Department of Computer Science, University of Otago

Technical Report OUCS-2016-04

A simulationist model of episode representations in working memory: syntactic interpretation, nested episodes and storage requirements

Authors:

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

Martin Takac
Centre for Cognitive Science, Comenius University, Slovakia

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: syntactic interpretation, nested episodes and storage requirements

Martin Takac[1,2] and Alistair Knott[1]
[1]Department of Computer Science, University of Otago, New Zealand
[2]Centre for Cognitive Science, Comenius University, Slovakia

November 22, 2016

Abstract

This report supplements Takac and Knott's (2016b) model of working memory (WM) for episodes and individuals (henceforth 'the main paper'). In Section 1 we introduce our interpretation of syntactic heads in relation to this WM model in more detail. In Section 2 we present some ideas about how the WM model can be extended to handle nested episodes. In Section 3 we discuss the storage requirements of the model, and assess whether it can be extended to represent a realistic number of episodes and individuals.

1 Applications of the WM model in a model of syntax

Our model of WM is intended to model certain aspects of language processing. In this section, we introduce in more detail our proposal that the WM model can 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)[1] 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 1.1; in Section 1.2 we propose that aspects of this account reflect structures in semantic WM, as it is conceived in our model.

1.1 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 *in-*

[1]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).

nate 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 1a, and the LF structure of its object noun phrase *the big cats* is shown in Figure 1b. 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 1a) and a noun (N) heads/projects an NP (see Figure 1b). 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.[2] 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 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 1a, 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.[3] 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 1b (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

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

[3]'Quantifying' determiners (e.g. *all*, *most*) also introduce referents, but we will not discuss these determiners here.

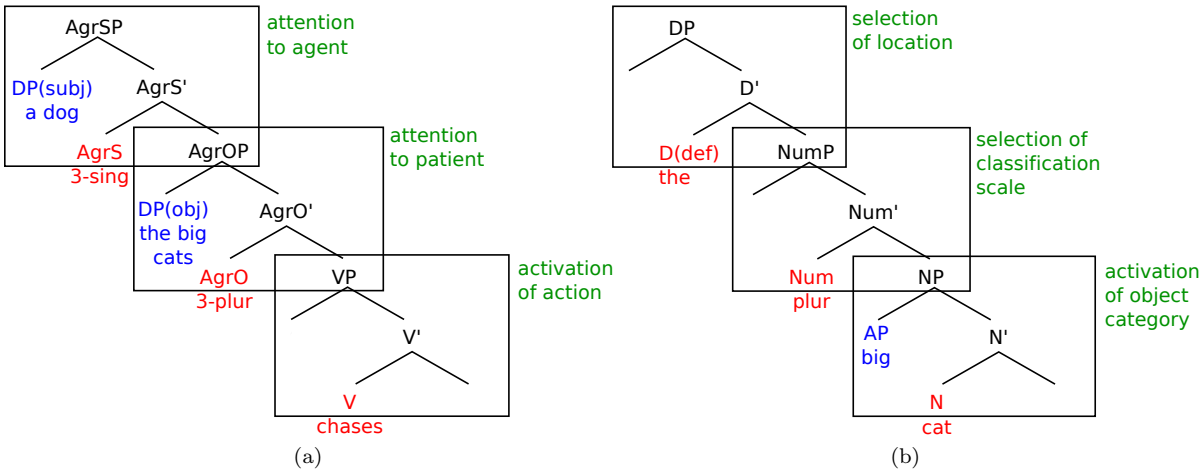

Figure 1: (a) LF structure of a transitive clause. (b) LF structure of a determiner phrase.

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 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.

1.2 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 our paper 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 1a. 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 1b. The highest XP (DP) selects a referential ele-

ment 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; 2015).

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, as shown in Figure 2, 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; 2015) 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 1a. 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 1a. 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 1b). There are also

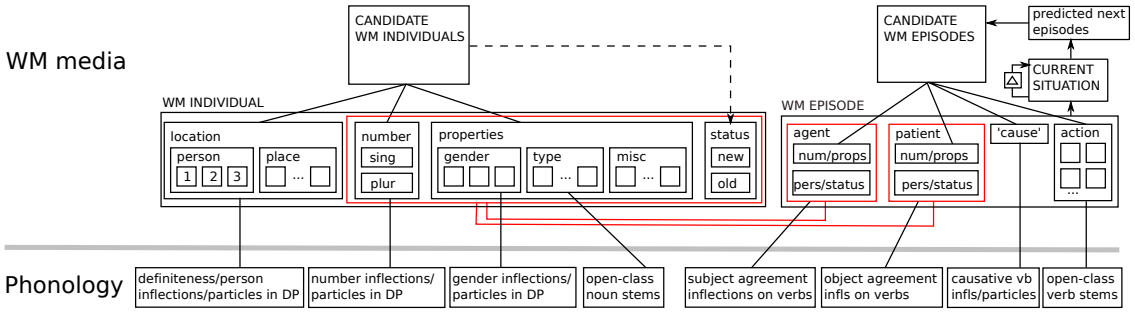

Figure 2: Interfaces between the semantic WM system and surface phonology

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 1b). 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 1.1 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 2). 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 2). 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 the main paper.)

