Department of Computer Science, University of Otago

Technical Report OUCS-2015-01

A simulationist model of episode representations in working memory: Technical appendix

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: Technical appendix

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

December 3, 2015

Abstract

In this report we present technical details of a novel connectionist model of episode representations in working memory, including details of encoding, training input and training regime, numerical parameters and results of simulations of different experimental tasks.

1 Introduction

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 (SM) routines with well-defined sequential structure. It differs from our previous model of WM for episodes (Takac and Knott, 2015) in many aspects:

1. Not only episodes, but also individuals are stored in semantic WM as prepared sensorimotor routines that can be internally replayed.

2. We distinguish between object types and tokens, i.e. the system can represent richer properties of individual token objects, including their current location.

3. We added a system for representing recently seen individuals, which supports discourse referencing to previous individuals (expressed in language with pronouns or definite articles).

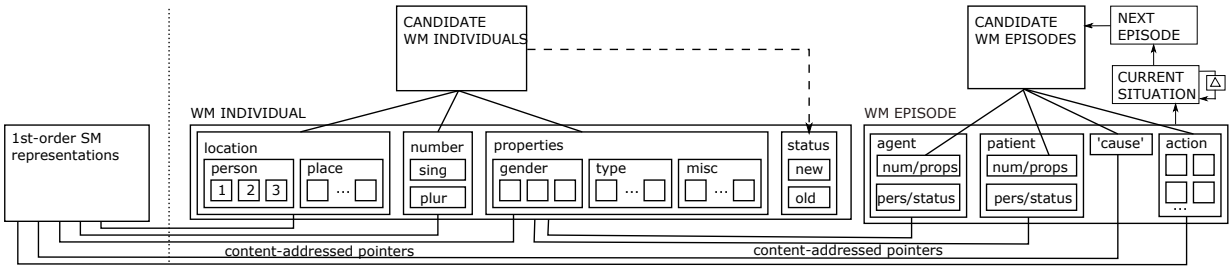

Figure 1: Architecture of the model of WM individuals and WM episodes.

4. The model provides a novel account of binding: specifically, of how individuals are bound to semantic roles in episodes (such as agent and patient). Agent and patient are represented in separate places, as content-addressed pointers to a common medium. In this way, cross-role generalization is possible, while the ability to represent different expectations about agents and patients is retained.

5. The model is extended with a situation update system able to predict a distribution of possible next episodes.

This report serves as supplementary material for a complete article (Takac and Knott, 2016), hence it is restricted to technical details of the model.

2 Architecture

Our model is illustrated in Figure 1. The WM media are above the 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**. This 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. 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). 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 Takac and Knott, 2015).

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. 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' (see Section 5 for details). 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 1. (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 and 'newness'.) 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 (see experiments in Section 11.1).

Once the WM episode medium is filled, it serves as input to the candidate WM episodes medium implemented as a self-organizing map or SOM (Kohonen, 1982). The primary function of this medium is to hold representations of multiple episodes active in parallel for various purposes, such as representing a probability distribution over remembered episodes consistent with the current content of the WM episode medium, or possible episodes typically following in the current situation context (as predicted by the Next episode prediction system, see below).

WM episode feeds another medium: the **current situation**. This is implemented as a recurrent self-organising map: specifically, an MSOM (Strickert and Hammer, 2005). Thanks to recurrent connections, units in the MSOM are organised to represent episodes in particular contexts, i.e. as whole situations. The MSOM activity represents a probability distribution of remembered situations consistent with the current WM episode, and serves as input to another medium, the **predicted next episode** layer, which is isomorphic to the candidate WM episodes layer. The current situation layer is linked to the predicted next episode layer via a fully connected network: during training this network learns to map the current situation to the content of the candidate WM episode system at the next time step: that is, it learns transition probabilities between pairs of successive episodes. The trained network generates a distribution of predicted next episodes. This distribution is conveyed to the candidate WM episodes layer via 1:1 connections. The distrinction in the candidate WM episodes layer supplies top-down expectations about the content of the next WM episode, which can be further propagated to the WM individual medium.

Now we will revisit individual components of the model in more detail.

3 First-order SM media

First-order SM media (below the grey line in Figure 1) are isomorphic to the WM individual and WM episode areas they are linked to. Perception of an episode (see Section 10) is simulated by generating a sequence of activity patterns in the first-order SM media that in turn generate identical activity patterns in the WM individual and WM episode areas. A particular way of encoding actions, objects, and their properties is specified below.

4 WM Individual

The WM individual layer consists of 69 units in areas encoding location, number, properties and status. Location consists of a set of 3 localist units for person (1, 2, 3) and further 36 units together coding a spatial position (place, see below). Number is coded by 2 units (for Sg, Pl). Properties area includes a set of 3 units for grammatical gender (Male, Female, Neutral), 12 units for type-specific features—animacy, type (person, dog, cat, bird,

cup, ball, chair) and hierarchy-imposing properties (is-human, can-fly, can-be-grabbed, see Table 3) and 11 units together encoding a colour of the object (see below). The status is coded by 2 units (new, old). Each set of units can either encode a single property unambiguously, or a probability distribution over properties.

