauses: it represents not only the content of the clausal complement, but also the special modal context within which this content must be interpreted. The propositional content of a speech action is special in two ways. Firstly, a hearer interpreting a spoken utterance is not committed to believing this content, but instead records it in a special context associated with the speaker's beliefs. (If Mary tells us that P, we are not committed to believing that P, only to the fact that Mary believes P.) Secondly, the content of an assertion is 'intensional', meaning that the actual words used to report it are important. (Say that Mary and John are both axe murderers: if Mary tells us she loves John, it is true that an axe murderer told us something, but it is certainly not true that Mary told us she loves an axe murderer.) If a physical utterance and its content are represented as successive WM episodes, then our model naturally creates a special context for the content of an utterance, since a *situation update* intervenes between them. The new situation is a function of the speaker's physical utterance, so it plausibly establishes a context representing this speaker's beliefs. Since this situation update also coincides with a transition into 'language mode', that is a mode where WM representations are activated by words rather than SM experiences, our model also naturally implements the fact that the content of an utterance must be interpreted intensionally. Models of nested clauses in which matrix and complement clauses are active simultaneously, in a single pattern of activity, do not permit such a straightforward implementation of modal contexts.

It is also worth noting that in a syntactic model, the syntactic projection (CP) that introduces the complement clause is a 'barrier' to head movement, preventing heads in the complement clause moving to the main clause and vice versa (see classically Chomsky, 1986). This is something that has to be stipulated in a stand-alone syntactic model. But in our account it is just a corollary of our proposal that head movement is a consequence of tonically active WM representations: the main clause and complement clause are read out from the WM episode medium at two different times, when the medium holds different tonically representations, so there is no opportunity for heads from one clause to be read out in the other one.

8.1.2. Subordinate clauses

There are many kinds of subordinate clause, but we will focus on the temporal subordinator *when* and the conditional subordinator *if*, in sentences of the form *[If/when episode1] episode2*. Again, we propose that the WM representation that encodes such sentences is a sequence of two WM episodes, active consecutively in the WM episode medium. And again, we propose that the special semantic contribution of the subordinate clause structure can be well conveyed by an account of the cognitive operations that intervene between the two WM episode representations.

In a standard account of sentence semantics, the meaning of a sentence is modelled as a function that updates the 'current discourse context' (see again Kamp, 1981; Heim, 1982). For instance a sentence reporting the episode *John kissed Mary* asserts that this episode occurs at the currently active temporal context (whatever that is), and also resets this temporal context to be the state that obtains after the asserted episode is completed. Subordinate clauses introduced by *if* and *when* are modelled as operations that set the current discourse context to a new value, so that their matrix clause updates a specified discourse context, rather than the default one. If the subordinate clause is introduced by *when*, it is presupposed that the episode it expresses has already happened, or will happen in the future; if it is introduced by *if*, there is no such presupposition.

In our model, there is a natural analogue of both kinds of context-resetting operation. As discussed in Section 2.3, we interpret references to 'the current discourse context' in language models as references to the currently active WM situation: so in our terms, subordinate clauses introduced by *if* and *when* signal an operation that sets the WM situation to some arbitrary new value. We propose that alongside the mechanism that updates the WM situation as a function of the episode just experienced, there is a competing mechanism which reactivates an arbitrarily distant situation from LTM, based on its resemblance to the current situation,[11] and establishes a special mode where WM episodes are retrieved from memory, rather than through SM experience. This proposal is supported by two recent strands of empirical work. There is good evidence that the brain can switch between alternative modes of connectivity, implemented in large-scale brain networks, and that one of these modes relates to retrieval of material from episodic memory (see Buckner *et al.*, 2008 for a review). There is also evidence for a network whose role is to interrupt an ongoing stream of SM experiences, that operates at the boundaries between experienced episodes (Corbetta and Shulman, 2008). We suggest this proposal offers a natural account of the semantics of subordinators like *if* and *when*. Specifically, we suggest that the subordinator signals the

[11] This resemblance could be determined by having the current WM situation produce a distribution of activity in the set of LTM times and/or LTM environments. If an LTM time or place becomes sufficiently active, this indicates that a situation similar to the current one occurred at that time/place, and this situation, along with its associated time/place, could be reactivated.

operation that interrupts processing and establishes a remote situation, and the subordinate clause identifies the newly established situation. (Since the new situation is a SOM unit, the episode associated with it can be reconstructed top-down, after which it can be rehearsed like a normal episode.) Following this, the episode that happened 'in' the restored situation can be recalled, via the next-episode prediction network.

In relation to the current discussion of multi-clause structures, the key point about this model is that a sentence with a subordinate clause reports a temporally extended sequence of operations, within which the semantics of the subordinate and main clauses are active at different times. The operation establishing a new WM situation and its associated circumstances strictly *precedes* the operation retrieving the episode that happened in this recalled situation. We propose that an agent communicates the sequentially structured experience of being reminded of a past situation simply by rehearsing this experience, including its sequential structure, in the language mode discussed in Sections 6.2. The rehearsal operation is a slightly higher level one, in that a pair of WM episodes are rehearsed, but there is still a single WM representation that supports the complete process, namely the retrieved situation, which links both to the antecedent and consequent WM episodes. In summary, the idea that semantic representations are rehearsed sequences extends naturally to an account of subordinate clauses that reset the current discourse context.

It is also worth noting that our model provides a good means for distinguishing the semantics of *if* and *when*. As already noted, *when* presupposes the episode denoted by the subordinate clause has occurred in the past, or will occur in the future: in our model, this means it will be associated with a specific time period. There is no such presupposition about the antecedent episode introduced by *if*; in this case, in our model, the retrieved situation should be a generic one, not associated with any particular time. On this model, the rule *If X, then Y* describes the operation of the agent's situation-update mechanism, without making any reference to the contents of episodic LTM. If, as we discussed in Section 2.3, this update mechanism resides in prefrontal cortex, our account of the semantics of this rule is consistent with Wallis *et al.*'s (2001) proposal that assemblies in prefrontal cortex encode general rules.

8.1.3. Relative clauses

As discussed in Section 6.2, our model is explicitly designed to represent the hierarchical relationship between a DP and its host clause. However, a clause can also be embedded in a DP, most obviously in a relative clause: it is important to make sure the model can be extended to account for nested structures of this sort. Again we suggest that it can—and furthermore, that the resulting model offers interesting advantages over existing models.

A speaker only produces a referential relative clause when this identifies some property of the intended referent that distinguishes it from distractors. We begin by situating this operation in a broader framework for DP planning, and then consider the relative clause mechanism specifically.

A sketched model of DP planning for referential DPs. Recall from Section 7.1 that the speaker can choose to report a referent using the WM individuals system, or alternatively by identifying an LTM individual the hearer knows about. In the former case, the generated DP must report either the introduction of a new WM individual or the reactivation of an existing one (which must then be identified uniquely). In the latter case, it can identify a known LTM individual by name, or by specifying properties unique to this LTM individual. In either case, our current model provides a good framework for planning a DP.

