Swann's Name: Towards a Dialogical Brain Semantics

Jonathan Ginzburg
CNRS, Université Paris-Cité
Laboratoire de Linguistique Formelle
yonatan.ginzburg@u-paris.fr

Chris Eliasmith
Centre for Theoretical Neuroscience
University of Waterloo
celiasmith@uwaterloo.ca

Andy Lücking
Goethe University Frankfurt
Text Technology Lab
luecking@em.uni-frankfurt.de

Abstract

The paper argues with reference to several examples that dialogical dynamic semantics, the idea that meaning arises from emergent public context, breaks down over extended temporal periods, ignoring as it does individual differences specifically with respect to memory dynamics. We argue, following several recent works, that this highlights the need for a semantics that is brain-based. We offer a sketch for such a semantics by developing a hybrid model that integrates work on memory–oriented dialogue semantics with work in the semantic pointer architecture for functional brain modelling.

1 Introduction

Dialogical dynamic semantics, the idea that meaning arises from emergent public context, can be effective for dialogue over short temporal periods. But over more extended temporal periods, dynamic semantics begins to break down, ignoring as it does individual differences specifically with respect to memory dynamics. Consider the following mundane story: I encounter my neighbour's daughter Swann when she gets locked out and learn her name. Two years pass: I encounter Swann occasionally, as I hear her close the entrance door, but I do not hear her name spoken. One morning I see Chloé, Swann's sister, and wonder: what is Chloe's sister's name? I remember it starts with 'S'. But I cannot remember the name. This lasts for a while. I see a list of names and know that they are not the name. Finally I see the name and recognize it. This *inner dialogue* can also be envisioned as a series of external dialogues:

(1) a. Dialogue 1: Neighbour: *This is Swann.* Me: *Nice to meet you.*

b. Interlude (time passes, events happen)

c. Dialogue 2: (I see Chloé) Me: *How is um* (pause, frowns) *your sister*? Chloé: *Swann*? Me: *Yes.*

(2) provides an additional illustration of the effect of time—dissociation between event-based, individual-based, and metalinguistic information, as exemplified in (2b), a dissociation backed by considerable clinical evidence (Greenberg and Verfaellie, 2010; Bastin et al., 2019).

(2) a. A: *Look, someone's broken the door handle.* B: *Right.* C: *Yeah it's this woman, Sloane.*

b. (a week later) D: *What had happened*? A: *What's her name, I forget, broke the door handle.* D: *and Bill was there too apparently.* B: *Who*? A: *Her partner.* B: *I don't know him.* A: *We met him last week. B: Oh, I see.*

We think cases such as these highlight the need, already outlined in several works (Eliasmith, 2013; Baggio, 2018; Hagoort, 2020; Macnamara and Reyes, 1994; Jackendoff, 2002; Seuren, 2009) for a semantics that is brain-based (where again one can appeal to the (biophysical/biochemical) neuronal and the neuron-network levels): generalizations about behaviour can occur at various levels (Marr, 1982; Bechtel, 2007; Eliasmith and Kolbeck, 2015); appealing also to brain-based levels need not mean that all explanations are most usefully stated at those levels—for instance, as we will see certain rules concerning dialogue coherence.

And yet, we think, nonetheless, that this data enables one to make stronger claims, namely that a brain-based account impacts also on the structure of the cognitive theory one can and should provide. In particular, it requires us to capture (i) the intrinsically associative character of memory (exemplified here by the speaker's thinking of Chloé's

Proceedings of the 28th Workshop on the Semantics and Pragmatics of Dialogue, September, 11–12, 2024, Trento, Italy.

sister when seeing Chloé, mirrored by corresponding external dialogue coherence) (ii) dissociative aspects in cognitive states (exemplified by forgetting Swann's name but not Swann and data in (2)), (iii) the pervasive nature of forgetting and the non-redundancy of reproviding (forgotten) information, and (iv) differences in communal memory emergent from individual divergences.

The paper is structured as follows: in section 2 we introduce relevant background about the various neural levels. We develop our account in two stages: in section 3 we apply an externalist, though memory-oriented dialogue framework NeuroKoS (Ginzburg and Lücking, 2022) to the data, which can only offer a partial account; in section 4 we discuss a simple model of the data using the Semantic Pointer brain-modelling framework (Eliasmith, 2013), which offers an account of the aspects which NeuroKos cannot handle.

2 Learning and Forgetting at the Neural Level

2.1 Short-term v. Working Memory v. Long-term Memory

The neuropsychological basis for short-term and long-term memory (STM, LTM) distinctions are both experimental (e.g., ability to recall number sequences or labelled pictures after a single presentation) and based on studies of patients, most notably the patient Henry Molaison (aka H.M.), well known for being high functioning despite lacking the ability to form new (episodic) memories that could persist beyond 45 minutes (Scoville and Milner, 1957; Milner and Klein, 2016; MacKay et al., 2013; Squire and Wixted, 2011). Working memory (WM) is a distinct though closely related notion to short-term memory amounting to 'an actively engaged system used to store information that is relevant to the current behavioral situation.' (Eliasmith, 2013, p. 211). Baddeley (1988, 2012) offered both arguments for the notion of WM and developed an influential framework, M-WM, which postulates a clear structure for WM (on which more below); an alternative to this was proposed by Cowan (2001), who emphasizes the capacity constraints of WM. Both Baddeley's episodic buffer and Cowan's focus of attention are chunk limited buffer stores, and both models by and large agree on a capacity limit of four chunks. An important issue such theories have contended with is whether working memory is a separate system (Baddeley) or merely a temporal slice from a unified memory system (Cowan, on one reading, though ultimately the differences between the frameworks are not large). Norris (2017) argues that STM/LTM are distinct systems given the need for (i) memory for previously unencountered information, (ii) storage of multiple tokens of the same type, and (iii) variable binding (in one sense of the term). Be that as it may, the exact relationship between WM (which is evinced in actual use) and STM/LTM is not fully clear. What is clear is that there are WM/LTM distinctions at neural and neural network levels.