Spatial position of objects (situated on a 100 x 100 grid) is coded by a population of 6x6 neurons with Gaussian receptive fields evenly covering the grid. The 'misc' area in Figure 1 represents a single property—colour. 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 4). 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).

5 Candidate WM individuals

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).

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. 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, as derived by Jazayeri and Movshon (2006). We outline the derivation below.

Let us assume an unknown sensory stimulus activates a population of N neurons with broad tuning curves f_i so that each neuron responds with activity n_i. Then the log likelihood that the unknown stimulus is equal to a particular value θ can be computed as a simple weighted sum of the responses of the neurons, where the activity of each neuron is weighted by the log of its own tuning function in θ:

$$\log L(\theta) = \sum_{i=1}^{N} n_i \log f_i(\theta) \tag{1}$$

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 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 use Equation 1 to 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 normalised (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 form of the expression in Equation 1 represents their cross-entropy, i.e.

$$\text{NLL}(\theta_j) = -\log L(\theta_j) = -\sum_{i=1}^{N} n_i \log f_i(\theta_j) = -\sum_{i=1}^{N} n_i \log w_i^{(j)} = H(\vec{n}, \vec{w}^{(j)}) \tag{2}$$

That means the most likely candidate J is the one with the smallest value of $\text{NLL}(\theta_J)$ (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 $\text{NLL}(\theta_j) > \tau$. 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 $\text{NLL}(\theta_j)$ with

$$\text{KL}(\vec{n}, \vec{w}^{(j)}) = H(\vec{n}, \vec{w}^{(j)}) - H(\vec{n}) \tag{3}$$

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

Now 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[1] τ to determine whether the currently perceived individual is novel.

If a likely-enough existing candidate is returned, it is updated (its weight values are replaced by the corresponding values in WM individual) and the WM individual's status is set to 'old'; otherwise a new entry in the layer is created (again with weights copied from the WM individual) and the WM individual's status is set to 'new'. Candidate units that have not been updated for the last 20 episodes are then 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.

As we have shown, KL divergence helps us to determine which of the remembered individuals is likely being perceived right now. However, the cWM-ind system serves yet another function. As we describe in Section 7.1, the candidate WM episode medium can generate top-down expectations in the agent or patient area of the WM episode medium that can be further copied to WM individual. Now the content of WM individual does not represent an *actually* perceived individual, but rather a blend of *expectations*, e.g. that an individual should be either cat or dog and either black or white. To evaluate which of the actual individuals remembered in the cWM-ind system matches this expectation best, we again use the average pairwise KL divergence, but now with swapped arguments, i.e. as $\text{KL}(\vec{w}^{(j)}, \vec{n})$, where $\vec{w}^{(j)}$ are the weights of j-th unit in cWM-ind system and $\vec{n}$ is the

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

content of the WM individual copied from the WM episode expectation. Units are made active inversely proportional to their KL value:

$$a_j = \exp(-c \cdot \mathrm{KL}(\vec{w}^{(j)}, \vec{n}) \tag{4}$$

where $c = 3$ is the sensitivity. Activities are then normalised so that the sum of their activities is 1:

$$A_j = \frac{a_j}{\sum_{k=1}^{K} a_k} \tag{5}$$

Now we refine the expectation in the WM individual by top-down propagating the combination of weight vectors of the active cWM-ind units mixed proportionally to their activities:

$$\vec{y} = \sum_{j=1}^{K} A_j \cdot \vec{w}^{(j)} \tag{6}$$

Refining the expectation in this way is important because the expectation copied from the WM episode medium does not contain place information—the expected location is supplied from cWM-ind medium. We test this experimentally—see Section 11.2.3.

6 WM episode

Agent and patient parts of the WM episode are isomorphic to the relevant parts of the 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 the person and status areas of the WM individual (the place representation from WM individual is not copied). 'Cause' is a single neuron that is 'on' for causative actions and 'off' otherwise. The 'action' area consists of 22 localist units for actions (see Table 1) and 11 units for their distributed featural codes.

7 Candidate WM episodes system

The **candidate WM episodes (cWM-ep)** medium is a self-organising map or **SOM** that takes input from the WM episode medium and is trained on episodes represented in this medium. It learns to represent episodes as localist units, organised so that similar episodes are close together in the map. Each localist unit can encode (in its incoming weights) a particular combination of representations in the agent, patient and action media, and thus can represent a complete episode by itself. We use a modified SOM that can also keep track of frequencies of episodes (by recording how often a particular unit became a winner—see below).

