Events as intersecting object histories: A new theory of event representation

Gerry T.M. Altmann
University of Connecticut

Zachary Ekves
University of Connecticut

We offer a new account of event representation based on those aspects of object representation that encode an object's *history*, and which convey the distinct states that an object has experienced across time – minimally reflecting the *before* and *after* of whatever changes the object undergoes as an event unfolds. Our intention is to account for the *content* of event representations. For an event that can be described as "*the chef chopped the onion*", the event *as a whole* is defined by the changes in state and location, across time, of the onion, the chef, and any instruments that (might have) mediated the interaction between the chef and the onion. We thus maintain that events are encoded as "ensembles of intersecting object histories" in which one or more objects change state. Our approach requires not just the distinction between object types and object tokens, but also between tokens and token-states (e.g. between that specific onion and its different states before, during, and after the chopping). These distinctions require an account of how object tokens are represented within the context of episodic and semantic memory, and how distinct object states are *bound* into a single object identity. We shall argue that the theoretical pieces, and their neural instantiation, are in place to develop a unified account of event representation in which such representation is simply a consequence of the mechanism for generating object tokens, their histories, and the binding of one to the other.

Keywords: event representation, object representation, episodic memory, semantic memory, types and tokens.

1. Introduction

Events change the world – even simple organisms like amoebae[1] encode, anticipate, and react to change (e.g. Saigusa, Tero, Nagaki, & Kuramoto, 2008). Our own (human) ability to notice, track, represent, recall, and communicate change is at the heart of human function – from our most peripheral sensory systems to the highest levels of cognitive representation and processing. Here, we consider the implications of representing change for theories of event cognition and event representation.

Mental Model theory (Johnson-Laird, 1983) and theories of situation models (e.g. van Dijk & Kintsch, 1983; Zwaan & Radvansky, 1998) consider event representation as a part of the construction of a mental situation – a "simulation" of relevant aspects of the world. A "situation" can be broader than an event in that it can consist of multiple, hierarchically-represented events, causally linked through their spatial and temporal relations (e.g. peeling an onion and then chopping it); however, a situation need not entail an event – a restaurant scene can correspond to a situation, even if the scene is unchanging (Barwise & Perry, 1983). In the event-indexing model of Zwaan, Langston, & Graesser

[1] Fortuitously, the word *amoeba* is derived from the Ancient Greek for *change*.

Gerry T.M. Altmann and Zac Ekves, Connecticut Institute for the Brain and Cognitive Sciences, and Department of Psychological Sciences, University of Connecticut.

ZE was supported by NSF IGERT grant DGE-1144399. Portions of this work have been presented at the Annual Meeting of the Cognitive Science Society, Philadelphia (2016) and at the Workshop on 'Event Representations in Brain, Language & Development', Nijmegen (2017). Section 3.3 extends work presented in Altmann (2017). The authors thank Yanina Prystauka, Emily Yearling, Eiling Yee, and the members of the 2015 Graduate Seminar on Event Cognition, UConn, for discussion and comments. We also thank GA Radvansky, Bernhard Hommel, and two further anonymous reviewers for their insightful comments on earlier versions of this paper.

Correspondence concerning this article should be addressed to Gerry T.M. Altmann, CT Institute for the Brain and Cognitive Sciences, 406 Babbidge Road, University of Connecticut, Storrs, CT 06269-1020, USA. E-mail*:* gerry.altmann@uconn.edu