2.2 Short-term and Long-term Learning at the Neural Level

Given the relative ease of access to their neural systems, the solidly established results on learning at the neural level have arisen from various invertebrates and from rodents. As exposited by Kandel et al. (2014) one can distinguish two classes of mechanisms: short/medium term changes in synaptic strength arising from specific patterns of electrical activity or the action of modulatory transmitters; long-lasting synaptic and behavioral memory plasticity requires epigenetic mechanisms—changing gene expression without modifying the underlying DNA: on the one hand the inhibition of miRNA-124 which facilitates the activation of CREB-1, which begins the process of memory consolidation, and on the other hand the delayed activation of piRNA, which leads to the methylation and consequent repression of the promoter of CREB-2. This allows CREB-1 to be active for a longer period of time.

2.3 Short-term and Long-term Learning at the Neural Network Level

As far as LTM goes, it is commonly assumed that memories are not stored in the hippocampus as such, but arise from the interaction of representations based at the hippocampus with neocortical information: sparsely-coded hippocampal neurons referencing and activating the neocortical neurons to re-create the content of an experience (Teyler and Rudy, 2007). Semantic memories, arising by generalisation across the neocortical representations of episodic memories are resistant to hippocampal damage. For a long period, the fact that performance on many explicit tasks is affected by temporally graded retrograde amnesia was explained by assuming that the hippocampus is only a temporary repository for memory whereas the neocor-

tex stores the memory (Squire and Wixted, 2011). More recently, evidence emerged that mediotemporal lobe lesions do not lead to a pattern of retrograde amnesia and also affect non-episodic, semantic memory. Sekeres et al. (2018) propose Transformation Trace Theory (TTT): *transformed* memories (i.e., ones shorn of detail) come to be represented in distributed neocortical networks from where they can be recovered without the involvement of the hippocampus; *detailed* episodic memories are always dependent on the hippocampus. The evidence for this is evidence that once a consolidated memory is reactivated, it can become labile and once again become susceptible to the effects of hippocampal disruption.

This leads to at least the following sources for forgetting, which models of forgetting need to tie into:

1. Non-consolidated short-term memories;
2. Detail modification during activation (Sekeres et al., 2018);
3. Loss as a result of neurogenesis (Weisz and Argibay, 2012; Epp et al., 2016);
4. Weight decay and synapse elimination (Richards and Frankland, 2017).

3 Towards an Account

3.1 Combining Memory and Dialogue GameBoards

As mentioned earlier, we draw on an earlier proposal, the only existing one to our knowledge, for combining externalist dialogue semantics with memory structure (Ginzburg and Lücking, 2020, 2022). But first, a brief explanation of externalist dialogue semantics, as conceived in the framework KoS (Ginzburg, 1994; Larsson, 2002; Purver, 2004; Fernández, 2006; Ginzburg, 2012)—formulated using the logical framework TTR (Cooper and Ginzburg, 2015; Cooper, 2023). Instead of assuming a single context to be operative, a collective notion is emergent (Stephens et al., 2010) from individual *Total Cognitive States* (TCS), one per participant. A TCS has two partitions, namely a *private*, and a *public* one, the DGB.

(3) $\text{TCS} =_{def} \begin{bmatrix} \text{public} : \mathit{DGBType} \\ \text{private} : \mathit{Private} \end{bmatrix}$

Dialogue gameboards (see (4) for the basic structure) track various aspects of the emerging context in terms of concrete real world entities and more abstract ones constructed in TTR. The parameters *spkr* and *addr* together with the addressing condition (at a given time) track verbal turns and mutual engagement; *vis-sit* represents the visual situation of an agent, including his or her focus of attention (*foa*), which can be an object (*Ind*), or a situation or event (*Rec*), relevant *inter alia* for processing gestural answers; *facts* represents the shared assumptions of the interlocutors; uncertainty about mutual understanding that remain to be resolved across participants—*questions under discussion*—are a key notion in explaining coherence and various anaphoric processes (Ginzburg, 2012; Roberts, 1996) and is tracked by the parameter *qud*; dialogue moves that are in the process of being grounded or under clarification are the elements of the *pending* list; already grounded moves are moved to the *moves* list, which captures expectations arising due to illocutionary acts—one act (querying, assertion, greeting) giving rise to anticipation of an appropriate response (answer, acceptance, counter–greeting), also known as adjacency pairs (Schegloff, 2007); finally, *mood* represents the publicly accessible emotional aspect of an agent that arises by publicly visible actions (such as non-verbal social signals, as well as by verbal exclamations), which can but need not diverge from the private emotional state:

(4) $\mathit{DGBType} =_{def}$

$$\begin{bmatrix} \text{spkr} & : & \mathit{Ind} \\ \text{addr} & : & \mathit{Ind} \\ \text{utt-time} & : & \mathit{Time} \\ \text{c-utt} & : & \text{addressing(spkr,addr,utt-time)} \\ \text{facts} & : & \mathit{Set}(\mathit{Prop}) \\ \text{vis-sit} & = & \left[\text{foa} : \mathit{Ind} \vee \mathit{Rec}\right] : \mathit{RecType} \\ \text{pending} & : & \mathit{List}(\mathit{LocProp}) \\ \text{moves} & : & \mathit{List}(\mathit{IllocProp}) \\ \text{qud} & : & \mathit{POSet}(\mathit{Question}) \\ \text{mood} & : & \mathit{Appraisal} \end{bmatrix}$$