The SOM we use in our experiments has N=400 (20x20) units. Each unit has a 90-dimensional vector $\vec{w}_i$ of incoming weights from the WM episode medium and a scalar

```
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 1: 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.

weight p_i reflecting the relative frequency of 'hits' for this unit, i.e. the proportion of times this unit was the most active unit.[2]

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

$$\vec{w}_i(t+1) = \vec{w}_i(t) + \gamma \cdot G(I,i) \cdot [\vec{x}(t) - \vec{w}_i(t)] \tag{7}$$

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 $G(I,i) = \exp(-\left\|r_I - r_i\right\|^2/\sigma^2)$ 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

$$a_i(t) = p_i(t) \cdot \exp(-c \cdot d^2(\vec{w}_i(t), \vec{x}(t))) \tag{8}$$

$$A_i(t) = \frac{a_i(t)}{\sum_{j=1}^{N} a_j(t)} \tag{9}$$

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 normalised 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.[3]

The SOM weights are initialised to random real numbers between 0 and 3. The learning rate γ decreases linearly from 1 to 0.5 during the first 5000 episodes, then stays constant at 0.5. The Gaussian neighbourhood size σ decreases 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).

[2] In fact, a scalar weight is increased each time a unit becomes the winner, hence it records the absolute frequency and the weights are normalised (i.e. divided by a scalar weight of a special unit increased each time any unit becomes a winner) when used. A more biologically plausible method where weights could not grow indefinitely would be to increase a scalar weight of the winning unit by a small amount and renormalise the scalar weights of all the units right then so that their sum is 1 in each step. In this way, the scalar weight would be biased towards more recent winners.

[3] The activities can be (approximately) normalised 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 normalise them by simple direct division.

7.1 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 the weights connecting the cWM-ep medium with the WM episode.[4] The resulting activity $\vec{y}$ in the WM episode is computed as

$$\vec{y} = \sum_{j=1}^{N} A_j(t) \cdot \vec{w}_j(t) \tag{10}$$

8 Current situation medium

The Current situation medium has a recurrent architecture: as shown in Figure 1, it takes as input the current episode, plus its own internal representation of the history of preceding episodes—'the current context'. It is implemented as an **MSOM** (Strickert and Hammer, 2005)—a type of SOM enhanced with recurrent connections. As in any SOM, 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). Together with the Next episode prediction network (see the next section) the current situation SOM can learn transition probabilities between episodes, and can therefore make predictions about the next episode, similar to those made by a trained simple recurrent network (SRN: Elman, 1990). An advantage of the MSOM over a SRN is that each MSOM unit explicitly remembers the situation it represents in its weights: this means that a situation can be *reconstructed* by top-down propagation from these weights back to the WM episode medium. This can in principle be used for reconstructing a remembered situation from a fragment presented in WM episode medium. Similar to the cWM-ep SOM, the activity $A_i(t)$ of each unit is computed as

$$a_i(t) = \exp(-c \cdot \mathrm{dist}_i(t)) \tag{11}$$

$$A_i(t) = \frac{a_i(t)}{\sum_{j=1}^{N} a_j(t)} \tag{12}$$

The distance $\mathrm{dist}_i(t)$ is a combination of the squared Euclidean distance between its regular weight $\vec{w}_i$ and the input vector $\vec{x}(t)$, and between the context weight $\vec{c}_i$ and the recursive context descriptor $\vec{c}(t)$ (for details see Strickert and Hammer (2005)):

$$\mathrm{dist}_i(t) = (1-\alpha) \cdot \|\vec{x}(t) - \vec{w}_i\|^2 + \alpha \cdot \|\vec{c}(t) - \vec{c}_i\|^2 \tag{13}$$

The context descriptor

$$\vec{c}(t) = (1-\beta) \cdot \vec{w}^{I_{t-1}} + \beta \cdot \vec{c}^{I_{t-1}} \tag{14}$$

[4] 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.

is a combination of the regular weight vector and the context weight vector of the winning unit I_{t-1} from the previous time step (the initial $\vec{c}(0)$ is set to zero vector).

During training, the regular weights $\vec{w}_i$ are updated using the standard SOM learning rule (7) and likewise the context weights

$$\vec{c}_i(t+1) = \vec{c}_i(t) + \gamma \cdot G(I,i) \cdot [\vec{c}(t) - \vec{c}_i] \tag{15}$$

where γ is the learning rate and G is a Gaussian neighbourhood function with linearly decreasing σ from 20 to 1 in the first 5000 episodes and then to 0.1 in episode 20000. In all our experiments, we use $\alpha = 0.6$, $\beta = 0.2$ and the learning rate γ linearly decreasing from 1 to 0.5 in the first 5000 episodes and constant thereafter.

9 Next episode prediction system

The Current situation medium represents a current episode in its context formed by preceding episodes. The Next episode prediction system (Next-ep) uses this medium to capture relations between the whole episodes, namely their transition probabilities. This system is implemented as a single layer of perceptrons with linear activation functions and the softmax function applied to their outputs. The input is the current situation medium (400 units), the output is 400 perceptrons together forming a layer isomorphic to the cWM-ep medium, which represents a distribution of possible next episodes. The weights are adapted using a standard Delta rule (Rosenblatt, 1962) with a constant learning rate 0.9.

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.

After some training, the result of prediction from Step 2 can be propagated/copied to the isomorphic c-WMep medium where it would represent a prior distribution of possible next episodes and can be further propagated top-down to the WM episode medium in a standard way (see Section 7.1) to generate prior expectations (see the experiment in Section 11.4).

10 Training

Our system is exposed to a continuous stream of episodes incoming via primary SM media. This modifies the state of the system in three ways:

1. Instantly—by direct flow of information, *copying* the activities from SM media to the WM individual medium and later to the WM episode medium, thus eliciting some immediate activity patterns;

2. By forward *propagation* via weighted connections in the Candidate WM individuals, Candidate WM episodes, Current situation, and Next episode prediction media, creating a pattern of activities in these media;

3. By long-term learning—modifying the connection weights in the Candidate WM individuals (one-shot update of an existing candidate unit, or addition of a new unit, in case of a novel individual), Candidate WM episodes SOM, Current situation MSOM, and Next episode prediction media (gradual learning).

10.1 Generation of episodes

The training happens all the time in all of the media in parallel. Episodes in the stream are generated stochastically from transcription rules (see Table 2). **Transitive** episodes involve an agent, patient and transitive action, e.g. *A woman stroked a cat*, **intransitive** episodes involve an agent and an intransitive action, e.g. *The dog sleeps*, and **causative** episodes involve an agent, patient and unaccusative action with a special 'cause' signal, e.g. *A man broke (caused to break) a chair*.[5] Each generated episode specifies types of the individuals involved (i.e. WOMAN, CAT, in case of *A woman stroked a cat*). Now it is necessary to determine token individuals of these types. We want to model a situation where a person can encounter novel individuals, but also re-encounter some of the recently seen individuals (so that s/he can learn to represent narratives such as *A man hit a dog. The dog bit the man. He ran.*, i.e. we need to create a basis for indefinite and definite articles and pronouns. In addition to that, there is a limited number ($N = 100$ in most of our experiments) of 'permanent' individuals that can feature in episodes repeatedly (to model familiar individuals), and a potentially open set of newly generated (unknown) individuals. There are two special permanent individuals of type PERSON representing 'me' (the system itself) and 'you' (its dialog interlocutor).

In practice, a list of recently generated individuals is maintained, and when an individual of a certain type is needed, it is drawn from the list (if available) with probability

[5] An unaccusative action is one that involves an object that is not semantically the agent of the action: examples would be 'breaking' or 'dropping'. (Causative actions can also produce actions that in other circumstances are volitional: for instance, a person can 'walk a dog', but we do not include these in our training episodes.)