When the DP references the WM individuals system, if the WM individual is flagged as 'new', it is rehearsed as-is, to generate an indefinite DP. If it is flagged as 'old', we envisage generation happens in two passes. In the first pass, the speaker activates a minimal WM individual featuring only the individual's number and gender, and uses this to retrieve all matching referents in the candidate WM individuals medium. If there is only one, a pronoun can be used. If not, the speaker activates a slightly richer WM individual featuring the individual's number and basic-level type, and again retrieves all the matching referents in the candidate WM individuals medium. If there is only one, there is no ambiguity, and this minimal WM individual can be rehearsed to create a definite DP. If there is more than one, then a property should be sought that distinguishes the intended referent from other WM individuals. This can be done straightforwardly in our architecture, by activating the properties of the intended referent while collectively *inhibiting* those of the distractors, and then picking the most active property.[12]

When the DP references the LTM individuals system, if the LTM individual has a name, it is used in place of a full DP. Otherwise, the speaker again creates a minimal WM individual featuring number and a basic-level type, and then identifies properties of the referent which make it unique *amongst all LTM individuals* of this selected type. Again this can be done straightforwardly in our system, by treating the WM individual as a query to semantic memory, to activate a set of distractor LTM individuals of the specified type, and then inhibiting the collective properties of these individuals, while positively activating the properties of the referent, and then selecting the most active property to include as a modifier in the DP.

Note that in this account of referring expression planning, the WM individual medium plays a central role. It is well positioned to determine the content of a DP, since it can hold queries both to WM and to LTM representations of individuals. At the same time, it underpins an account of the syntax of DPs (in particular of syntactic heads in the DP), as discussed in Section 6.

Generation of a relative clause. In some network models (e.g. Stewart and Eliasmith, 2012), the semantics of a sentence containing a relative clause is a single pattern of neural activity, in which representations of matrix and nested clauses are active simultaneously. In the model we envisage, the semantics of the matrix and embedded clauses are active in the WM episode medium at different times. Specifically, we propose that rehearsal of the matrix episode is briefly *interrupted* by rehearsal of the relative clause episode. Our motivation for this model is that the relative clause serves a very different purpose from the matrix clause. The matrix clause conveys an episode the speaker has experienced, involving various participants, whose properties and/or identity the speaker apprehended directly. A relative clause functions to identify one of these participants *to the hearer*: its content need have nothing to do with the experience the speaker wants to report. We argue there is no cognitive representation in which the semantics of the matrix clause is combined with that of a relative clause, and that by activating them at separate times, the distinct functions of these clauses can be better modelled.

To illustrate, consider *The dog [that chased Mary] bit John.* The purpose of the main clause here is to convey an experience the speaker has just had, in which a certain dog bit John. The purpose of the relative clause is to identify the dog in question to the hearer, in a case where there are several candidate dogs. We propose that the speaker begins by rehearsing the WM episode conveyed by the matrix clause *The dog bit John.* This WM episode contains pointers to two WM individuals: a token dog, and a token person (John), each associated with an LTM individual. During rehearsal, the speaker loads each WM/LTM individual pair in turn, creating the context for two separate DP planning scenarios of the kind just sketched above. During the first of these, when producing a DP to refer to the specified token dog, a property of this dog is sought that distinguishes it from a set of distractor dogs. In the current scenario, the selected property is an *episode-based property* retrieved from semantic LTM, of the kind discussed in Section 7.2, namely 'x chased Mary'. The question now is how this episode is expressed verbally, given that the WM episode medium already holds 'the dog bit John', and is halfway through rehearsing this episode.

The basic scheme we have in mind is a very traditional one, originally proposed by Miikkulainen (1996), which recruits a network implementing a general-purpose *stack*, with push and pop operations (Pollack, 1990).[13] At the point when the relative clause is to be expressed, the matrix episode 'the dog bit John' is pushed onto this stack, along with a pointer to the position at which rehearsal should be resumed, and the episode-level property 'x chased Mary' in the candidate WM

[12] We envisage that the distractor individuals are activated in the candidate WM individuals medium; then their properties are collectively activated in the properties medium; then the active properties are inhibited; finally, the properties of the intended referent are positively activated on top of this pattern. The properties that are active after this will be those only possessed by the referent.

[13] Miikkulainen models sentence interpretation rather than sentence generation, but the stack has a similar role in both cases.

episodes SOM is reconstructed in its place in the WM episode medium. This is then rehearsed, generating the relative clause. When rehearsal is complete, 'the dog bit John' is popped from the stack back into the WM episode medium, and its rehearsal is resumed. The crucial point in this model is that the two clauses are generated by distinct semantic representations, activated by distinct mechanisms, and active at different times, in accordance with their very different pragmatic roles in the sentence.

8.2. Space and capacity analysis of the model

We conclude by discussing the storage requirements of our proposed network. (We will not include the network that implements LTM for object locations in our calculations, since this is not the focus of the current paper.)

We first make a rough estimate of the size of the media that represent object tokens and properties, which in our model are copied to the WM episode medium. As our estimate for the size of the 'properties' medium, we will use the size of the penultimate layer in a high-performing convolutional neural network for visual object classification (Simonyan and Zisserman, 2014), 4000 units, scaled by a factor of 2.5 to account for other sensory modalities, yielding a total of 10,000 units. As our estimate for the number of LTM individuals that can be individually distinguished, we will use a figure of 20,000, which encompasses a relatively small number of well-known individuals in a personal network (on the order of 2000, according to Killworth *et al.*, 1990), plus 18,000 miscellaneous token objects (roughly 20 instances of each of the roughly 900 basic-level nouns identified in WordNet by Izquierdo *et al.*, 2007 Table 2).

Based on these estimates, the 'agent' and 'patient' media in the WM episode must each hold 30,000 units. In addition, the 'action' medium must hold a repertoire of actions. We estimate that 2000 action categories that are represented, based on a cross-linguistic measure of verb vocabulary size (Tang and Nevins, 2013).[14] There are therefore 62,000 units in a realistically sized WM episode medium.

We now consider the appropriate size of the candidate WM episodes SOM. Recall that this medium does not need to represent all possible episodes: only recalled and expected episodes (and these can in some cases be represented as generic episodes rather than token episodes). Neither does this medium need to represent episodes with 'nested' episodes: as discussed in Section 8.1, the relations between these episodes are stored in the situation SOM for complement and subordinate clauses, and retrieved from episode-based properties for relative clauses. We estimate the candidate WM episodes SOM must hold around 10^6 episodes, based on the 10^6 'common-sense axioms' in Cyc's knowledge base (Lenat, 1995). We cannot expect perfect efficiency in the candidate WM episodes SOM: in our experiments, 25% of its units never 'won' the competition to represent a WM episode. Erring on the side of caution, we estimate that the scaled-up candidate WM episodes SOM must contain 10^7 units.

We now turn to the number of connections implied by these estimates. The WM episode and candidate WM episodes media are fully connected, so there are $62,000 \times 10^7 = 62 \times 10^{10}$ connections between these media. Also, recall from Section 7.2 that episode-based properties are stored in links between the LTM individuals layer and the candidate WM episodes SOM, resulting in an additional $20,000 \times 10^7 = 20 \times 10^{10}$ connections.