(1995; see Zwaan & Radvansky, 1998, for further discussion), events are fundamental components of situation models, and are indexed, encoded, and updated on each of (at least) five dimensions: *time* (one event relative to another, and to the time of narration), *space* (spatial relations between protagonists, and between events in the situation model), *protagonists* (the animate or inanimate objects that are a part of the situation), *causation* (causal relationships between events or states), and *intentionality* (the protagonists' goals, which causally shape their actions). Changes along these dimensions determine how and when information across events is integrated and updated into the situation model. Much of the contemporary work on situation models has influences from Mental Models theory (Johnson-Laird, 1983), discourse models (e.g. van Dijk & Kintsch, 1983), Situation Semantics (Barwise and Perry, 1983), and Situation Calculus (McCarthy, 1963) which conceives of actions as a special kind of event, but events as not necessarily entailing actions (see below). *Event models* (see Radvansky & Zacks, 2011; 2014) focus more on the events themselves rather than, specifically, their integration into situation or mental models: Research on event models has tended to focus on how the continuous input stream is segmented into a series of discrete events, as a function of moment-by-moment changes in the predictability (at multiple timescales) of what may happen next (Event Segmentation Theory (EST); Zacks, Speer, Swallow, Braver, & Reynolds, 2007; see Reynolds, Zacks, & Braver, 2007, and Altmann & Mirković, 2009 for further discussion); how information is maintained/recalled within and across event boundaries; and the conditions in which interference obtains between different sources of information within and across those boundaries (e.g. The Event Horizon Model; e.g. Radvansky, 2012; for review of both theories together, see Radvansky & Zacks, 2017). Specifically, EST defines an event as "a segment of time at a given location that is conceived by an observer to have a beginning and an end" (Zacks & Tversky, 2001; p. 3). Event *models* (structured representations of events) enable predictions about what will likely happen next; within an event, prediction error will be low (i.e. what will happen next is relatively predictable), but prediction error will increase as the immediate future becomes less certain, as happens at the boundaries between events. Increases in prediction error trigger the updating of the current event model (if the prediction were perfect, there would be nothing to update – we return to this below), and this in turn leads to the perception of an event boundary (e.g. Zacks et al., 2007). The Event Horizon Model takes EST as its starting point and proposes a number of principles that govern how event models are structured, how they are recalled, and when shifts occur between one and another; sequences of events are organized according to their causal relationships; only the current event model is in working memory and the ease of retrieval of previous models Event segmentation serves as a form of chunking mechanism – limiting the degree to which information in one chunk interferes with information in another (Pettijohn, Thompson, Tamplin, Krawietz, & Radvansky, 2016) – but information that is common across events is subject to retrieval interference (e.g. Radvansky & Zacks, 1991).

These approaches to event cognition specify some of the "ingredients" of a full account of event representation: Events occur across time and space, involve protagonists and objects, have causal structure (something causes the event, or the event causes one or more subsequent events, and the objects taking part in the event generally have different causal roles), and (on occasion) are causally related to the protagonists' intentions (in the case of animate, and presumably sentient, protagonists). But what is missing from these approaches is an account of the *representational content* that distinguishes, for example, an event that includes *chopping* an onion from an event that instead includes *peeling* an onion. In each of these cases, the initial state of the onion is potentially the same, but the resultant (and intermediary) states differ (i.e. there are changes in its physical features and other properties) and these differences distinguish between the two events. And in the case of simply *lifting* an onion, the intrinsic states of the onion remain the same, but what changes are its extrinsic state (location) and the states of the lifter (the temporal characteristics of all these different events are also different, something that distinguishes, for example, traffic that *speeds* along a highway from traffic that *crawls* along the highway).

The relevance of object-state change for event representation is recognized within the linguistic tradition: Accounts of lexical semantics posit that the meanings of verbs include reference to objects' changes of state as entailed by the actions denoted by the verb (e.g. Dowty, 1979; Rappaport Hovav & Levin, 1998; Vendler, 1957; Warglien, Gärdenfors, and Westera, 2012). Such accounts can in principle capture the internal structure of events, with one or more participants in the event undergoing some change between the initial and end states (see especially Warglien et al., 2012). Thus, whereas the meaning of *"cut"* entails (amongst other things) a change in state from uncut to cut (possibly resulting in two separable parts of the original object), the meaning of "*fold*" entails a change in state from an original geometry to a new geometry in which one part of the object now overlaps the other. Crucially, however, one cannot simply equate events with the meanings of individual verbs or the changes that happen to individual objects – events are, as alluded to above, *ensembles* of object representations that encode the multiplicity of interactions and dependencies that, taken together, constitute the event. These ensembles are the closest correspondence in the account we develop below to the earlier notion of "event model". These ensembles capture not just the relationship between the participants in an event and the changes they undergo, but also the relationships between one event and another (see Section 3.6). Moens and Steedman (1988) developed an account of the mapping from language onto structures that reflect contingencies between events, which they describe in terms of a tripartite structure consisting of the goal event (the "culmination"), a "preparatory process" (loosely, a causal antecedent), and a "consequent state" – this structure is essentially a causal trajectory (as distinct from a temporal one, although they couch their account in terms of contingency rather than causality – see Section 6 for further discussion). And in respect of the distinction between types, tokens, and token-states, Kratzer (1995) points out a distinction relevant for understanding certain linguistic phenomena between "stage-level" and "individual-level" predicates that refer, broadly speaking, to properties associated with token-states (stage-level) or tokens and types (individual-level). While the linguistic (and philosophical) traditions have addressed issues of event representation from different perspectives, whether temporal, causal, referential, or to an extent representational, there remains the issue of how the human cognitive and perceptual apparatus does itself apprehend, encode, and retrieve event knowledge, and how it does so as a function of whether the event is experienced directly, or via language.