```
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 2: 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).

$p_{reattend} = 0.99$, otherwise a new individual is generated and added to the list. Non-permanent individuals that have not been reused for past 20 steps are removed from the list.[6]

10.2 Generation of properties of individuals

Each individual has specific properties such as number, gender, colour and location (see below). Some properties non-randomly correlate with types:

- Gender—all animate individuals (people, dogs, cats, birds) are male or female, while all things (cups, chairs, balls) are neutral.
- Colour—is stochastically chosen from Gaussian distributions centred on 11 basic colours for things; on pink, yellow, and black for people; on black, white and brown for animals.
- Location—all individuals are placed on positions on a 100x100 grid so that all things are placed (randomly) in the lower half of the grid, all people in the top left quadrant, all animals in the top right quadrant (with white dogs in the upper half of this quadrant and black dogs in its lower half). These regularities will help us to verify that the candidate WM individual system produces correct expectations about locations (see Section 11.2.3).

In addition to that, each individual has fixed type-specific features (see Table 3). These features are represented more strongly to impose a similarity hierarchy (PEOPLE (DOGS CATS) BIRDS) ((CUPS BALLS) CHAIRS), so that, for instance, dogs are more similar to cats than to birds, and so on.

As we mentioned, the system is not repeatedly exposed to a fixed training set, but to a continuous stream of generated episodes. However, for measuring purposes, we divide this stream into 40 epochs—each epoch involves 500 training episodes.

Experiencing an episode involves presenting its components in a specific sequence in the model's SM media. For transitive episode, the sequence is (AGENT→PATIENT→TRANS-ACTION), for intransitive it is (AGENT→INTRANS-ACTION) and for causative it is (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 then sent to the WM individuals medium, activating the different components of a WM individual representation one by one.