TCSs and in particular DGBs change as a result of private perception and public interaction, which can be described in terms of *conversational rules* (Larsson, 2002). We exemplify here three rules (minor variants of rules in Ginzburg, 2012, Chapters 4,6) that will play a role subsequently. The first exemplifies coherence at the level of Moves, the second the emergence of presuppositions, the third the coherence of clarification questions:

(5) a. **Interlocutor introduction rule**: given that the LatestMove is Introduce(A,B,C), this licenses the next move to be Greet(B,C).

b. **FACTS update following assertion acceptance**: if the LatestMove is Accept(A,p), this licenses FACTS:=FACTS $\cup\{p\}$

c. **Confirmation question emergence**: if A's utterance u is (a sub-utterance of) the maximal element of Pending, QUD can be updated with the question *did A mean c by u?* (c some potential referent/content).

KoS provides a theory of meaning for highly context dependent elements such as non-sentential utterances (6a,b), filled pauses (6c), and non-verbal social signals such as smiles or frowns (6d,e), which figure further below.

(6) a. *yes* $\mapsto$ p (p? is MaxQUD);

b. *right* $\mapsto$ Understand(A,u) (u is MaxPending, A current speaker);

(both Ginzburg, 2012)

c. *um* $\mapsto$ Makes λxMeanNextUtt(spkr,Pending,x) MaxQUD (Ginzburg et al., 2014)

d. smile: Given A as speaker, s as smilable event, $\mapsto$ Pleasant(s,A)

e. frown: Given A as speaker, f as frownable event, q : Question $\mapsto$ Raise(f,q,A)

(both Ginzburg et al., 2020)

The essence of the proposal of Ginzburg and Lücking (2020, 2022) is to tie the externally–oriented data structure used to describe dialogue dynamics, the dialogue gameboard (Ginzburg, 2012), with working and long-term memory. Thus, they propose to 'break up' the dialogue gameboard into WM and LTM components, building on models for WM (Baddeley, 2012) and LTM (Bastin et al., 2019), respectively—see Fig. 1 for a graphical summary. In particular, they proposed to (i) view conversations as episodes tracked in episodic memory, (ii) distinguish within LTM the following components: (a) episodic memory typically associated with the hippocampus, (b) entity-based memory (based in the perirhinal cortex, Bastin et al., 2019), and (c) semantic memory (mainly localized in the posterior region of the left temporal lobe, Saumier and Chertkow, 2002, though the specific regions involved in semantic memory retrieval depend on whether sensorimotor or abstract amodal features are accessed, Reilly et al., 2016).[1]

Characterizing the emergence of LTM is of course highly complex—Ginzburg and Lücking (2022) offered one simplified rule concerning episodic memory, but said nothing about entity and semantic memory. We refine very slightly their rule concerning episodic memory and offer two very simplified rules concerning entity and semantic memory. Events undergo appraisal which leads to both updates in the current emotional makeup of the cognitive state (both in the private and in the public parts) and to creating episodic indices in the hippocampus, which are in effect vertices in a network connecting to percepts of events stored neocortically. We assume that such indices are created for events with positive pleasantness above a threshold or negative pleasantness above a larger threshold—which yields a bias for long-term memory of enjoyable events or of highly unpleasant ones. The rule in (7) creates a fresh index and associates it with the current event in working memory, originating either in Pending or in vis-sit:

(7)
$$\begin{bmatrix} \text{pre} : \begin{bmatrix} \text{e=MaxPending} \vee \text{vis-sit} : \textit{RecType} \\ \text{c1} : \text{Private.Mood.pleasant.affect.pve} \geq \theta_1 \\ \quad \vee \text{Private.Mood.pleasant.affect.nve} \geq \theta_2 \end{bmatrix} \\ \text{effects} : \begin{bmatrix} \text{n} = \text{card(HC-Indices)+1} : \mathbb{N} \\ \text{HC-indices} := \text{HC-Indices} \cup \langle \text{n,pre.e} \rangle \end{bmatrix} \end{bmatrix}$$

Although Tulving (1972) suggested that semantic memory was in some sense prior to episodic, recently it has been common to view both entity and semantic memory as emerging from decontextualized episodic traces (and existing in parallel) (Greenberg and Verfaellie, 2010).

We define an *individual-oriented* subpart of a record type as in (8a) and exemplify it as in (8b):

(8) a. Assume l_1 is a label of the record type i and $i \sqsubseteq \left[\text{l}_1 : \text{Ind}\right]$ and for no other label l_i in

[1] From a formal point of view one might say that an entity–oriented semantics has already been proposed in Irene Heim's File-Change Semantics (Heim, 1982), though in that case the episodes are represented within each individual file, which emerges with the utterance of an indefinite. So there is no dissociation and of course no means to deal with forgetting or associative memory. The same is true for related *mental files* approaches (e.g. Maier, 2016).