Within cognitive psychology, theories of event cognition rarely, if at all, make reference to the encoding, or retrieval, of changes of state (but see Sakarias & Flecken, 2019, who offer empirical data for the cognitive saliency of such changes both in verbal and non-verbal encoding of simple events), and yet, to understand a sentence such as "*the boy cut the paper*", or to understand the corresponding event immediately after directly observing it, requires the representation of the paper's states before and after its cutting (as well as requiring a representation of the boy and of the instrument of the cutting, and changes in their respective states time-locked to the transitioning states of the paper across the unfolding of the cutting event). Theories of embodied cognition (see Wilson, 2002, for review of their basic tenets) would propose that events are understood as mental "simulations" of actions and objects, grounding such representations in the substrates that support sensorimotor experience (e.g. Altmann, 1997; Barsalou, 1999; Barsalou, Simmons, Barbey, & Wilson, 2003). Such accounts postulate that an event is understood, when recalled or described through language, in terms of what would be the sensorimotor correlates of having directly experienced the event (or some abstraction of such correlates across experience). This is not to say that all possible experiences of the same event (direct or indirect through language) result in the same sensorimotor trace(s) – Tamplin, Krawietz, Radvansky, and Compeland (2013) demonstrated, in the context of a memory task, that reading about an event is not the same as participating in that same event; in the latter case, goal-directed behaviors that are absent during reading may lead to the suppression of task-irrelevant information which, during reading, may not be suppressed. Similarly, observing an event without participating in it, or reading about that event, or recalling it, all differ in respect of the event's temporal properties – temporal relations that are experienced in real-time during direct experience are not experienced in analogous time when recalling the event (see Section 4 below) nor when reading about the event (and the latter two's temporal dynamics also differ). Thus, the "representational products" of language comprehension will not be same as those due to the direct experience of an event, or due to its recall. Nonetheless, there will be commonalities across the different kinds of experience. Similarly for embodiment accounts of action understanding: actions are understood through their representation in the sensorimotoric cortical substrates that control action and perception (cf. procedural accounts of knowledge representation, e.g. Winograd, 1972), and their understanding depends on task context (such as whether an action is being planned in service of some goal, or whether it is merely being observed). We return to action control and action planning, and the representations that support these, in Section 5 when we discuss Hommel, Müsseler, Aschersleben, & Prinz's (2001) Theory of Event Coding. Like actions, objects are understood and represented through their sensorimotoric properties, i.e. in terms of the perceptions and interactions they afford – their *affordances* (e.g. Gibson, 1979; Glenberg, 1997). But while such embodiment, or sensorimotoric, accounts maintain that mental states (and the simulations that are entailed) will reflect first the 'before', and subsequently the 'after', of an event (either when experiencing the event directly or when comprehending a linguistic description of that event), some additional theoretical machinery is required to explain how, when the sensorimotor experience of chopping an onion is simulated or re-enacted, the resultant state of the onion is not divorced from its initial state. Without this additional machinery, simulation becomes the representational equivalent of a "running commentary", and without some memory of prior states, the simulation ends up with a representation of the world divorced from any history of how it came to be that way; *understanding* an event minimally requires not simply knowing what is the resultant state, but knowing that it differs from an earlier state. In other words, an integral part of event understanding is knowing that an event has occurred, and that the world was in one state but now is in another. Moreover, and this is true of any theory of event representation, not just embodiment theories: The sensorimotor (or other) encoding must reflect also the *specific instantiations* of the objects that took part in the event – i.e. individuated *tokens*, and not just the *types* of objects that took part. That is, the event representation, however it manifests in the brain, must reflect the episodic and temporal instantiations of the participants, and their interactions, in the event. Here we distinguish the instantiation of an actual event in an episodic memory structure from *generalized event* knowledge that reflects "semantic" knowledge of how, in general, the world changes as a function of typical protagonist interactions (e.g. Elman & McRae, 2019; McRae, Hare, Elman, & Ferretti, 2005; Metusalem, Kutas, Urbach, Hare, McRae, & Elman, 2012). This latter knowledge is related to the concepts of *schema* (Bartlett, 1932) and *scripts* (Schank & Abelson, 1977), to which we return below.