[6]It is necessary to distinguish between the list maintained by the individual generator that simulates the external environment and a list of candidate WM individuals maintained by the cognitive system itself: these are different lists with their own mechanisms of update and removal.

```
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 3: 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.

11 Experiments

In this section we describe experiments focusing on different aspects/tasks in the model. Unless specified otherwise, the regularities described above are present in training sets in all our experiments. For some experiments, we introduced further regularities—these are described with the experiments.

11.1 Testing the sequence-based binding scheme

To demonstrate the new scheme for binding semantic roles in episode representations to individuals, 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. Once the first-order SM signals are recreated, they are compared to the representation of the individual that was recorded in the original episode. Among all occurrences of agents and patients in 500 episodes in each training epoch we evaluate the proportion of cases when these representations match. Across all training epochs, the representations of individuals were correctly matched in 99.7–100% of cases; this demonstrates that our proposed binding mechanism is effective.

11.1.1 Replaying unseen episodes

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 episodes that were not encountered during training (i.e, the episodes were previously unseen combinations of object types and the types were instantiated with novel token individuals). We tested on 100 unseen episodes in each training epoch; 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.

11.2 Testing the network's prediction/generalisation abilities

The network can make several kinds of prediction; we will focus on three progressively more complex predictions.

11.2.1 Expectations about actions

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 exploited the following regularities in 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 (with the rest of the WM episode medium unfilled/inactive), 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 Section 7.1). Figure 2 shows that these distributions are correctly weighted towards the actions encountered during training.[7]

11.2.2 Prior expectations about agents

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 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

[7]These graphs show particular instances of the three episode fragments; however, a similar pattern was also obtained when averaging across multiple instances with different token individuals.

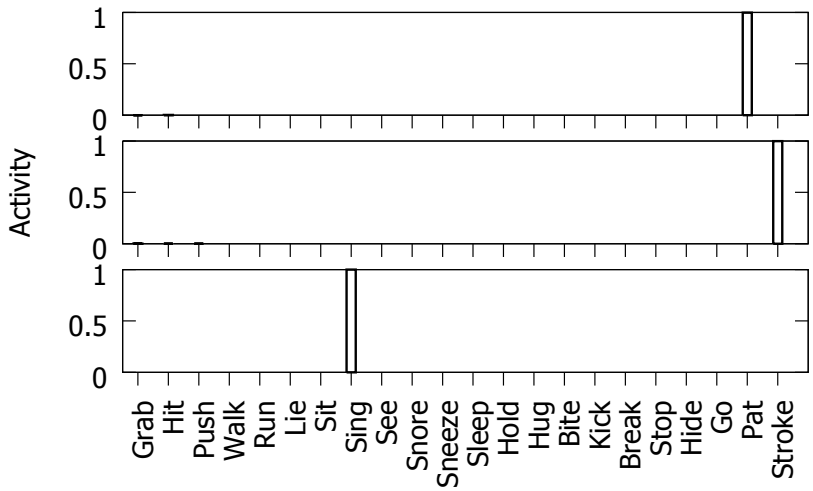

Figure 2: Action types predicted in the cWM-ep layer for 3 episode fragments. From top to bottom: PERSON→DOG→?, PERSON→CAT→?, BIRD→?.

its scalar weight, see Section 7).[8] We reconstructed a distribution over types/properties in the agent part of the WM episode, again based on the pattern of activities of all SOM units (see Section 7.1). 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 Section 5). 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 3 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.