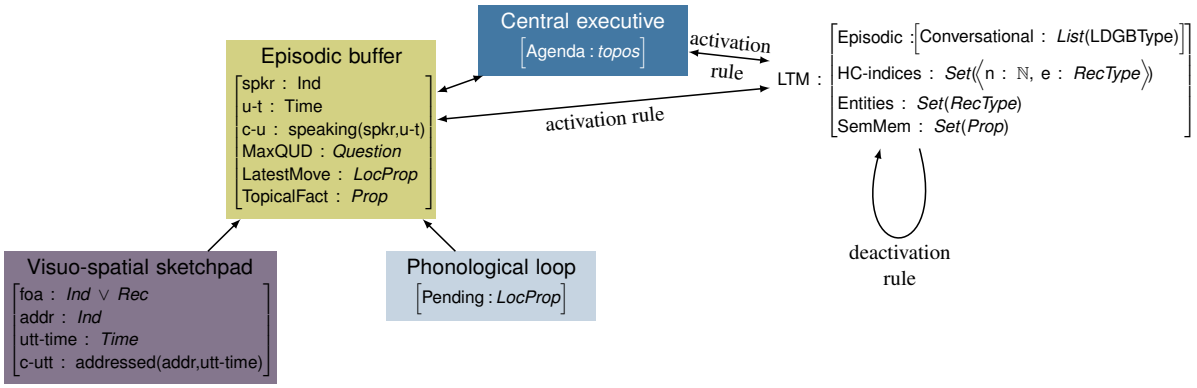


Figure 1: Fusing M-WM and DGB, and adding LTM.

i is it the case that $i \sqsubseteq \left[\mathrm{l}_i : \mathrm{Ind}\right]$ and assume r is a record type such that for some j $r = i \overset{\cdot}{\wedge} j$ ('merge'), then i is an *individual-oriented* subpart of r.

b. $i = \begin{bmatrix} \mathrm{x} & : \mathit{Ind} \\ \mathrm{C} & : \mathit{faceshape} \\ \mathrm{c1} & : \mathrm{C(x)} \\ \mathrm{c}_{\mathit{name}} & : \mathrm{Name(Emmo,x)} \end{bmatrix}$,

$r = \begin{bmatrix} \mathrm{x} & : \mathit{Ind} \\ \mathrm{C} & : \mathit{faceshape} \\ \mathrm{c1} & : \mathrm{C(x)} \\ \mathrm{c}_{\mathit{name}} & : \mathrm{Name(Emmo,x)} \\ \mathrm{y} & : \mathit{Ind} \\ \mathrm{c2} & : \mathrm{Hammer(y)} \\ \mathrm{t} & : \mathit{Time} \\ \mathrm{c3} & : \mathrm{Hold(x,y,t)} \end{bmatrix}$

We will assume that entities emerge in LTM as individual-oriented parts of episodes from episodic memory:

(9) **Entity memory update**: If $\langle n, r \rangle \in$ HC-Indices and i is an individual-oriented part of r, then Entities $:=$ Entities $\cup \{i\}$

The principle we sketch for the emergence of semantic memory involves a subcase of the FACTS update rule (5b) above. We assume that assertions communicating stative information update semantic memory. This is of course quite crude, but presumably a more refined typing of propositions can offer a reasonable starting point for such a procedure.

(10) **Semantic memory update**: If $p \in$ FACTS and p : *StativeProposition*, then SemMem $:=$ SemMem $\cup \{p\}$

We mention one additional principle, which we will not attempt to formalize in the current setup, but which is (partially) formalizable in the neural setup of section 4. It is intuitively correct for *inner dialogue*, and we think reasonably extensible to interactive dialogue:

(11) **Associative topics:** If q is a question whose similarity to MaxQUD $\geq \theta$, $Ask(q)$ is licensed as the LatestMove

3.2 Initial Account

We return to our initial example repeated here as (12):

(12) Dialogue 1: Neighbour: *This is Swann.* Me: *Nice to meet you.*

Given the tools we have, we can explain the following: the coherence of my response to the neighbour's introduction (on the basis of the **Interlocutor introduction rule**, (5a)); the update of entities with the individual Swann (as an update of entity memory, see (9)), the update of semantic memory with Swann's name (as an update of semantic memory, see (10)).

For the second dialogue repeated here as (13):

(13) Dialogue 2: (I see Chloé) Me: *How is um* (pause, frowns) *your sister*? Chloé: *Swann*? Me: *Yes.*

we can explain how the self-repair question introduced by a filled pause licences a frown (see (6)); we can explain the coherence of Chloé's confirmation request (see (5.c)). On the other hand, **we do not have a means of explaining my inability to recall Swann's name** (since it is in my semantic memory), **nor the restorative effect of Chloé's utterance on the availability of Swann's name. Nor do we have a means of explaining why I think of Swann when I see Chloé**; my asking

about Swann could be explained if we had a means of formalizing our rule of associative topics, as a question similar to asking how Chloé is. We suggest that dealing with these unresolved issues requires a brain-oriented semantics, to which we now turn.

4 Adding a Neural Level

4.1 The Semantic Pointer Architecture

We draw on the Semantic Pointer Architecture (SPA) approach to cognition (Eliasmith, 2013). The idea in a nutshell is the following: an input current is nonlinearly encoded within a population of neurons according to each neuron's tuning curve and spiking pattern. The encoded input can either be reconstructed by other populations of neurons by weighted linear decoding (the pair of encoding and decoding defines a neural *representation*), or transformed (by another weighted linear decoding). We employ vectors as a means for representing symbols, dubbing them *semantic pointers* (SPs), since we construe them as compressed representations that carry partial semantic content. Certain transformations can be defined on the class of SPs. One of the most important transformations is *circular convolution* (Plate, 1991), which *binds* two or more vectors into an output vector $\mathbf{v}$ without increasing dimensionality but ensuring also that the input vectors can be *unbound* or *decoded* from $\mathbf{v}$, albeit with some noise.[2,3]