In the following sections we posit a theoretical mechanism, and its consequences, that would enable the representational distinction between types, tokens, and the transformations that those tokens undergo across time, and which lead to those tokens being in different token-states at different times. We propose object histories, and how over some part of that history objects change state, as a fundamental primitive of event representation. In addition, we shall argue that the representations of an object's initial and final states (and possibly its entire trajectory through state space) are *simultaneously active* (albeit in different proportions) during recall, communication, or comprehension of events in which that object participated. Much of what follows emphasizes *object* representation more than *event* representation (in fact, object representation in service of event representation). This reflects our argument that object representations play a fundamental role in event cognition. Similarly, much of what follows emphasizes the direct experience of external events. However, our intention is to specify an account of event representation that transcends modality – we necessarily blend theoretical concepts from visual object perception with concepts from linguistic representation and, indeed, cognitive representation. We apply principles deduced from direct experience to cases where an event is experienced through narrative (c.f. sentence comprehension) or is imagined in the absence of perceptual input. Importantly, constraints on how we must comprehend events through narrative will also motivate parts of our account as they apply to direct experience. Related, we necessarily blend concepts from perception and the real-time uptake of information from the environment (as when we directly observe an unfolding event) with concepts from memory and the relationship between real-time encoding and longer-term representation (as when we comprehend a sentence that describes an event – a proxy for direct observation of that event, or as when we recall an event). Before considering our account in detail (henceforth "IOH", for Intersecting Object Histories), we review the currently available empirical evidence in support of this simultaneous activation of multiple object-state representations during event comprehension.

2. Multiple object-state representations during language comprehension

There are *a priori* reasons for supposing that multiple object-state representations must be *simultaneously* activated during event comprehension. The fact that we can use language to direct mental time travel back to prior states of the world and its objects indicates that these representations of alternative states can be independently selected. Of course, this does not entail that the *before* and *after* states of an object need be *simultaneously* active (let alone the intermediary states also): In principle, after hearing "*the chef chopped the onion*" one might maintain just a single representation of a chopped onion; if the sentence continued "*but first she weighed the onion*", one could *recreate* the appropriate representation of the unchopped version, in much the same way as one might, when viewing a photograph of a chopped onion, *infer* that it had existed at some prior time in its intact form. In this case, one need never hold both representations simultaneously in mind. This would predict that only a single representation would be activated during the second clause, regardless of whether it began with "*and then*" or "*but first*". Below, we review evidence to the contrary. But regardless of the evidence, if just a single representation of the appropriate object state was active after hearing "*the chef chopped the onion*", how could one understand *what had just happened*? To do that would require knowing, among other things, that the currently chopped onion had existed in some particular other state beforehand. And if, instead, the burden of explanation were shifted away from a representation of the prior object state (i.e. the onion before it was chopped) and onto some representation of the action denoted by the verb "*chopped*", how could such a representation denote the change in state that the world has just undergone *without* entailing some representation of the prior state of the world? And when observing events directly, how else could one know that an event had in fact occurred, and importantly, *what* had just occurred, if one did not have a memory for the before and after? A theoretical understanding of the phenomenology of event comprehension (whether through narrative or through direct experience) would seem to require multiple object-state representations (or a mechanism that, when cued, would activate some representation of the transition from one state to another – again, requiring multiple object-state representations). This in turn leads to the following question about the ontology of action representation: if, as the evidence below suggests (and theory requires), the cognitive system represents the spatiotemporal properties of object states, and does so in the context of other objects' spatiotemporally defined states, what more is required, representationally speaking, to encode *action*? Before considering this question further, we turn to the empirical evidence for the representation of multiple object states during event comprehension.

Altmann and Kamide (2009), and subsequently Hindy, Kalenik, Altmann, and Thompson-Schill (2012) explored the empirical consequences for the cognitive system of having to track multiple representations of the same object as it changed from one location to another (Altmann & Kamide, 2009) or from one state to another (Hindy et al., 2012). Their assumption was that the representation of an object in one location or state and the representation of that same object in a different location or state must be distinct (if not, how could one refer to one location/state or the other without confusing them?). Altmann & Kamide (2009) proposed that, like multiple meanings of an individual word, or multiple compatible completions of an unfolding word fragment, these multiple, mutually exclusive, representations may be in *competition* with one another when one representation must be chosen to the exclusion of the other(s). Evidence of such competition would provide strong evidence for simultaneous activation of object-state representations; for competition to occur between representations of object state, the representations must be simultaneously active.