2 Semantic representations containing nested propositions

Our model already accounts for some types of hierarchical structure in syntactic representations. Through its account of the interface be-

tween WM individuals and WM episodes, it accounts for how DPs, with their own internal structure, can appear at positions within clauses. 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 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.

2.1 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. 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 a specified interlocutor)', 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 com-

plete 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 (e.g. van der Velde and de Kamps, 2006; Rohde, 2002; Stewart and Eliasmith, 2012): 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.

2.2 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 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 6 of the paper, 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,[4] 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 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 Section 2.1. 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.

[4]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.

2.3 Relative clauses

As discussed in Section 1.2, 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 semantic memory and DP planning, and then consider the relative clause mechanism specifically.

2.3.1 A sketched model of semantic memory

We will first sketch how a system of semantic memory can be added to our WM model. Semantic memory is a form of long-term memory (LTM): we envisage it as a medium that holds explicit knowledge about the properties of individuals. In our proposed model, the key new 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 can activate an LTM individual (if there is one whose properties match well enough). For an initial implementation of LTM individuals, see Takac and Knott (2016a).

LTM individuals in language 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 1.2. Another method is to generate a proper noun directly from the active LTM individual (if an associated proper noun is known).

Episode-based properties 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' 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-based property from an actual epsiode, because they are stative facts, rather than episodes—and yet episode-based properties must ultimately be identified by experiencing actual episodes. Our model of WM suggests an interesting model of episode-based properties in semantic LTM, which we will briefly outline.

While the representation of episode-based 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-based 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. The new type of link allows the semantic memory system to record episode-based 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 our account of relative clauses.

2.3.2 A sketched model of DP planning for referential DPs

We assume that the speaker can choose to report a referent using the WM individuals system, or alternatively by identifying an LTM individual that is known to the hearer. 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.[5]

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 1.

2.3.3 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 static 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*

[5] 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.

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 2.3.1, 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).[6] 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-based property 'x chased Mary' in the candidate WM 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.

3 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).[7] 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 2, 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

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

[7] 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.

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 2.3.1 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 6.1 of the main paper), 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.[8] If 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.[9]

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 is represented in the MSOM as the weights of the winning unit in the previous situation, as discussed in Section 3 of the main paper, 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.[10] 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.[11] 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 has reasonable complexity, given that it models substantial parts of LTM as well as semantic WM.

4 Acknowledgments

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

References

Abney, S. (1996). Statistical methods in linguistics. In J. Klavans and P. Resnick, edi-

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

[9] 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).

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

[11] 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.

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

tors, *The Balancing Act: Combining Symbolic and Statistical Approaches to Language*. MIT Press.

Azevedo, F., Carvalho, L., Grinberg, L., Farfel, J., Ferretti, R., Leite, R., Jacob Filho, W., Lent, R., and 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.

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

Buckner, R., Andrews-Hanna, J., and 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.

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

Chomsky, N. (1986). *Barriers*. MIT Press, Cambridge, MA.

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

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

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

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

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

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

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

Hauser, M., Chomsky, N., and 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. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Blackwell Publishers.

Izquierdo, R., Suárez, A., and 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.

Kalchbrenner, N., Grefenstette, E., and 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 J. Groenendijk, T. Janssen, and M. Stokhof, editors, *Formal Methods in the Study of Language*, pages 277–322. Mathematical Center Tract 135, Amsterdam.

Killworth, P., Johnsen, E., Bernard, H., Shelley, G., and 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 T.-W. Hung, editor, *Language and Action*, pages 107–124. Springer, Berlin/Heidelberg.

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

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

Mayberry, M. and Miikkulainen, R. (2008). Incremental nonmonotonic sentence interpretation through semantic self-organization. Technical Report AI08-12, Department of Computer Sciences, The University of Texas at Austin.

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

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

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

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

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

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 S. Rothstein, editor, *Syntax and Semantics 25: Perspectives in Modern Phrase Structure*, pages 37–62. Academic Press, New York.

Rohde, D. (2002). *A Connectionist Model of Sentence Comprehension and Production*. Ph.D. thesis, School of Computer Science, Carnegie Mellon University.

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.

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

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

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

Takac, M. and Knott, A. (2016a). Mechanisms for storing and accessing event representations in episodic memory, and their expression in language: a neural network model. In *Proceedings of the 38th Annual Meeting of the Cognitive Science Society (CogSci)*, pages 532–537.

Takac, M. and Knott, A. (2016b). A simulationist model of episode representations in working memory. Submitted manuscript.

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

Tang, K. and 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., and 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 M. Nespor and J. Mascaró, editors, *Grammar in progress*. Foris, Dordrecht.

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

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

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

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

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