If vectors $\mathbf{d}$ and $\mathbf{e}$ are bound into $\mathbf{p}$, $\mathbf{p} = \mathbf{d} \circledast \mathbf{e}$, "$\circledast$" being circular convolution, then $\mathbf{d}$ can be approximately recovered from $\mathbf{p}$ by binding $\mathbf{p}$ with the inverse of $\mathbf{e}$: $\mathbf{d} \approx \mathbf{p} \circledast \mathbf{e}'$ ($\mathbf{e}'$ being the inverse of $\mathbf{e}$). Encoding, decoding, and transforming are dynamic processes in time and are implemented using the software tool Nengo (Bekolay et al., 2014), which also allows for "biological compilation" in terms of neural simulations.[4]

The SPA has successfully been applied to a number of cognitive tasks, including the representation of concepts (Blouw et al., 2016), memory (Gosmann and Eliasmith, 2021),and emotion (Thagard et al., 2023), and underlies the world's largest functional brain model to date (Eliasmith et al., 2012).

4.2 SPA and Symbolic Representation

A key feature of the SPA is that it enables a systematic correspondence of symbolic and neural content in a way that meets Jackendoff's challenges for cognitive neuroscience (Jackendoff, 2002; Gayler, 2004). In recent work Larsson et al. (2023) show how to map TTR entities into SPA ones, offering a mapping that covers basic types, perceptual and cache-based judgements, singleton types, record types, meet types and merging of record types, ptypes, and subtyping.

4.3 Completing the Account

We employ the SPA to propose a simple model that completes our account of the simple name forgetting episode (1), and (12) and (13), respectively.[5]

Adding a neural level allows us to offer *rudimentary* accounts of desiderata (i) to (iv) from section 1, in particular a gradual emergence of forgetting. The current model is simplified as a brain model in a variety of aspects: no WM (so no short-term learning); consolidation is assumed to happen; there is no coupling between dialogue cognitive states; perfect perception/communication is assumed—no processing of vision or language is integrated into the account.

The model represents certain perceptual input (visual and linguistic) and resultant memory traces as semantic pointers. It models recollection of an entity's property P (e.g., x's name) as (i) finding the vector most similar to the current percept and (ii) unbinding the entity bound to P. If recollection is successful, (a) the entity found is updated with the information originating with the current percept and (b) a smile is triggered,[6] otherwise a frown is

[2] *Vector Symbolic Architectures* (VSA; Gayler, 2004) define symbolic operations on high-dimensional numerical vectors. See Schlegel et al. (2022) for a very useful survey of Vector Symbolic Architectures.

[3] *Circular convolution* $C = A \circledast B$ is defined as in (i) in a space of dimension D, whereas the inverse of a vector is defined as in (ii), and we use the notation B' for B^{-1}

(i) **Circular convolution**
$c_j = \sum_{k=0}^{D-1} b_k a_{j-k(modD)}$
for $j \in \{0, \ldots, D-1\}$

(ii) **Inverse for circular convolution**
$a_j^{-1} = a_{D-j(modD)}$
where $j \in \{0, \ldots, D-1\}$
In other words: $\langle a_0, a_1, \ldots, a_{D-1}\rangle^{-}1 = \langle a_0, a_{D-1}, \ldots, a_1\rangle$

[4] https://www.nengo.ai/

[5] The code for the model is available here: https://github.com/aluecking/Swanns-Name. Note that you might obtain numbers that differ from those given in this paper due to the random initialization of vectors.

[6] In a more detailed model, the motor neurons responsible for the action sequence responsible for a smile would be

triggered.

Initially there is input about three girls, Swann ('girl_52'), Anais ('girl_53'), and Chloé ('girl_54'). Swann and Anais are differentiated in terms of their names and food preferences and have the same hair colour and are neighbours (of the observer), whereas Chloé has Swann's properties bundled with being her sister:[7]

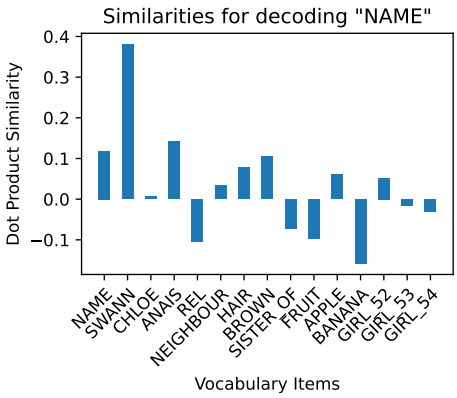
(14) a. GIRL_{52} = NAME ⊛ SWANN + REL ⊛ NEIGHBOUR + HAIR ⊛ BROWN + FRUIT ⊛ APPLE

b. GIRL_{53} = NAME ⊛ ANAIS + REL ⊛ NEIGHBOUR + HAIR ⊛ BROWN + FRUIT ⊛ BANANA

c. GIRL_{54} = GIRL_{52} + SISTER_OF + NAME ⊛ CHLOE

At this point, the state views Chloé and Swann as similar (their dot product is 0.59), and recalls Swann's name (the vector associated with the name SWANN is most similar to the decoded vector with a dot product of 0.38), as indicated in (15) for decoding NAME and in Fig. 2, where the most similar items when unbinding all its properties are shown:

(15) The name "Swann" is recalled:

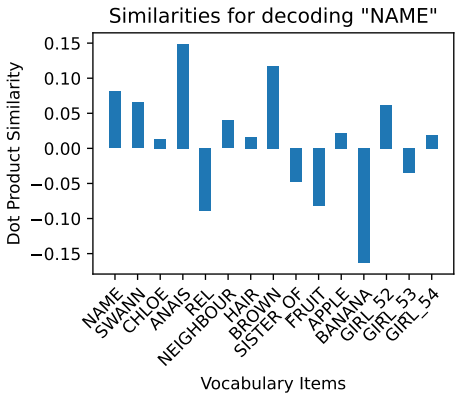


Subsequently there is input about Swann solely with respect to her hair and being a neighbour:

(16) REL ⊛ NEIGHBOUR + HAIR ⊛ BROWN + FRUIT ⊛ APPLE

This has the effect that the entity representing Swann has the properties associated with her hair and neighbourliness boosted. At this point, the state does not recall Swann's name (its similarity is below the threshold), as shown in (17) for NAME and in Fig. 3 for all properties.

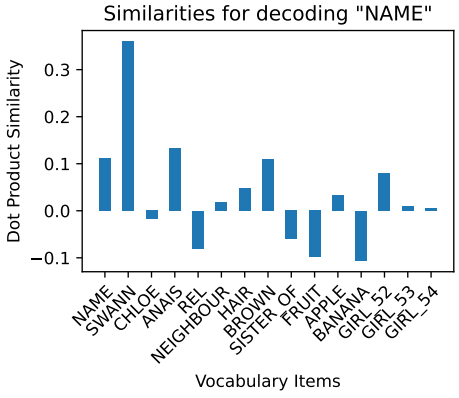
(17) The name "Anais" would be wrongly recalled, although very weakly (it is below the forgetting threshold of 0.3):



In terms of the sources of forgetting collected at the end of subsection 2.3 we can think of this as modeling forgetting by weight decay due to modification during activation.

Subsequently there is visual input about Chloé; Chloé and Swann remain similar, in other words Swann is associated (triggered as a possible topic) Finally, there is verbal input of Swann's name, which leads to it being recalled again as her name – see (18) and Fig. 4.

(18) The name "Swann" is regained:



5 Conclusions and Future Work

In this paper we have argued with reference to several concrete examples that dialogical semantics needs to be brain-oriented to account for a number of fundamental properties of cognition including forgetting and memory associativity. We have offered an initial synthesis of dialogue semantics where cognitive states are expressed in terms of external entities, though formulated with attention to the brain's memory structure, with a vector-based semantics that can be compiled into neurons and

triggered.

[7] All vectors are normalized, i.e., of unit length.

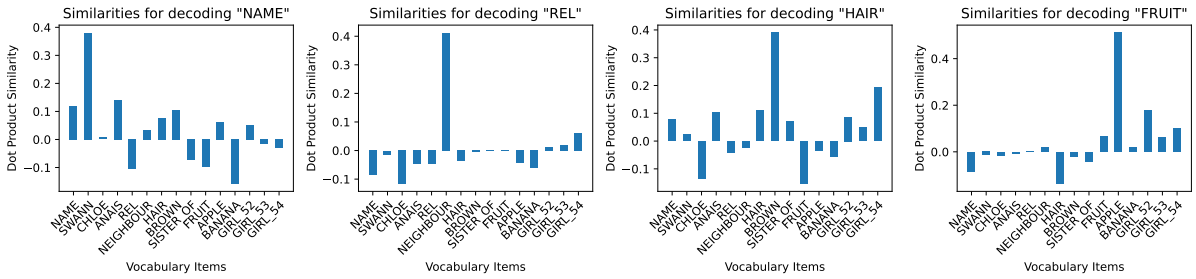


Figure 2: Unbinding the properties of the initial semantic pointer girl_52

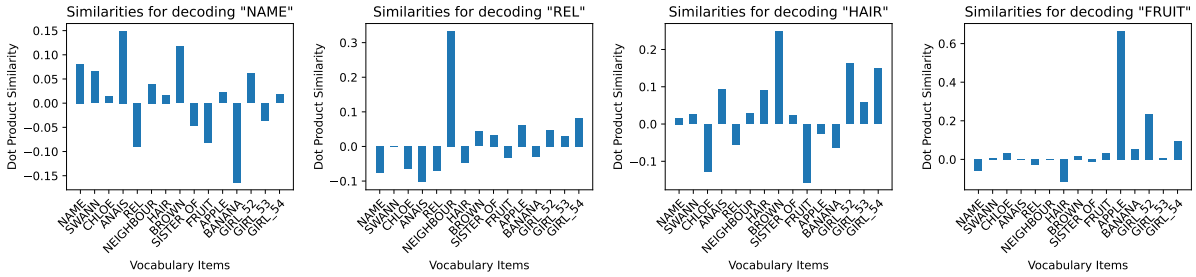


Figure 3: Unbinding the properties of girl_52 after updating REL and HAIR, but not NAME: the name 'Swann' counts as forgotten since it is not the most similar item any more and is below a similarity score of 0.3