Hindy et al. (2012) investigated competition between object-states by contrasting sentence pairs such as the following:

(1) *The chef will chop the onion. And then, she will smell the onion.*
(2) *The chef will weigh the onion. And then, she will smell the onion.*

At the final "*onion*" in (1) there is an ambiguity regarding which is the appropriate state that is intended (in this case, the chopped state). There is no such ambiguity in (2), in the sense that the onion is (presumably) in the same state as it was before the weighing (and while it changes state insofar as it changes location during the weighing, its intrinsic state – its geometry and other featural properties – remain largely the same; we thus consider (1) to entail "substantial" change, and (2) to entail, at most, "minimal" change – see below). In (3) below, there is again a "state ambiguity" at the final "*onion*", but this time, the intended state is the unchopped state:

(3) *The chef will chop the onion. But first, she will smell the onion.*

In two fMRI studies, Hindy et al. (2012) contrasted sentences similar to those shown in (1) to (3). In their first study, they contrasted "*chop/weigh*" and "*And then…/But first…*" in a 2x2 design. Participants read each pair of sentences and responded if the second of the pair was impossible given the first (a typical foil might be "*The man smashed the glass. And then, he poured the wine into the glass*"). For each participant, Hindy et al. also established which voxels in the brain were most responsive to conflict in a Stroop color-word interference task (see MacLeod, 1991, for review). In our version of the task, participants had to respond to the color in which the word was printed; to do so required resolving the conflict between this color (e.g. red) and the meaning of the color-word (e.g. "*green*"). January, Trueswell, & Thompson-Schill (2009) had previously found that these same voxels (i.e. sensitive to Stroop color-word interference, and in left posterior ventrolateral prefrontal cortex, pVLPFC) were active when syntactically ambiguous sentences were initially misinterpreted and required subsequent correct resolution. More generally, left pVLPFC has been found to be sensitive to semantic competition (e.g. Thompson-Schill, Bedny, & Goldberg, 2005), to selection of context-appropriate meanings of ambiguous words (Metzler, 2001; Hindy, Hamilton, Houghtling, Coslett, & Thompson-Schill, 2009), and to completion of sentences that permit multiple alternative responses (Robinson, Blair, & Cipolotti, 1998; Robinson, Shallice, & Cipolotti, 2005). Our rationale was that if these same Stroop-sensitive voxels in pVLPFC are also sensitive to the distinction between (1) and (2) above, this would be evidence of competition between object-state representations. This is in fact what we found: Not only was there a difference between (1) and (2) in the activation of Stroop-sensitive voxels, the magnitude of this difference was predicted by separate ratings of the degree to which the critical object on each trial was deemed to be changed by the event (for example, chopping an onion changes an onion by more than chipping a glass changes a glass – even for the "minimal" changes, there was variation in the degree to which the objects were rated as changing). We interpret this last observation in the context of models of distributed memory (e.g. Allport, 1985), in which the greater the featural overlap between two representations, the greater their overlap within the representational substrate, and consequently, the less scope there is for each representation to compete with the other – only non-overlapping (i.e. distinguishable) components of the representation can actively compete.

We concluded that the differential activation observed in this study was an indication of competition between alternative object-states. We argued in Hindy et al. (2012) that these effects were incompatible with explanations in terms of differences in memory load (left pVLPFC is insensitive in other studies to memory load manipulations) or the maintenance of only a single object-state representation accompanied by on-the-fly computation of the contextually appropriate state. The latter argument relied on the finding that the Stroop-sensitive voxels were as responsive to the "*chop*"/"*weigh*" difference after "*And then*" as they were after "*But first*" (example (3) above); if only a single representation of the onion's state was represented after "*And then*" (e.g. its final state), or after "*But first*" (e.g. its initial state), there would be no reason for a competition effect in the second sentence (Solomon, Hindy, Altmann, & Thompson-Schill, 2015, demonstrated that the competition effects we observed in Hindy et al. (2012) were indeed due to retrieval processes at the end of the second clause).

In a second study, Hindy et al. (2012) contrasted (4) and (5) below:

(4) *The girl will stamp on the egg. And then, she will look down at the egg.*
(5) *The girl will stamp on the penny. And then, she will look down on the penny.*

In these cases, unlike (1) and (2), the verb remained the same but the object ("*egg*" or "*penny*") changed (given what we know about these objects and their relative fragility). Exactly the same pattern of fMRI results was found. Importantly, in this study, neither the sentences nor foils included a reverse temporal connective ("*But first…*"), meaning that participants never had to explicitly retrieve the initial intact state of the egg. And yet, we still observed the same competition as in the first study.