We now consider the size of the current situation SOM. This network holds localist representations of all situations the agent encounters, which as before can be generic types of situation or specific token situations in episodic LTM. The main purpose of a situation representation is to hold information about the episode that occurred (or will occur) 'in' this situation. On the assumption that situations encode 'discourse contexts', which are updated after each eventive sentence in a narrative (see Section 2.3), we will use a textual method to estimate the number of distinct situations that must be stored. A typical novel contains around 2,600 sentences.[15] If

[14] While there may also be units encoding action types, we assume their number is small in comparison with the number of units representing token actions.

[15] Source: the now-defunct Amazon 'text stats'.

Units		Links	
Properties	1×10^4	→ agent, patient	2×10^4
LTM individuals	2×10^4	→ WM episode	4×10^4
		→ cWM-ep-SOM	20×10^{10}
		→ properties	20×10^7
WM episodes	6.2×10^4	→ cWM-ep SOM	62×10^{10}
cWM-ep SOM	1×10^7		
current situation SOM	2.6×10^6	→ WM-ep/prev-sit'n/LTM times/env's	42.6×10^{10}
		→ predicted next episode	2.6×10^{13}
previous situation	6.2×10^4		
LTM time periods	2×10^4		
LTM environments	2×10^4		
predicted next episode	1×10^7	→ cWM-ep SOM	1×10^7
Total (approx)	2.28×10^7	Total (approx)	2.72×10^{13}

Table 1: Estimated dimensions of a realistically sized network

we assume that a person's inventory of situations equates to detailed knowledge of 100 novels, excluding overlapping generic situations, the situations SOM must store 2.6×10^5 situations, and allowing for the same degree of redundancy as the episodes SOM, should hold 2.6×10^6 units.[16]

The current situation MSOM takes input from four media: the WM episode, the layer representing the previous situation, the set of LTM time periods and the set of LTM environments. The previous situation layer is represented in the MSOM as the weights of the winning unit in the previous situation, as discussed in Section 4, and thus holds 6.2×10^4 units. We estimate that token LTM environments stand in a 1:1 relation with LTM individuals, because places can be reconstrued as objects; on this basis there are 20,000 LTM environments. We estimate there are a similar number of LTM time periods.[17] The situation SOM therefore receives input from a total of $6.2 \times 10^4 + 6.2 \times 10^4 + 2 \times 10^4 + 2 \times 10^4 = 16.4 \times 10^4$ units. The number of connections into the situations SOM is $16.4 \times 10^4 \times 2.6 \times 10^6 = 42.6 \times 10^{10}$. The situations SOM provides output to a medium the same size as the candidate WM episodes buffer (10^7 units), requiring an additional $2.6 \times 10^6 \times 10^7 = 2.6 \times 10^{13}$ connections. This medium is mapped by 1:1 links to the candidate WM episodes buffer itself, requiring an additional 10^7 connections.

As summarised in Table 1, we estimate our model when scaled up will require around 2.28×10^7 units and 2.72×10^{13} connections. Following Stewart and Eliasmith (2012) we assume each unit in our model is implemented by a local assembly of 100 actual neurons: on this basis the network would require 2.28×10^9 neurons.[18] There are at least 8.6×10^{10} neurons in the human brain (Azevedo *et al.*, 2009), and at least 1.6×10^{14} synapses (Tang *et al.*, 2001); our projected network uses less than 10% of available neurons, and less than 20% of available synapses. Even if we assume each unit in the network corresponds to an assembly of 100 actual neurons, the network can still be accommodated within the number that has reasonable complexity, given that it models substantial parts of LTM as well as semantic WM.

[16] In this scheme, there are fewer possible situations than possible episodes, by a factor of 100. We consider this reasonable, given that many situations are generic. (Note that a generic situation can still make specific predictions: for instance, if the situation represents a circumstance in which there is a cat, it can predict episodes involving 'this cat', which will apply in any particular case to a specific token cat).

[17] This estimate is based on a hierarchical model of time periods currently under development.

[18] We assume a small-world architecture (see e.g. Downes *et al.*, 2012) in which these assemblies communicate as wholes with the wider network, and accordingly we do not adjust the total number of synapses.

```
HumanType=(THuman:3;Animate:2;Human:2)
DogType=(TDog:3;Animate:2)
CatType=(TCat:3;Animate:2)
BirdType=(TBird:3;Animate:2;Flies:2)
CupType=(TCup:3;Inanimate:2;Grabbable:2)
ChairType=(TChair:3;Inanimate:2)
BallType=(TBall:3;Inanimate:2;Grabbable:2)
```

Table A.1: Featural representation in the 'type' area of WM individual. Features starting with 'T' represent type-specific properties, others represent general binary properties. The numbers represent 'strength' of the feature, computationally equivalent to the number of identical units coding the same feature.

9. Acknowledgments

We thank the NZ Marsden Fund for Grant 13-UOO-048, and Slovak VEGA grant 1/0898/14 for their support.

Appendix A. Technical details of the model

Appendix A.1. Representation of objects/actions and episode generation

The following tables show the featural representation of object types in the WM individual medium (Table A.1) and actions in the WM episode medium (Table A.2). Each feature is represented by a separate unit. Table A.3 shows the transcription rules for stochastic generation of episodes.

Appendix A.2. Learning in the candidate WM episode system

During training, the SOM's incoming weights $\vec{w}_i$ are updated using the standard SOM learning rule (Kohonen, 1982)

$$\vec{w}_i(t+1) = \vec{w}_i(t) + \gamma \cdot G(I,i) \cdot [\vec{x}(t) - \vec{w}_i(t)] \tag{A.1}$$

where $\vec{x}(t)$ is the input in the current time step t (the content of WM episode medium), γ is the learning rate, G is a Gaussian neighbourhood function $G(I,i) = \exp(-\|r_I - r_i\|^2/\sigma^2)$ with the width σ, I is the index of the winning neuron, and r_I, r_i are lattice coordinates of neurons I, i.

For the purposes of training, the winner is determined as the unit I with the minimal Euclidean distance between its weight vector $\vec{w}_I(t)$ and the current input $\vec{x}(t)$.

The activity $A_i(t)$ of each unit is then computed as

$$a_i(t) = p_i(t) \cdot \exp(-c \cdot d^2(\vec{w}_i(t), \vec{x}(t))) \tag{A.2}$$

$$A_i(t) = \frac{a_i(t)}{\sum_{j=1}^{N} a_j(t)} \tag{A.3}$$

The Gaussian term $\exp(-c \cdot d^2(\vec{w}_i(t), \vec{x}(t)))$ reflects the likelihood that the current input $\vec{x}(t)$ corresponds to an episode remembered in the weights $\vec{w}_i(t)$ of the i-th unit (the parameter c expresses the sensitivity of the Gaussian), $p_i(t)$ is its frequency-based prior. The activities are then normalized to sum to 1, so the computation follows the Bayesian rule and the overall activity in the candidate WM episode SOM can be interpreted as a probability distribution over possible remembered episodes corresponding to the current WM episode input.[19]