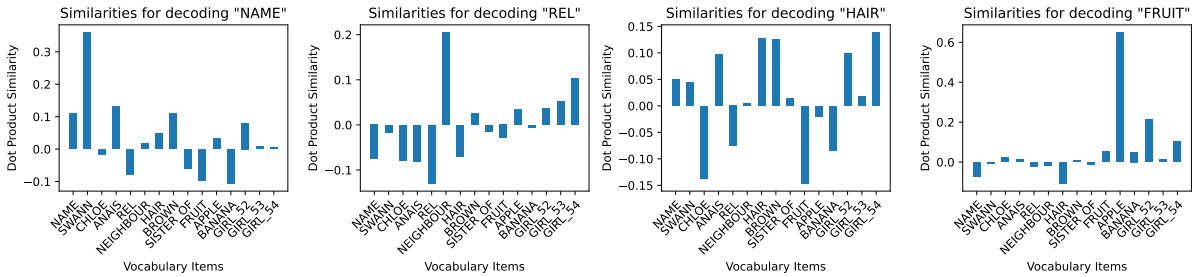


Figure 4: Unbinding the properties of girl_52 after updating NAME: the name is regained

neuron networks. The explanation we offer for the example we deal with in detail shows the need for a model that operates at various distinct levels, both the external and the neural. It is important to emphasize that such a model will clearly not be modular. For instance, our rule concerning associative topics makes reference to both a level of external content and to the neural level—more precisely the level where associations need to be computed, but the neural level is probably the more plausible level for this.

The neural model utilized here is very simplified, as we have pointed out, bypassing perception and working memory, in contrast to various existing work using the SPA architecture—see Borst et al. (2023) for a model demonstrating biological plausibility through the use of spiking neurons, and accounting for both human behavior and neuroimaging data across a whole task. In future work we hope to incorporate utterance processing and perception; an initial task being to provide neuralized versions of conversational rules.

Acknowledgments

Many thanks to Robin Cooper and Staffan Larsson for discussion and to three anonymous reviewers for TrentoLogue for very useful comments. We gratefully acknowledge support by the French *Investissements d'Avenir-Labex EFL* program (ANR-10-LABX-00) and by the *Deutsche Forschungsgemeinschaft* (DFG), grant number 502018965.

References

Alan Baddeley. 1988. Cognitive psychology and human memory. *Trends in neurosciences*, 11(4):176–181.

Alan Baddeley. 2012. Working memory: Theories, models, and controversies. *Annual Review of Psychology*, 63:1–29.

Giosuè Baggio. 2018. *Meaning in the brain*. MIT Press.

Christine Bastin, Gabriel Besson, Jessica Simon, Emma Delhaye, Marie Geurten, Sylvie Willems, and Eric Salmon. 2019. An integrative memory model of recollection and familiarity to understand memory deficits. *Behavioral and Brain Sciences*, pages 1–66.

William Bechtel. 2007. *Mental mechanisms: Philosophical perspectives on cognitive neuroscience*. Psychology Press.

Trevor Bekolay, James Bergstra, Eric Hunsberger, Travis DeWolf, Terrence Stewart, Daniel Rasmussen, Xuan Choo, Aaron Voelker, and Chris Eliasmith. 2014. Nengo: a Python tool for building large-scale functional brain models. *Frontiers in Neuroinformatics*, 7.

Peter Blouw, Eugene Solodkin, Paul Thagard, and Chris Eliasmith. 2016. Concepts as semantic pointers: A framework and computational model. *Cognitive Science*, 40(5):1128–1162.

Jelmer P Borst, Sean Aubin, and Terrence C Stewart. 2023. A whole-task brain model of associative recognition that accounts for human behavior and neuroimaging data. *PLOS Computational Biology*, 19(9):e1011427.

Robin Cooper. 2023. *From Perception to Communication: a Theory of Types for Action and Meaning*. Oxford University Press.

Robin Cooper and Jonathan Ginzburg. 2015. Type theory with records for natural language semantics. In Shalom Lappin and Chris Fox, editors, *The Handbook of Contemporary Semantic Theory*, 2 edition, chapter 12, pages 375–407. Wiley-Blackwell, Oxford, UK.

Nelson Cowan. 2001. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and brain sciences*, 24(1):87–114.

Chris Eliasmith. 2013. *How to Build a Brain: A Neural Architecture for Biological Cognition*. Oxford University Press, Oxford.

Chris Eliasmith and Carter Kolbeck. 2015. Marr's attacks: On reductionism and vagueness. *Topics in Cognitive Science*, 7(2):323–335.

Chris Eliasmith, Terrence C. Stewart, Xuan Choo, Trevor Bekolay, Travis DeWolf, Yichuan Tang, and Daniel Rasmussen. 2012. A large-scale model of the functioning brain. *Science*, 338(6111):1202–1205.

Jonathan R Epp, Rudy Silva Mera, Stefan Köhler, Sheena A Josselyn, and Paul W Frankland. 2016. Neurogenesis-mediated forgetting minimizes proactive interference. *Nature communications*, 7(1):10838.

Raquel Fernández. 2006. *Non-Sentential Utterances in Dialogue: Classification, Resolution and Use*. Ph.D. thesis, King's College, London.

Ross W Gayler. 2004. Vector symbolic architectures answer jackendoff's challenges for cognitive neuroscience. *arXiv preprint cs/0412059*.

Jonathan Ginzburg. 1994. An update semantics for dialogue. In H. Bunt, editor, *Proceedings of the 1st International Workshop on Computational Semantics*. ITK, Tilburg University, Tilburg.

Jonathan Ginzburg. 2012. *The Interactive Stance: Meaning for Conversation*. Oxford University Press, Oxford.

Jonathan Ginzburg, Raquel Fernández, and David Schlangen. 2014. Disfluencies as intra-utterance dialogue moves. *Semantics and Pragmatics*, 7(9):1–64.

Jonathan Ginzburg and Andy Lücking. 2020. On laughter and forgetting and reconversing: A neurologically-inspired model of conversational context. In *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue (WeSSLLI)*, Brandeis University.