A subsequent study (Solomon et al., 2015) using the same items and reading task as in Hindy et al. (2012) added an additional condition to ask whether the effects we had observed in Hindy et al. (2012) were due to the representation of multiple distinct states of the *same* object or due to the representation of multiple distinct states regardless of whether they were represented across the same object or different objects. We contrasted the following conditions:

(6) *The chef will weigh an onion. And then, she will smell the onion.*
(7) *The chef will chop an onion. And then, she will smell the onion.*
(8) *The chef will chop an onion. And then, she will smell another onion.*

Crucially, in both (7) and (8), an onion is chopped, requiring the representation of both the unchopped and chopped states (unlike in 6). In (7), the onion that is smelled at the end of the second sentence is the same onion as was chopped in the first sentence, and thus the chopped state must be retrieved at the expense of, and hence in competition with, the unchopped state (they are mutually exclusive). In (8), however, the state of the onion referred to at the end of the second sentence ("*another onion*") is not in competition with the unchopped (or chopped) state of the onion referred to in the first sentence. Our hypothesis here was that if our previously observed effects were due to competition between object states that pertain to the *same* object, we should observe such competition in (7) but not in (8). And of course, we anticipated little or no competition in (6) where the onion undergoes minimal or no change. If, however, our prior results were due to the representation of multiple states *regardless* of whether they were bound to the same object, we should observe competition in both (7) and (8). We found, in fact, that the effect of competition in (7) was completely absent in (8), indicating that the competition obtains only between distinct representational states of the *same* object, and does not obtain if these distinct representations correspond to distinct objects. This finding also rules out an account of the competition effect as being due to competition between broader representations of the situations (rather than between the specific object's states) before and after the event; the situations were identical except for the same/different token manipulation.

In this last study, we also found that differential activation of stroop-sensitive voxels in left pVLPFC, as a function of degree of change, correlated with differences in early visual cortex. We interpreted this to suggest that the competition we observed in this and our previous studies (Hindy et al., 2012; see also Hindy, Solomon, Altmann, & Thompson-Schill, 2015) was based on alternative sensory (and presumably sensorimotor) features associated with the different object states. Following Grill-Spector and Malach (2004), we assume that whereas late visual cortex encodes abstract visual information relevant for object identity (invariant to changes in viewpoint as well as invariant to object state; Hindy et al., 2012), early visual cortex encodes and retains information about specific visual features (that distinguish between one visual state and another) even in their absence (Harrison & Tong, 2009). Thus we interpret the effects we observed in early visual cortex as indicative of the encoding of sensory features that distinguish between one object state and another.

To summarize the empirical data: We consistently found the same result in our fMRI studies (Hindy et al., 2012; Hindy et al., 2015; Solomon et al., 2015); voxels sensitive to Stroop-conflict are sensitive to object-state changes, and in particular, to the *degree of change* that the object underwent in the event described by the language. The most parsimonious explanation of these results, given what is known about pVLPFC, is that this sensitivity, like other examples of pVLPFC activation in prior studies, reflects *competition*; in this case, between multiple representations of object-state. We therefore conclude that, as predicted, event comprehension (at least as operationalized in these studies) does indeed entail simultaneous activation of multiple object-state representations.

3. IOH: The Intersecting Object Histories account of Event Representation

Having established the *a priori* need for multiple object-state representations, and reviewed the available empirical evidence for their simultaneous activation during event comprehension (at least as evidenced during language understanding), we turn now to why we believe that object-state representations (more precisely, object histories – trajectories of object state through space and time) are the fundamental representational primitive of event representation.