[19] The activities can be (approximately) normalized in a biologically plausible way by receiving a global inhibitory signal proportional to their cumulated activity coming from a special layer—see O'Reilly and Munakata (2000), chapter 3.5. However, we normalize them by simple direct division.

```
Grab=(TGrab:3;Manual:2)
Hit=(THit:3;Manual:2)
Push=(TPush:3;Manual:2)
Pat=(TPat:3;Manual:2)
Stroke=(TStroke:3;Manual:2)
Walk=(TWalk:3;Self-Movement:2)
Run=(TRun:3;Self-Movement:2)
Lie=(TLie:3;Self-Position:2)
Sit=(TSit:3;Self-Position:2)
Sing=(TSing:3;Mouth:2)
See=(TSee:3;Sensory:2)
Snore=(TSnore:3;Physiological:2)
Sneeze=(TSneeze:3;Physiological:2)
Sleep=(TSleep:3;Physiological:2)
Hold=(THold:3;Arms:2)
Hug=(THug:3;Arms:2)
Bite=(TBite:3;Mouth:2)
Kick=(TKick:3;Leg:2)
C+Break=(TBreak:3;Causative:2;Result:2)
C+Stop=(TStop:3;Causative:2;Result:2)
C+Hide=(THide:3;Causative:2;Result:2)
C+Go=(TGo:3;Causative:3;Self-Movement:2)
```

Table A.2: Featural representation of actions. Features starting with 'T' represent type-specific properties, others represent general binary properties. The numbers represent 'strength' of the feature, computationally equivalent to the number of identical localist units coding the same feature.

The SOM weights are initialized to random real numbers between 0 and 3. The learning rate γ is linearly decreasing from 1 to 0.5 during the first 5000 episodes, then stays constant at 0.5. The Gaussian neighbourhood size σ decreased linearly from 20 to 1 during the first 5000 episodes, then to 0.1 during the next 15000 episodes. The sensitivity c of the Gaussian activation term is set to 1. In order to smooth the priors, each scalar weight has an initial value of 1 (i.e., at the beginning we assume a uniform prior $p_i = 1/400$ for each unit).

Appendix A.3. Top-down reconstruction from the active SOM

If we interpret the current activity pattern in the SOM as a probability distribution of possible episodes (as remembered in weights of the SOM's units), we can compute expected values of the episode representation by propagating the activities top-down via weights connecting the cWM-ep medium with the WM episode.[20] The resulting activity $\vec{y}$ in the WM episode is computed as

$$\vec{y} = \sum_{j=1}^{N} A_j(t) \cdot \vec{w}_j(t) \tag{A.4}$$

Appendix A.4. Determining a winner in the cWM-ind system

Let us first assume that the population of N neurons with activities n_i elicited by an unknown stimulus reside in the WM individual layer as a part of it that represents a single property e.g. person, number, place, gender, or colour. Let us further assume that the cWM-ind layer contains representations of K individuals, where the j-th unit remembers the actual value θ_j of the same

[20] Another option would be to combine the weight vectors of just K most active units with an extreme case $K = 1$, i.e. reconstructing only the most probable candidate. However, in all our experiments we combine activities of all 400 units.

```
Episode -> Transitive:86 | Intransitive:24 | Causative:40

Human -> Man | Woman
Dog -> WDog | BDog
Animal -> Dog | Cat
Animate -> Human | Dog | Cat | Bird
AnimateWODogCat -> Human | Bird
Agent -> Human | Dog | Cat
Thing -> Cup | Chair | Ball
Patient -> Human | Dog | Cat | Cup | Chair | Ball | REFL

Transitive -> TrHuman:38 | TrAnimal:48
TrHuman -> TrHumanAnim:14 | TrHumanThing:21 | TrPatM:1 |
           TrPatF:1 | TrStroke:1
TrAnimal -> Animal Patient AnimalTrAction
TrHumanAnim -> Human AnimateWODogCat HumanTrAction
TrHumanThing -> Human Thing HumanTrAction
TrPatM -> Man BDog Pat
TrPatF -> Woman WDog Pat
TrStroke -> Human Cat Stroke
HumanTrAction -> Grab | Hit | Push | See | Hold | Kick | Hug
AnimalTrAction -> Hit | Push | See | Bite

Intransitive -> IntrWOBird:24 | IntrBird
IntrWOBird -> Agent IntrAction
IntrBird -> Bird Sing
IntrAction -> Walk | Lie | Sneeze | Sit | Sleep | Sing | Run |
              Snore

Causative -> CausHumanOnAnimates:2 | CausAnimalOnAnimates:20 |
             CausOnThings:18
CausHumanOnAnimates -> Human Human CausActionOnAnimates
CausAnimalOnAnimates -> Animal Animate CausActionOnAnimates
CausOnThings -> Agent Thing CausActionOnThings
CausActionOnThings -> C+Break | C+Hide
CausActionOnAnimates -> C+Stop | C+Go
```

Table A.3: Transcription rules for stochastic episode generation. '|' character separates alternatives; each alternative is generated with the probability proportional to the number after the colon (if omitted, a default value of 1 is assumed). REFL means reflexive patient (i.e. identical with the agent individual).

property for a particular individual in its weights $\vec{w}^{(j)} = \left(w_i^{(j)}\right) = (f_i(\theta_j))$ for $i = 1..N$. We can then evaluate for each remembered individual how likely it is that it is currently perceived in WM individual. If both n_i and $f_i(\theta_j)$ population codes are normalized (i.e. the sum of n_i for $i = 1..N$ equals to 1, and the same holds for $f_i(\theta_j)$), they can be conceived as probability distributions and the negative log likelihood of θ_j represents their cross-entropy, i.e.

$$\mathrm{NLL}(\theta_j) = -\log L(\theta_j) = -\sum_{i=1}^{N} n_i \log f_i(\theta_j) = -\sum_{i=1}^{N} n_i \log w_i^{(j)} = H(\vec{n}, \vec{w}^{(j)}) \tag{A.5}$$

That means the most likely candidate J is the one with the smallest value of $\mathrm{NLL}(\theta_J)$ (which is always non-negative). However, it is well possible that the currently perceived individual is novel, i.e. none of the remembered ones. To be able to determine that, we need to set a threshold τ such that an individual is considered novel, if $\mathrm{NLL}(\theta_j) > \tau$. The problem is that even for a perfect match $\vec{n} = \vec{w}^{(j)}$ their cross-entropy is not zero, but it is equal to the entropy $H(\vec{n})$. In order to be able to use an absolute threshold τ, it is reasonable to substitute the measure $\mathrm{NLL}(\theta_j)$ with

$$\mathrm{KL}(\vec{n}, \vec{w}^{(j)}) = H(\vec{n}, \vec{w}^{(j)}) - H(\vec{n}) \tag{A.6}$$

which is their Kullback-Leibler (KL) divergence. This measure is zero if and only if $\vec{n} = \vec{w}^{(j)}$.

Because a WM individual does not consist of just a single property, but several population codes—for person, place, number, type, and colour, the KL divergence is determined pairwise between respective areas of WM individual and the weights of each candidate unit in cWM-ind system for each property and then they are averaged to yield a single KL value, which is then compared to the threshold[21] τ to determine whether the currently perceived individual is novel. For more details, see Takac and Knott (2015b).

References

Abney, S., 1996. Statistical methods in linguistics. In: Klavans, J., Resnick, P. (Eds.), The Balancing Act: Combining Symbolic and Statistical Approaches to Language. MIT Press.

Abraham, W., Logan, B., Greenwood, J., Dragunow, M., 2002. Induction and experience-dependent consolidation of stable long-term potentiation lasting months in the hippocampus. Journal of Neuroscience 22, 9626–9634.