Jonathan Ginzburg and Andy Lücking. 2022. The integrated model of memory: a dialogical perspective. In *Proceedings of the 26th Workshop on the Semantics and Pragmatics of Dialogue (DubDial)*, Dublin Technical University.

Jonathan Ginzburg, Chiara Mazzocconi, and Ye Tian. 2020. Laughter as language. *Glossa: a journal of general linguistics*, 5(1):1–51.

Jan Gosmann and Chris Eliasmith. 2021. CUE: A unified spiking neuron model of short-term and long-term memory. *Psychological Review*, 128(1):104–124.

Daniel L. Greenberg and Mieke Verfaellie. 2010. Interdependence of episodic and semantic memory: Evidence from neuropsychology. *Journal of the International Neuropsychological society*, 16(5):748–753.

Peter Hagoort. 2020. The meaning-making mechanism(s) behind the eyes and between the ears. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 375(1791):20190301.

Irene Heim. 1982. *The Semantics of Definite and Indefinite Noun Phrases*. Ph.D. thesis, University of Massachusetts, Amherst.

Ray Jackendoff. 2002. *Foundations of Language*. Oxford University Press, Oxford, UK.

Eric R Kandel, Yadin Dudai, and Mark R Mayford. 2014. The molecular and systems biology of memory. *Cell*, 157(1):163–186.

Staffan Larsson. 2002. *Issue based Dialogue Management*. Ph.D. thesis, Gothenburg University.

Staffan Larsson, Robin Cooper, Jonathan Ginzburg, and Andy Lücking. 2023. TTR at the SPA: Relating type-theoretical semantics to neural semantic pointers. In *Proceedings of the Fourth Workshop Natural Logic meets Machine Learning*.

Donald G. MacKay, Laura W. Johnson, and Chris Hadley. 2013. Compensating for language deficits in amnesia ii: H.M.'s spared versus impaired encoding categories. *Brain Sciences*, 3(2):415–459.

John Macnamara and Gonzalo E. Reyes, editors. 1994. *The Logical Foundations of Cognition*. Number 4 in Vancouver Studies in Cognitive Science. Oxford University Press, New York.

Emar Maier. 2016. Attitudes and mental files in discourse representation theory. *Review of Philosophy and Psychology*, 7(2):473–490.

David Marr. 1982. *Vision*. Freeman, San Francisco.

Brenda Milner and Denise Klein. 2016. Loss of recent memory after bilateral hippocampal lesions: memory and memories—looking back and looking forward. *Journal of Neurology, Neurosurgery & Psychiatry*, 87(3):230–230.

Dennis Norris. 2017. Short-term memory and long-term memory are still different. *Psychological bulletin*, 143(9):992.

Tony Plate. 1991. Holographic reduced representations: Convolution algebra for compositional distributed representations. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence*, IJCAI'91, pages 30–35.

Matthew Purver. 2004. *The Theory and Use of Clarification in Dialogue*. Ph.D. thesis, King's College, London.

Jamie Reilly, Jonathan E. Peelle, Amanda Garcia, and Sebastian J. Crutch. 2016. Linking somatic and symbolic representation in semantic memory: the dynamic multilevel reactivation framework. *Psychonomic Bulletin & Review*, 23:1002–1014.

Blake A. Richards and Paul W. Frankland. 2017. The persistence and transience of memory. *Neuron*, 94(6):1071–1084.

Craige Roberts. 1996. Information structure in discourse: Towards an integrated formal theory of pragmatics. *Working Papers in Linguistics-Ohio State University Department of Linguistics*, pages 91–136. Reprinted in Semantics and Pragmatics, 2012.

Daniel Saumier and Howard Chertkow. 2002. Semantic memory. *Current Neurology and Neuroscience Reports*, 2:516–522.

Emanuel Schegloff. 2007. *Sequence Organization in Interaction*. Cambridge University Press, Cambridge.

Kenny Schlegel, Peer Neubert, and Peter Protzel. 2022. A comparison of vector symbolic architectures. *Artificial Intelligence Review*, 55(6):4523–4555.

William Beecher Scoville and Brenda Milner. 1957. Loss of recent memory after bilateral hippocampal lesions. *Journal of neurology, neurosurgery, and psychiatry*, 20(1):1121.

Melanie J Sekeres, Gordon Winocur, and Morris Moscovitch. 2018. The hippocampus and related neocortical structures in memory transformation. *Neuroscience letters*, 680:39–53.

Pieter A. M. Seuren. 2009. *Language from within: Vol. 1. Language in cognition*. Oxford University Press, Oxford.

Larry R. Squire and John T. Wixted. 2011. The cognitive neuroscience of human memory since H.M. *Annual Review of Neuroscience*, 34:259–288.

Greg J Stephens, Lauren J Silbert, and Uri Hasson. 2010. Speaker–listener neural coupling underlies successful communication. *Proceedings of the national academy of sciences*, 107(32):14425–14430.

Timothy J Teyler and Jerry W Rudy. 2007. The hippocampal indexing theory and episodic memory: updating the index. *Hippocampus*, 17(12):1158–1169.

Paul Thagard, Laurette Larocque, and Ivana Kajić. 2023. Emotional change: Neural mechanisms based on semantic pointers. *Emotion*, 23(1):182.

Endel Tulving. 1972. Episodic and semantic memory. In E. Tulving and W. Donaldson, editors, *Organization of memory*. Academic Press, New York.

Victoria I Weisz and Pablo F Argibay. 2012. Neurogenesis interferes with the retrieval of remote memories: forgetting in neurocomputational terms. *Cognition*, 125(1):13–25.