11.2.3 Predictions about properties and locations

The WM episodes and WM individuals systems also interact in generating useful predictions about the locations and properties of individuals encountered during episode percep-

[8] Another possibility based on predictions from the Next episode prediction system is explored in Section 11.4.

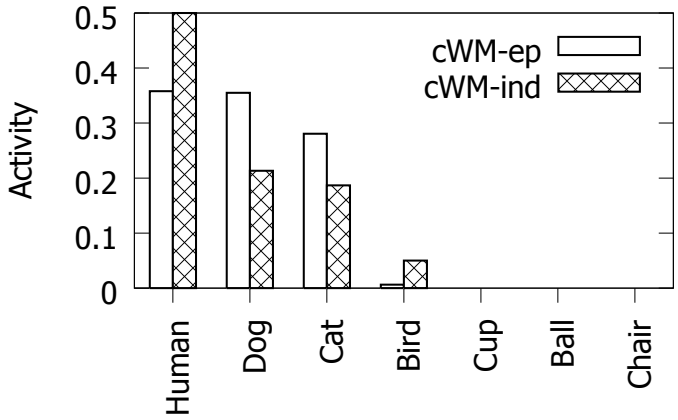

Figure 3: Prior expectations on the agent type generated top-down by the cWM-ep and cWM-ind layers.

tion. To test these, we exploited some further regularities 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, all white dogs were located in the upper half of the top-right quadrant of the grid and black dogs in the lower half of this 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 used the SOM activity to reconstruct predicted distributions of patient features (see Section 7.1). 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 4 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 5. There is a general bias towards the top-right quadrant, since dogs always appear there. 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.

11.3 Generalization to unseen object types

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

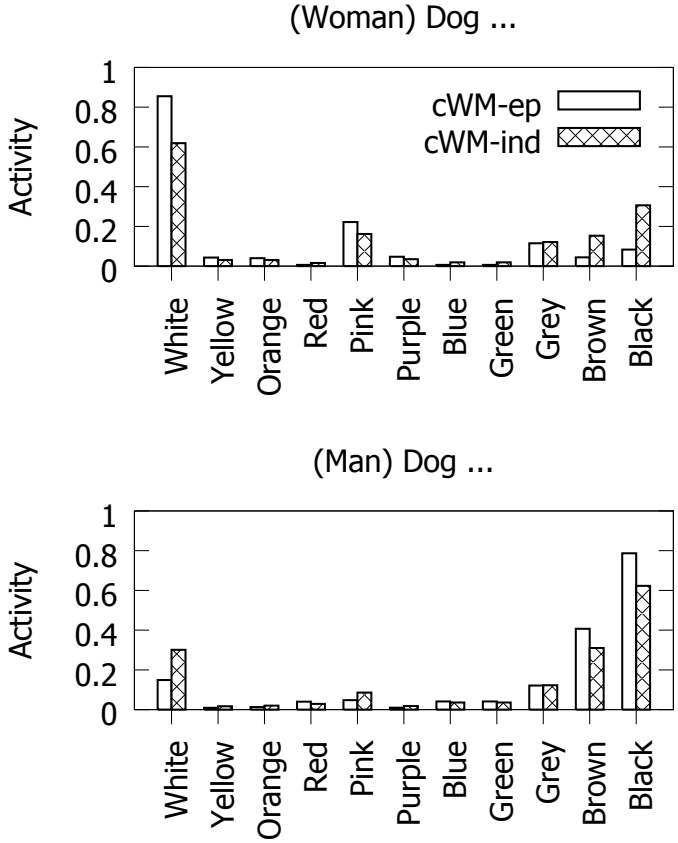

Figure 4: Expectations about the colour of the patient generated by the cWM-ep and cWM-ind layers for episode fragments WOMAN→DOG (top) and MAN→DOG (bottom).

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 2 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 6, the action for the unseen cat target is biased towards 'pat' while the action for the unseen ball target is biased towards 'grab'.

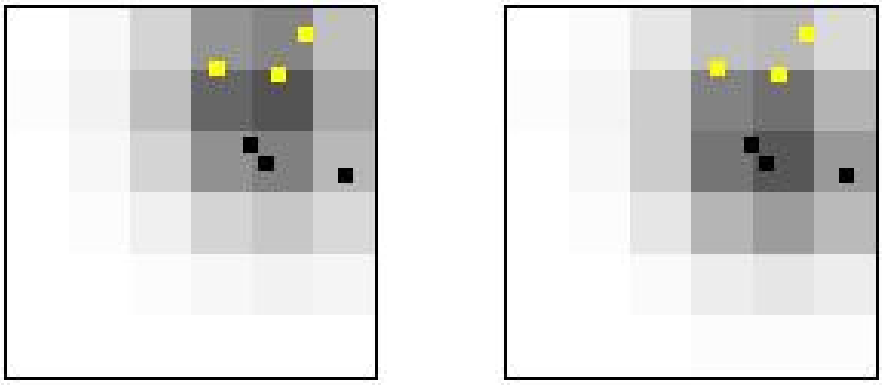

Figure 5: Expectations about location of the patient generated by the cWM-ind layer for for episode fragments WOMAN→DOG (left) and MAN→DOG (right). Darker areas mean stronger expectations. Black (yellow) dots represent actual locations of currently present black (white) dogs.

11.4 Predicting sequences of episodes

Recall that, thanks to recurrent connections, the Current situation MSOM represents episodes together with their context (that in a compressed form encodes preceding episodes). Activity in this medium serves as input to the Next episode prediction system that generates a distribution of possible next episodes in its output layer isomorphic to cWM-ep system (see Sections 8 and 9). 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 (see Section 7.1). Figure 7 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 8), 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

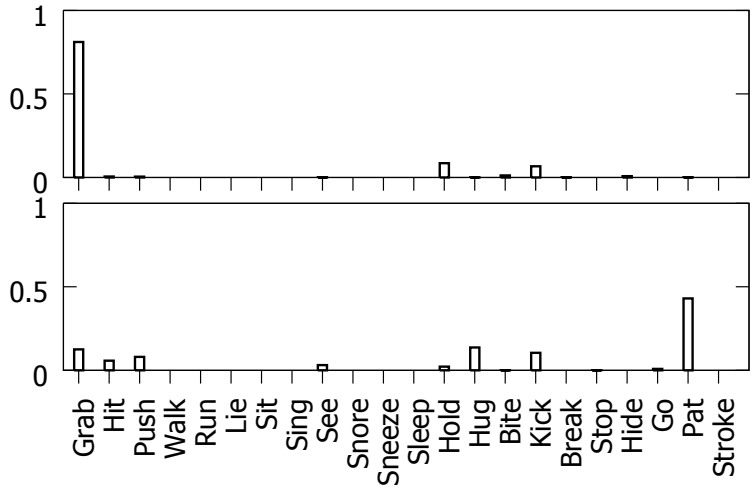

Figure 6: Action types predicted in the cWM-ep layer for episode fragments PERSON→BALL→? (top) and PERSON→CAT→? (bottom).

involves other actions too.