Alvarez, G., Cavanagh, P., 2004. The capacity of visual short-term memory is set both by visual information load and by number of objects. Psychological Science 15, 106–111.

Andersen, R., Buneo, C., 2002. Intentional maps in posterior parietal cortex. Annual Review of Neuroscience 25, 189–220.

Anderson, D., Vogel, E., Awh, E., 2011. Precision in visual working memory reaches a stable plateau when individual item limits are exceeded. Journal of Neuroscience 31, 1128–1138.

Averbeck, B., Chafee, M., Crowe, D., Georgopoulos, A., 2002. Parallel processing of serial movements in prefrontal cortex. PNAS 99 (20), 13172–13177.

Averbeck, B., Lee, D., 2007. Prefrontal correlates of memory for sequences. Journal of Neuroscience 27 (9), 2204–2211.

Averbeck, B., Sohn, J., Lee, D., 2006. Activity in prefrontal cortex during dynamic selection of action sequences. Nature Neurosceience 9 (2), 276–282.

[21] Because individuals do not change properties in our experiment, we use the threshold close to zero, namely $\tau = 0.005$.

Awh, E., Barton, B., Vogel, E., 2007. Visual working memory represents a fixed number of items, regardless of complexity. Psychological Science 28, 622–628.

Azevedo, F., Carvalho, L., Grinberg, L., Farfel, J., Ferretti, R., Leite, R., Jacob Filho, W., Lent, R., Herculano-Houzel, S., 2009. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. Journal of Comparative Neurology 513 (5), 532–541.

Baddeley, A., 1992. Working memory. Science 255 (5044), 556–559.

Baddeley, A., 2000. The episodic buffer: A new component of working memory? TICS 4 (11), 417–423.

Baddeley, A., Hitch, G., 1974. Working memory. In: Bower, G. (Ed.), The psychology of Learning and Motivation. Academic Press, pp. 48–79.

Baeg, E., Kim, Y., Huh, K., Mook-Jung, I Kim, H., Jung, M., 2003. Dynamics of population code for working memory in the prefrontal cortex. Neuron 40, 177–188.

Ballard, D., Hayhoe, M., Pook, P., Rao, R., 1997. Deictic codes for the embodiment of cognition. Behavioral and Brain Sciences 20 (4), 723–767.

Barone, P., Joseph, J.-P., 1989. Prefrontal cortex and spatial sequencing in macaque monkey. Experimental Brain Research 78, 447–464.

Barsalou, L., 2008. Grounded cognition. Annual Review of Psychology 59, 617–645.

Barwise, J., Cooper, R., 1981. Generalized quantifiers and natural language. Linguistics and Philosophy 4 (2), 159–219.

Brady, T., Konkle, T., Alvarez, G., 2011. A review of visual memory capacity: Beyond individual items and toward structured representations. Journal of Vision 11 (5), 1–34.

Braver, T., Cohen, J., 2000. On the control of control: The role of dopamine in regulating prefrontal function and working memory. In: Monsell, S., Driver, J. (Eds.), Attention and Performance XVIII: Control of cognitive processes. MIT Press, pp. 713–737.

Buckner, R., Andrews-Hanna, J., Schacter, D., 2008. The brain's default network: Anatomy, function and relevance to disease. Annals of the New York Academy of Sciences 1124, 1–38.

Burgess, N., Hitch, G., 2005. Computational models of working memory: Putting long-term memory into context. Trends in Cognitive Sciences 9 (11), 535–541.

Caza, G., Knott, A., 2012. Pragmatic bootstrapping: A neural network model of vocabulary acquisition. Language Learning and Development 8, 1–23.

Chang, F., 2002. Symbolically speaking: A connectionist model of sentence production. Cognitive Science 26, 609–651.

Charniak, C., 2011. The brain as a statistical inference engine—and you can too. Computational Linguistics 37 (4), 643–655.

Chomsky, N., 1986. Barriers. MIT Press, Cambridge, MA.

Chomsky, N., 1995. The Minimalist Program. MIT Press, Cambridge, MA.

Cisek, P., Kalaska, J., 2005. Neural correlates of reaching decisions in dorsal premotor cortex: Specification of multiple direction choices and final selection of action. Neuron 45, 801–814.

Collins, M., 1996. A new statistical parser based on bigram lexical dependencies. In: Proceedings of the 34th Meeting of the ACL. Santa Cruz, pp. 184–191.

Cooper, R., Shallice, T., 2000. Contention scheduling and the control of routine activities. Cognitive Neuropsychology 17 (4), 297–338.

Corbetta, M., Patel, G., Shulman, G., 2008. The reorienting system of the human brain: from environment to theory of mind. Neuron 3, 306–324.

Cowan, N., 1999. An embedded-process model of working memory. In: Miyake, A., Shah, P. (Eds.), Models of working memory: mechanisms of active maintenance and executive control. Cambridge University Press, Cambridge, UK, pp. 62–101.

Csibra, G., Gergely, G., 2009. Natural pedagogy. Trends in Cognitive Sciences 13 (4), 148–153.

Curtis, C., D'Esposito, M., 2003. Persistent activity in the prefrontal cortex during working memory. Trends in Cognitive Sciences 7 (9), 415–423.

Curtis, C., Rao, V., D'Esposito, M., 2004. Maintenance of spatial and motor codes during oculomotor delayed response tasks. Journal of Neuroscience 24, 3944–3952.

Damasio, A., Damasio, H., 1994. Cortical systems for retrieval of concrete knowledge: The convergence zone framework. In: Koch, C., Davis, J. (Eds.), Large-scale Neuronal Theories of the Brain. MIT Press, Cambridge, MA, pp. 61–74.

DArdenne, K., Eshel, N., Luka, J., Lenartowicz, A., Nystrom, L., Cohen, J., 2012. Role of prefrontal cortex and the midbrain dopamine system in working memory updating. Proceedings of the National Academy of the USA 109, 19900–19909.

Davis, G., Holmes, A., 2005. The capacity of visual short-term memory is not a fixed number of objects. Memory and Cognition 33, 185–195.

Deco, G., Rolls, E., 2008. Neural mechanisms of visual memory: a neurocomputational perspective. In: Luck, S., Hollingworth, A. (Eds.), Visual Memory. Oxford University Press, Oxford, UK, pp. 247–289.

D'Esposito, M., 2007. From cognitive to neural models of working memory. Philosophical Transactions of the Royal Society B 362, 761–772.

Diana, R., Yonelinas, A., Ranganath, C., 2007. Imaging recollection and familiarity in the medial temporal lobe: A three-component model. Trends in Cognitive Sciences 11 (9), 379–386.

Dominey, P., 1997. An anatomically structured sensory-motor sequence learning system displays some general linguistic capacities. Brain and Language 59, 50–75.

Downes, J., Hammond, M., Xydas, D., Spencer, M., Becerra, V., Warwick, K., Whalley, B., Nasuto, S., 2012. Emergence of a small-world functional network in cultured neurons. PLoS Computational Biology 8 (5), e1002522.

Dowty, D., 1991. Thematic proto-roles and argument selection. Language 67 (3), 547–619.