The central claim of IOH is that event comprehension and encoding (whether occurring as we directly experience an event or as we learn about it through language) are built upon dynamic representations of intersecting object histories; an individual event is represented through an *ensemble* of such representations. These representations are dynamic not because they reflect individual objects' trajectories through space, featural state, *and time*, but because the representations themselves change, as we shall describe below. These representations of intersecting object histories capture the spatiotemporal contiguities between different objects and their respective changes of state. They intersect by virtue of their co-occurrence in (near) space, time, and representational substrate. This latter claim reflects one of the basic tenets of contemporary models of distributed memory (e.g. Allport, 1985) – that concepts that overlap in aspects of meaning also overlap in the neural substrates that support the encoding of those aspects of meaning. This has the natural consequence that objects become related as a function of those intersecting object histories and will, through that representational overlap, come to prime one another (cf. Kalénine, Mirman, Middleton, & Buxbaum, 2012; Moss, Ostrin, Tyler, & Marslen-Wilson, 1995; Yee & Sedivy, 2006; Yee et al. 2010). Below, we describe a theoretical framework in which such physical overlap, as well as abstraction from episodic to semantic representation, is a natural consequence of the encoding of tokens and token-states. For now, the critical point is that ensembles of intersecting object histories constitute the representational primitives of event understanding – that is, other aspects of event structure, such as participant roles, causation, and indeed action, are representations that emerge from these representational primitives through a process of abstraction. And while object histories may decompose into lower-level features that define each object, its spatiotemporal dynamics, and the state it is in or has been in, these lower-level features do not, in and of themselves, define aspects of event structure *independently* of the object and its history.

In the following sections, we assume a number of theoretical principles, outlined in Altmann and Mirković (2009), which are based on theoretical insights derived from Elman's (1990) implementation of a simple recurrent network (SRN) and shared with other dynamical systems (e.g. Elman, 1990; 1993; Tabor & Tanenhaus, 1999). Our focus on theoretical principles embodied in the SRN is not a claim that the human mind is an SRN, or that object-state representations are encoded within an SRN; rather, the SRN embodies principles that capture important aspects of human cognition (see Altmann & Mirković, 2009, for further discussion of these principles, and Reynolds et al., 2007, for their application to event segmentation, and Elman & McRae, 2019, for an implementation in an SRN of generalized event knowledge and its deployment during event cognition. In the latter model, the input to the model is given in terms of participant roles (such as agent, patient, instrument, etc), actions (corresponding to e.g. cutting, giving, reading, etc.), and context (location, time); this contrasts with our approach here, in which categories such as participant roles and actions are emergent rather than primitive, and in which object tokens and token states are central to the account). Specifically, we "borrow" from the SRN and subsequent work the assumption that the cognitive system engages in (among other things) predicting upcoming input; that its input at any one moment in time is a product of both the concurrent external input and its prior internal states reflecting past inputs (*recurrence*; see below); and that the discrepancy between its predictions and the actual input (i.e. *prediction error*) results in modifications to the internal structure of the system, and the emergence of hierarchical representations, that better reflect the dependencies between the successive inputs that it receives across time (c.f. experiential learning). It is these principles, not the SRN itself, which we believe underpin human event cognition. Indeed, there are a number of limitations inherent to the original SRN architecture and associated algorithms that make it unsuitable as a computational instantiation of aspects of the theory we shall outline below (beyond its implementation within a 'toy domain'), including limits on scalability, temporal resolution, and its relationship to the putative brain mechanisms that underpin human cognition. The discussion below will draw on insights and related findings from the neurobiology of memory; our aim is to develop an account of event representation – the IOH – which is not only computationally plausible, at least in principle, but for which there may also be a plausible neurobiological grounding.

3.1 The representational status of actions and participant roles

The claim that participant roles and actions are representations that emerge through a process of abstraction across our experience of intersecting object histories, might at first seem untenable: On the face of it, more is needed to explain event understanding than simply knowing that an object changed state: To understand what happens to the onion in the event described as "*the chef chopped the onion*" requires more than a representation of the onion at one moment in time, and the representation of it at a subsequent moment in time in a different state. In addition to the change in state of the onion (tracked across time), we need, of course, to represent the chef also (hence the notion that event representations are *ensembles* of representations). But should we not also represent the action that caused the change in state of the onion, as well as the fact that it was the chef that executed this action, and most likely (if we did not perceive all of the event first-hand, or if we learned about it through language) with an instrument with which to do the chopping? While it is indeed the case that we need to represent the chef and the likely involvement of a knife or other instrument, we do not in fact need to represent the action itself or the chef's role as the agent of the action; the actions we observe or learn about through language are no more than changes in object states through time: the chef changes state, the instrument used for the chopping changes state, the onion changes state, and the temporal properties of these changes, as well as their spatiotemporal contiguities, define both the action and, indeed, the event itself. The knife, in a chopping event, for example, determines our perspective on "what happened"; it is an integral defining part of the event (if it is merely waved towards the onion, which then magically becomes chopped, we would interpret that event differently) – but the knife is itself undergoing changes (in location) across time which intersect with changes that the chef and the onion each undergo. Each participant in the event follows a *trajectory* of change in state across time[2], and it is the ensemble of these trajectories (that is, the ensemble of their representations) that constitutes the event representation; the *intersection* of these trajectories constitutes the event within this representation. This is not to say that all trajectories are equal – for the onion, there is a culmination point and a resultative state; the chef's trajectory of change does not necessarily lead to a resultative state that is different from her initial state. But this difference is the basis for the implicit encoding of different kinds of participant roles, which we discuss below. For now, we maintain that nothing more is required to represent an event than the representation of the multiple objects and their respective states across time and space.