We also inspected the activity patterns in the cWM-ep medium and Current situation after presenting the episode (PERSON→DOG→PAT) in the WM episode medium. While the activity pattern in the cWM-ep was identical, the Current situation medium responded with different (though neighbouring) winners: a unit with map coordinates $[16, 2]$ in the condition A and $[1, 15]$ in the condition B, because the episode appeared in different contexts.

We were also interested whether the system could capture the knowledge that the dog and the man in the predicted episode are the same individuals as in the preceding 'a man patted a dog' episode, which, on the level of episodes, can be expressed via predicting their properties like number, colour, gender correctly. In our episode, 'Yellow' men (Pl) patted a black female dog (Sg). However, in the predicted 'dog bites person' episode the predicted black dog (Sg) was male with odds 80:20, and the predicted male person was in singular (odds 70:30) and the unit for pink colour outperformed the yellow one (25:20). This is caused by low 'granularity' of the cWMep system. In order to capture all the properties correctly, it would need to store all their combinations in weights of separate units. Given the current cWM-ep capacity (400 units), the system often generalises across individuals while correctly distinguishing types (see also discussion in Section 12).

Finally, we inspected what happens after the expectation in the WM episode is relayed through the cWM-ind system. With expectations about the dog agent, the system activated units for 3 individual black dogs (one male and two female, one of the female being the correct one) and in the resulting expectation the gender skewed back toward female (odds 62:38). Regarding the male patient, the expectation stayed with Sg man with dominantly pink colour. The reason is the cWMind system tries to match the expectation copied from cWM-ep system as well as possible (and biases it by the properties of matching

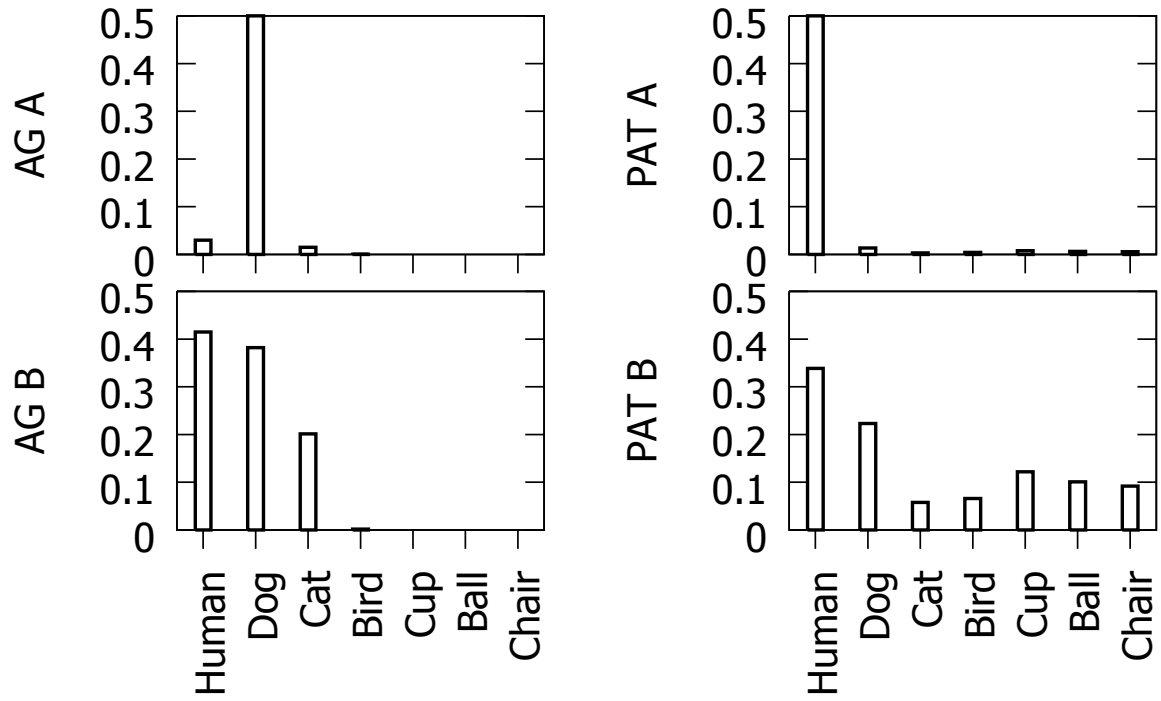

Figure 7: Prior expectation about the agent and patient of an episode following the episode (PERSON→DOG→PAT) in the condition A (the man hitting the dog previously, top row) and B (a different previous episode, bottom row).

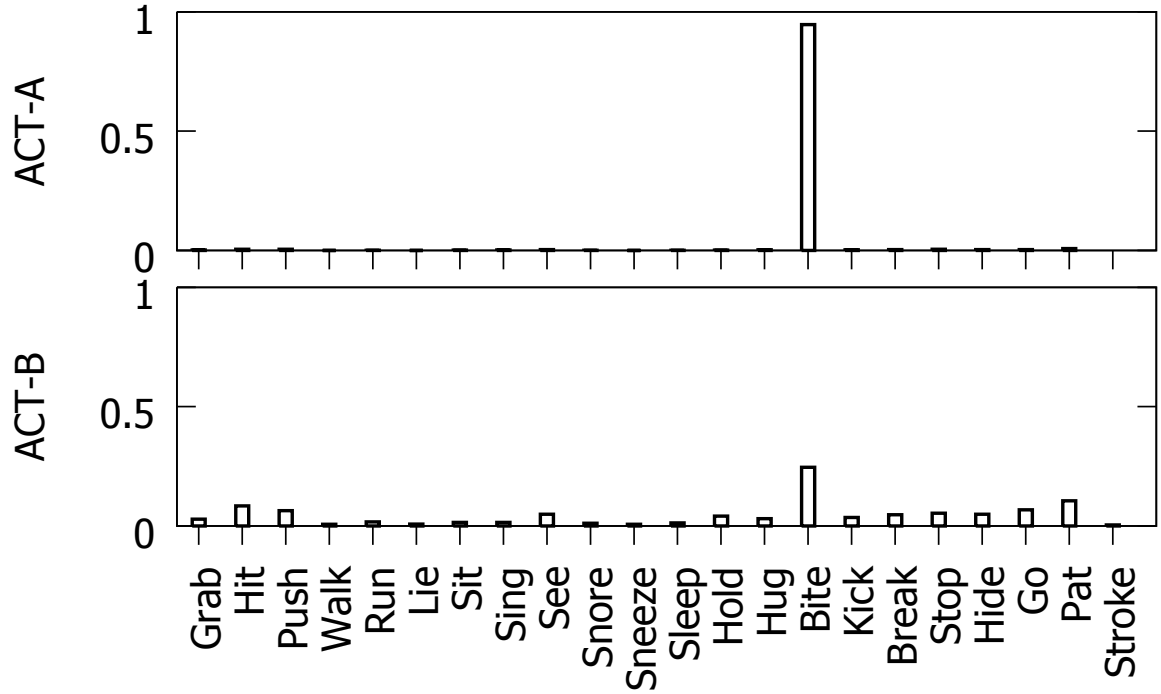

Figure 8: Prior expectation about the action of an episode following the episode (PERSON→DOG→PAT) in the condition A (the man hitting the dog previously, top row) and B (a different previous episode, bottom row).

individuals it currently remembers).

12 Discussion

In this report, we have provided technical details of a novel model of representation of episodes and individuals in working memory. However, the model goes beyond working memory: we envisage that the Current situation and Next episode media also participate in long-term memory (LTM). In this sense, information stored in the connection weights is long-term, while a current pattern of activities constitute a working memory representation. Also, the model is far from being complete; in the future we plan to enhance it with other LTM media for LTM individuals and spatial and temporal contexts.