Eichenbaum, H., 2006. Memory binding in hippocampal relational networks. In: Zimmer, H., Mecklinger, A., Lindenburger, U. (Eds.), Handbook of Binding and Memory. Oxford University Press, New York, pp. 25–52.

Eichenbaum, H., Yonelinas, A., Ranganath, C., 2007. The medial temporal lobe and recognition memory. Annual Review of Neuroscience 30, 123–152.

Eliasmith, C., 2013. How to build a brain: An architecture for neurobiological cognition. Oxford University Press, New York.

Elman, J., 1990. Finding structure in time. Cognitive Science 14, 179–211.

Epstein, R., Higgins, S., Jablonski, K., Feiler, A., 2007. Visual scene processing in familiar and unfamiliar environments. Journal of Neurophysiology 97, 3670–3683.

Feigenson, L., 2008. Parallel non-verbal enumeration is constrained by a set-based limit. Cognition 107, 1–18.

Feldman, J., Narayanan, S., 2004. Embodiment in a neural theory of language. Brain and Language 89 (2), 385–392.

Fiebach, C., Friederici, A., Smith, E., Swinney, D., 2007. Lateral inferotemporal cortex maintains conceptual-semantic representations in verbal working memory. Journal of Cognitive Neuroscience 19 (12), 2035–2049.

Fink, G., Halligan, P., Marshall, J., Frith, C., Frackowiack, R., Dolan, R., 1996. Where in the brain does visual attention select the forest and the trees? Nature 382, 626–628.

Fletcher, P., Henson, R., 2001. Frontal lobes and human memory—Insights from functional neuroimaging. Brain 124, 849–881.

Frank, S., Haselager, W., van Rooij, I., 2009. Connectionist semantic systematicity. Cognition 110, 358–379.

Fuentemilla, L., Penny, W., Cashdollar, N., Bunzeck, N., Düzel, E., 2010. Theta-coupled periodic replay in working memory. Current Biology 20, 1–7.

Grosu, A., 1988. On the distribution of genitive phrases in Romanian. Linguistics 26, 931–949.

Hauser, M., Chomsky, N., Fitch, T., 2002. The faculty of language: What is it, who has it, and how did it evolve? Science 298, 1569–1579.

Heim, I., 1982. The semantics of definite and indefinite noun phrases. Ph.D. thesis, University of Massachusetts, distributed by Graduate Linguistic Student Association.

Heim, I., Kratzer, A., 1998. Semantics in Generative Grammar. Blackwell Publishers.

Hickok, G., Bellugi, U., 2010. Neural organization of language: Clues from sign language aphasia. In: Guendouzi, J., Loncke, F., Williams, M. (Eds.), Handbook of Psycholinguistic & Cognitive Processes: Perspectives in Communication Disorders. Taylor and Francis, pp. 685–706.

Howard, M., Kahana, M., 2002. A distributed representation of temporal context. Journal of Mathematical Psychology 46, 269–299.

Itti, L., Arbib, M., 2006. Attention and the minimal subscene. In: Arbib, M. (Ed.), Action to Language via the Mirror Neuron System. Cambridge University Press, Cambridge, UK, pp. 289–346.

Izquierdo, R., Suárez, A., Rigau, G., 2007. Exploring the automatic selection of basic level concepts. In: Proceedings of the International Conference on Recent Advances on Natural Language Processing (RANLP'07). Borovetz, Bulgaria.

Jackendoff, R., 2010. The parallel architecture and its place in cognitive science,. In: Heine, B., Narrog, H. (Eds.), Oxford Handbook of Linguistic Analysis. Oxford University Press, Oxford, pp. 583–605.

Jazayeri, M., Movshon, A., 2006. Optimal representation of sensory information by neural populations. Nature Neuroscience 9 (5), 690–696.

Kalchbrenner, N., Grefenstette, E., Blunsom, P., 2014. A convolutional neural network for modelling sentences. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

Kamp, H., 1981. A theory of truth and semantic representation. In: Groenendijk, J., Janssen, T., Stokhof, M. (Eds.), Formal Methods in the Study of Language. Mathematical Center Tract 135, Amsterdam, pp. 277–322.

Kemmerer, D., 2012. The cross-linguistic prevalence of sov and svo word orders reflects the sequential and hierarchical representation of action in brocas area. Language and Linguistics Compass 6 (1), 50–66.

Kiani, R., Shadlen, M., 2009. Representation of confidence associated with a decision by neurons in the parietal cortex. Science 324, 759–764.

Killworth, P., Johnsen, E., Bernard, H., Shelley, G., McCarty, C., 1990. Estimating the size of personal networks. Social Networks 12, 289–312.

Knott, A., 2012. Sensorimotor Cognition and Natural Language Syntax. MIT Press, Cambridge, MA.

Knott, A., 2014a. How infants learn word meanings and propositional attitudes: a neural network model. In: Hung, T.-W. (Ed.), Language and Action. Springer, Berlin/Heidelberg, pp. 107–124.

Knott, A., 2014b. Syntactic structures as descriptions of sensorimotor processes. Biolinguistics 8, 1–52.

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

Kullback, S., Leibler, R., 1951. On information and sufficiency. Annals of Mathematical Statistics 22 (1), 79–86.

Lee, H., Simpson, G., Logothetis, N., Rainer, G., 2005. Phase locking of single neuron activity to theta oscillations during working memory in monkey extrastriate visual cortex. Neuron 45, 147–156.

Lee-Hand, J., Knott, A., 2015. A neural network model of causative actions. Frontiers in Neurorobotics 9, Article 4.

Lenat, D., 1995. Cyc: A large-scale investment in knowledge infrastructure. Communications of the ACM 38, 33–38.

Leslie, S.-J., Gelman, S., 2012. Quantified statements are recalled as generics: Evidence from preschool children and adults. Cognitive Psychology 64, 186–214.

Lisman, J., 1999. Relating hippocampal circuitry viewpoint to function: Recall of memory sequences by reciprocal dentate-CA3 interactions. Neuron 22, 233–242.

Luck, S., Vogel, E., 1997. The capacity of visual working memory for features and conjunctions. Nature 390, 279–281.

Manns, J., Eichenbaum, H., 2009. A cognitive map for object memory in the hippocampus. Learning and Memory 16 (10), 616–624.

Mante, V., Sussillo, D., Shenoy, K., Newsome, W., 2013. Context-dependent computation by recurrent dynamics in prefrontal cortex. Nature 503, 78–84.

Martin, R., Crowther, J., Knight, M., Tamborello, F., Yang, C.-L., 2010. Planning in sentence production: Evidence for the phrase as a default planning scope. Cognition 116, 177–192.

Martin, R., Freedman, M., 2001. Short-term retention of lexical-semantic representations: Implications for speech production. Memory 9, 261–280.

Martin, R., Shelton, J., Yaffee, L., 1994. Language processing and working memory: Neuropsychological evidence for separate phonological and semantic capacities. Journal of Memory and Language 33, 83–111.

Martin, R., Yan, H., Schnur, T., 2014. Working memory and planning during sentence production. Acta Psychologica 152, 120–132.

Mayberry, M., Miikkulainen, R., 2008. Incremental nonmonotonic sentence interpretation through semantic self-organization. Tech. Rep. AI08-12, Department of Computer Sciences, The University of Texas at Austin.

McClelland, J., St. John, M., Taraban, R., 1989. Sentence comprehension: A parallel distributed processing approach. Language and Cognitive Processes 4 (3–4), 287–335.

Mendoza-Halliday, D., Torres, S., Martinez-Trujillo, J., 201. Sharp emergence of feature-selective sustained activity along the dorsal visual pathway. Nature Neuroscience 17 (9), 1255–1265.

Meyers, E., Freedman, D., Kreiman, G., Miller, E., Poggio, T., 2008. Dynamic population coding of category information in inferior temporal and prefrontal cortex. Journal of Neurophysiology 2008 (100), 1407–1419.

Miikkulainen, R., 1996. Subsymbolic case-role analysis of sentences with embedded clauses. Cognitive Science 20, 47–73.

Miller, E., Cohen, J., 2001. An integrative theory of prefrontal cortex function. Annual Review of Neuroscience 24, 167–202.

Moore, T., Armstrong, K., 2003. Selective gating of visual signals by microstimulation of frontal cortex. Nature 421 (6921), 370–373.

Moscovitch, M., Kapur, S., Köhler, S., Houle, S., 1995. Distinct neural correlates of visual long-term memory for spatial location and object identity: A positron emission tomography study in humans. Proceedings of the National Academy of the USA 92, 3721–3725.

Müller, N., Kleinschmidt, A., 2007. Temporal dynamics of the attentional spotlight: Neuronal correlates of attentional capture and inhibition of return in early visual cortex. Journal of Cognitive Neuroscience 19 (4), 587–593.

Muller, R., Kubie, J., 1987. The effects of changes in the environment on the spatial firing of hippocampal complex-spike cells. Journal of Neuroscience 7, 1951–1968.

Nakamura, K., Kubota, K., 1995. Mnemonic firing of neurons in the monkey temporal pole during a visual recognition memory task. Journal of Neurophysiology 74, 162–178.

Navon, D., 1977. Forest before trees: The precedence of global features in visual perception. Cognitive Psychology 9, 353–383.

Nieder, A., Miller, E., 2004. A parieto-frontal network for visual numerical information in the monkey. Proceedings of the National Academy of Sciences of the United States of America 101 (19), 7457–7462.

O'Reilly, R., Frank, M., 2006. Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. Neural Computation 18, 283–328.

O'Reilly, R., Munakata, Y., 2000. Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain. MIT Press, Cambridge, MA.

Papagno, C., 2011. Naming and the role of the uncinate fasciculus in language function. Current Neurology and Neuroscience Reports 11 (6), 553–559.

Pickering, M., Traxler, M., 1998. Plausibility and recovery from garden paths: An eye-tracking study. Journal of Experimental Psychology: Learning, Memory and Cognition 24 (4), 940–961.

Poch, C., Fuentemilla, L., Barnes, G., Düzel, E., 2011. Hippocampal theta-phase modulation of replay correlates with configural-relational short-term memory performance. Journal of Neuroscience 31 (19), 7038 7042.

Pollack, J., 1990. Recursive distributed representations. Artificial Intelligence 46 (1–2), 77–105.

Pollard, C., Sag, I., 1994. Head-Driven Phrase Structure Grammar. University of Chicago Press.

Posner, M., Rafal, R., Choate, L., Vaughn, J., 1984. Inhibition of return: Neural basis and function. Cognitive Neuropsychology 2, 211–228.

Potter, M., Lombardi, L., 1990. Regeneration in the short-term recall of sentences. Journal of Memory and Language 29, 633–654.

Potter, M., Lombardi, L., 1998. Syntactic priming in immediate recall of sentences. Journal of Memory and Language 38, 265–282.

Pouget, A., Beck, J., Ma, W.-J., Latham, P., 2013. Probabilistic brains: knowns and unknowns. Nature Neuroscience 16 (9), 1170–1178.

Quiroga, R., Reddy, L., Koch, C., Fried, I., 2005. Invariant visual representation by single neurons in the human brain. Nature 435, 1102–1107.

Rainer, G., Asaad, W., Miller, E., 1998. Memory fields of neurons in the primate prefrontal cortex. Proceedings of the National Academy of Sciences of the United States of America 95 (25), 15008–15013.

Ranganath, C., 2010. A unified framework for the functional organization of the medial temporal lobes and the phenomenology of episodic memory. Hippocampus 20, 1263–1290.

Rao, S., Rainer, G., Miller, E., 1997. Integration of what and where in the primate prefrontal cortex. Science 276 (5313), 821–824.

Reilly, R., 1992. Connectionist technique for online parsing. Network 3 (1), 1–37.

Ritter, E., 1991. Two functional categories in noun phrases: Evidence from Modern Hebrew. In: Rothstein, S. (Ed.), Syntax and Semantics 25: Perspectives in Modern Phrase Structure. Academic Press, New York, pp. 37–62.

Ro, T., Farné, A., Chang, E., 2003. Inhibition of return and the human frontal eye fields. Experimental Brain Research 150, 290–296.

Robertson, L., Lamb, M., Knight, R., 1988. Effects of lesions of temporo-parietal junction on perceptual and attentional processing in humans. Journal of Neuroscience 8, 3757–3769.

Rosenblatt, F., 1962. Principles Of Neurodynamics. Spartan Books, Washington.

Schacter, D., Norman, K., Koutstaal, W., 1998. The cognitive neuroscience of constructive memory. Annual Review of Psychology 49, 289–318.

Schluppeck, D., Curtis, C., Glimcher, P., Heeger, D., 2006. Sustained activity in topographic areas of human posterior parietal cortex during memory-guided saccades. Journal of Neuroscience 26 (19), 5098–5108.

Schmidt, D., Krause, B., Weiss, P., Fink, G., Shah, N., Amorim, M., Müller, H., Berthoz, A., 2007. Visuospatial working memory and changes of the point of view in 3D space. NeuroImage 36, 955–968.

Schweppe, J., Rummer, R., 2007. Shared representations in language processing and verbal short-term memory: The case of grammatical gender. Journal of Memory and Language 56, 336–356.

Schweppe, J., Rummer, R., Bormann, T., Martin, R., 2011. Semantic and phonological information in sentence recall: Converging psycholinguistic and neuropsychological evidence. Cognitive Neuropsychology 28 (8), 521–545.

Semenza, C., 2006. Retrieval pathways for common and proper names. Cortex 42 (6), 884–891.

Semenza, C., 2011. Naming with proper names: The left temporal pole theory. Behavioural Neurology 24, 277–284.

Shadlen, M., Newsome, W., 2001. Neural basis of a perceptual decision in the parietal cortex (area lip) of the rhesus monkey. Journal of Neurophysiology 86 (4), 1916–1936.

Shetreet, E., Palti, D., Friedmann, N., Hadaraf, U., 2011. Cortical representation of verb processing in sentence comprehension: Number of complements, subcategorization, and thematic frames. Cerebral Cortex 17, 1958–1969.

Shima, K., Isoda, M., Mushiake, H., Tanji, J., 2007. Categorization of behavioural sequences in the prefrontal cortex. Nature 445, 315–318.

Shivde, G., Anderson, M., 2011. On the existence of semantic working memory: Evidence for direct semantic maintenance. Journal of Experimental Psychology: Learning, Memory, and Cognition 37 (6), 1342–1370.