One remaining issue to discuss is the representational capacity of the model, namely its SOMs (cWM-ep and Current situation). The theoretical upper bound for episode types[9] is 368, so the cWM-ep SOM can represent each episode type by a different unit. By exposing the trained cWM-ep to 500 stochastically generated episodes and recording the winning units for each episode, we found out that it indeed never represented two different episode types by the same units. However, in many cases units generalised across individual properties such as colour or gender. To test the hypothesis that, given enough units, the SOM would distinguish individuals, we ran another experiment where all episodes involved just dogs, cats and 2 transitive, 2 intransitive and 2 causative actions (i.e. 20 episode types), moreover the individuals were fixed—the world consisted of 7 dogs and 3 cats that featured in all the episodes. We inspected the SOM winning units in the same way as above, and now the system distinguished between individuals: all episodes involving the same individuals and action were represented by the same units, while an episode of the same type involving different individuals was represented by a different unit. A (frequent) exception to this was generalization across individuals in one of the roles, e.g. episodes involving a particular individual dog as agent biting one of two individual cats in the patient role. Clearly, the cWM-ep medium cannot store all possible combinations of all possible individuals in different roles, but this is not its purpose anyway. The medium is flexible enough to adapt to its capacity limits by generalising over smaller and less frequent differences. The same holds for the Current situation medium, where the combinatorial possibilities of episodes in different contexts would be even more demanding.

References

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

Fink, G., Halligan, P., Marshall, J., Frith, C., Frackowiack, R., and Dolan, R. (1996).

[9] 4 types of animate agents, 7 types of patients, 8 intransitive, 8 transitive, and 4 causative actions make 4x8+4x7x8+4x7x4=368 episode types. In reality, there are less episode types, because not all their combinations are allowed (see Table 2).

Where in the brain does visual attention select the forest and the trees? *Nature*, **382**, 626–628.

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

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

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

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

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

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

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

Robertson, L., Lamb, M., and 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.

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

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

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. (2016). A simulationist model of episode representations in working memory: role-binding, hierarchical structure and inference. Submitted to Cognition.

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

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