Siegel, M., Warden, M., Miller, E., 2009. Phase-dependent neuronal coding of objects in short-term memory. Proceedings of the National Academy of Sciences of the USA 106 (50), 21341–21346.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. arXiv Technical Report.

Slevc, J., 2011. Saying whats on your mind: Working memory effects on sentence production. Journal of Experimental Psychology: Learning, Memory, and Cognition 37 (6), 1503–1514.

Squire, L., 1987. Memory and brain. Oxford University Press, New York.

Sreenivasan, K., Vytlacil, J., DEsposito, M., 2014. Distributed and dynamic storage of working memory stimulus information in extrastriate cortex. Journal of Cognitive Neuroscience 26 (5), 1141–1153.

Sridharan, D., Levitin, D., Chafe, C., Berger, J., Menon, V., 2007. Neural dynamics of event segmentation in music: Converging evidence for dissociable ventral and dorsal networks. Neuron 55, 521–532.

Stewart, T., Eliasmith, C., 2012. Compositionality and biologically plausible models. In: Werning, M., Hinzen, W. (Eds.), The Oxford Handbook of Compositionality. Oxford University Press, New York.

Stokes, M., 2015. 'Activity-silent' working memory in prefrontal cortex: a dynamic coding framework. Trends in Cognitive Sciences 19 (7), 394–405.

Strickert, M., Hammer, B., 2005. Merge SOM for temporal data. Neurocomputing 64, 39–71.

Sutskever, I., Vinyals, O., Le, Q., 2014. Sequence to sequence learning with neural networks. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N., Weinberger, K. (Eds.), Advances in Neural Information Processing Systems 27. Curran Associates, Inc., pp. 3104–3112.

Takac, M., Benuskova, L., Knott, A., 2012. Mapping sensorimotor sequences to word sequences: A connectionist model of language acquisition and sentence generation. Cognition 125, 288–308.

Takac, M., Knott, A., 2015a. A neural network model of episode representations in working memory. Cognitive Computation 7 (5), 509–525.

Takac, M., Knott, A., 2015b. A simulationist model of episode representations in working memory: Technical appendix. Tech. Rep. OUCS-2015-01, Dept of Computer Science, University of Otago.

Takahashi, E., Ohki, K., Kim, D.-S., 2013. Dissociation and convergence of the dorsal and ventral visual working memory streams in the human prefrontal cortex. NeuroImage 65, 488–498.

Takeda, M., Naya, Y., Fujimichi, Y., Takeuchi, D., Y, M., 2005. Active maintenance of associative mnemonic signal in monkey inferior temporal cortex. Neuron 48, 839–848.

Tang, K., Nevins, A., 2013. Quantifying the diachronic productivity of irregular verbal patterns in Romance. UCL Working Papers in Linguistics 25, 289–308.

Tang, Y., Nyengaard, J., De Groot, D., Gundersen, H.-J., 2001. Total regional and global number of synapses in the human brain neocortex. Synapse 41 (3), 258273.

Taraldsen, T., 1990. D-projections and N-projections in Norwegian. In: Nespor, M., Mascaró, J. (Eds.), Grammar in progress. Foris, Dordrecht.

Tomasello, M., 2003. Constructing a Language: A Usage-Based Theory of Language Acquisition. Harvard University Press, Cambridge, MA.

Traxler, M., 2011. Introduction to Psycholinguistics: Understanding Language Science. Wiley-Blackwell, Hoboken, NJ.

Traxler, M., 2014. Trends in syntactic parsing: anticipation, bayesian estimation, and good-enough parsing. Trends in Cognitive Sciences 18 (11), 605–611.

Treisman, A., Gelade, G., 1980. A feature integration theory of attention. Cognitive Psychology 12, 97–136.

Tulving, E., 1983. Elements of Episodic Memory. Oxford University Press, New York.

Turk-Browne, N., Simon, M., Sederberg, P., 2012. Scene representations in parahippocampal cortex depend on temporal context. Journal of Neuroscience 32 (21), 7202–7207.

Tyler, L., Moss, H., 2001. Towards a distributed account of conceptual knowledge. Trends in Cognitive Sciences 5 (6), 244–252.

van der Velde, F., de Kamps, M., 2006. Neural blackboard architectures of combinatorial structures in cognition. Behavioral and Brain Sciences 29, 37–108.

Wallenstein, G., Eichenbaum, H., Hasselmo, M., 1998. The hippocampus as an associator of discontiguous events. Trends in Neurosciences 21, 317–323.

Wallentin, M., Weed, E., Østergaard, L., Mouridsen, K., Roepstorff, A., 2008. Accessing the mental space—spatial working memory processes for language and vision overlap in precuneus. Human Brain Mapping 29, 524–532.

Walles, H., Knott, A., Robins, A., 2008. A model of cardinality blindness in inferotemporal cortex. Biological Cybernetics 98 (5), 427–437.

Walles, H., Robins, A., Knott, A., 2014. A perceptually grounded model of the singular-plural distinction. Language and Cognition 6, 1–43.

Wallis, J., Anderson, K., Miller, E., 2001. Single neurons in prefrontal cortex encode abstract rules. Nature 411, 953–956.

Warden, M., Miller, E., 2007. The representation of multiple objects in prefrontal neuronal delay activity. Cerebral Cortex 17, i41–i50.

Warden, M., Miller, E., 2010. Task-dependent changes in short-term memory in the prefrontal cortex. Journal of Neuroscience 30 (47), 15801–15810.

Webb, A., Knott, A., MacAskill, M., 2010. Eye movements during transitive action observation have sequential structure. Acta Psychologica 133, 51–56.

Wei, Z., Wang, X.-J., Wang, D.-H., 2012. From distributed resources to limited slots in multiple-item working memory: A spiking network model with normalization. Journal of Neuroscience 32 (33), 11228–11240.

Wilken, P., Ma, P., 2004. A detection theory account of change detection. Journal of Vision 4 (12), 1120–1135.

Wilson, F., Scalaidhe, S., Goldman-Rakic, P., 1993. Dissociation of object and spatial processing domains in primate prefrontal cortex. Science 260 (5116), 1955–1958.

Yee, E., Chrysikou, E., Thompson-Schill, S., 2013. The cognitive neuroscience of semantic memory. In: Ochsner, K., Kosslyn, S. (Eds.), The Oxford Handbook of Cognitive Neuroscience Volume 1: Core Topics. Oxford University Press, Oxford, UK, pp. 353–374.

Zacks, J., Braver, T., Sheridan, M., Donaldson, D., Snyder, A., Ollinger, J., Buckner, R., Raichle, M., 2001. Human brain activity time-locked to perceptual event boundaries. Nature Neuroscience 4 (6).

Zacks, J., Speer, N., Swallow, K., Braver, T., Reynolds, J., 2007. Event perception: A mind-brain perspective. Psychological Bulletin 133 (2), 273–293.

Zamparelli, R., 2000. Layers in the Determiner Phrase. Garland, New York.

Zhang, Y., Meyers, E., Bichot, N., Serre, T., Poggio, T., Desimone, R., 2011. Object decoding with attention in inferior temporal cortex. Proceedings of the National Academy of Sciences of the USA 108 (21), 8850–8855.