This last claim may seem at odds with the intuition that actions are themselves representationally basic. We define a representational description as basic if representations at that level do not decompose into constituent representations which themselves are not described at that same level of description. Informally, we can ask what it is, in the outside world, that we would *measure* to describe an action or event that has taken place; there are parameters we might measure, but our claim, as we proposed earlier, is that these parameters are uniquely defining of objects and their spatial, intrinsic, or temporal properties – that is, events and actions derive from these other measurable qualities (in fact, measurable by our perceptual systems) which do not, independently of objects and their properties, define actions or events. In this sense, object histories should more correctly be described as representationally "more basic" than actions or events. Another perspective on this same argument is exemplified by the individual frames of an animated cartoon (or indeed, of a movie). In the three successive (and hence only slightly different) stills shown in Figure 1, there are no actions – just objects changing their spatiotemporal configuration. We "project" actions onto this dynamic stimulus (when the stills are presented in succession at e.g. 24 frames per second), and our claim is that, just as actions are not directly represented in the perceptual stream[3], so they are not representational primitives – rather we contend that they are emergent abstractions across the input, in much the same way as representations can emerge as abstractions across a temporally varying sequence of words given to an SRN (Elman, 1990) – there, the distributional characteristics of the input (an unparsed sequence of words across time) are abstracted across to yield higher-level hierarchical representations that reflect the distributional characteristics (and hence categories) of words. The emergence of these categories is accompanied, in the SRN, by the ability to better predict the likely characteristics of the upcoming input. That is, these emergent representations are not simply a non-functional by-product of some other function, but rather their emergence reflects the manifestation of new abilities, including the ability to generalize in ways that would not be possible without that category structure. The emergent representation of hierarchical and overlapping structure has the advantage also of offering the equivalent of a form of representational compression – information common to distinct entities need not be duplicated, and their overlap in representational (and potentially physical) space reduces the informational burden. Abstraction across experience thus enables compressed representations relative to an unstructured list of individual experiences. Abstract representations are key to the ability to generalize to novel episodes of experience. Thus, we do not claim that actions have no role to play in event representation. Rather, we claim that actions, as representational components of *event representation*, are emergent abstractions across the more basic intersection of object histories.

There are advantages to rethinking the (traditional) status of actions within event cognition: If they *are* a representational "building block" for event representation, how are we to represent events due to one inanimate object causing a change in another – as when a tree crashes down on a powerline? Typically, we do not consider such events as entailing action – actions entail intention (c.f. Pacherie, 2008), and intentions entail goal states rather than actions *per se*. Of course, this may just be a matter of nomenclature (if we permit the tree to have "acted upon" the powerline). Regardless of nomenclature, parsimony favors a single theoretical mechanism for encoding events that does not require one kind (action-based) for volitional events, and another kind (object-based) for non-volitional events. In Section 5 below we consider the role of actions, and more importantly, action goals from the

Figure 1. Three successive frames from *Garfield and Friends* (1990). We interpret as action the spatiotemporal dynamics of the perceptual input across successive frames – the changing physical configurations of the depicted objects through time are interpreted as Garfield hitting Odie with a pie, and simultaneously, Odie hitting Garfield with a pie. Original Still Images © Paws Inc. Reproduced with permission.

[2] We take changes of location to be changes of state also, although the representational consequences are different; extrinsic changes in state – i.e. changes in location, or in the physical context in which an object is observed, require encoding of that context. We return to this point below when discussing the construction of tokens and token-states on-the-fly; the same mechanism that allows such on-the-fly construction enables also the encoding of an object token's context

[3] Motion *is* represented in the perceptual stream, with neural mechanisms specialized for, or at least sensitive to, biological motion (e.g. Grossman, Donnelly, Price, Pickens, Morgan, Neighbor, & Blake, 2000). However, motion is not action, even if for some actions it is a component part.

perspective of action planning, action understanding, and the relationship between action and perception (cf. Theory of Event Coding (TEC); Hommel, Müsseler, Aschersleben, Prinz, 2001); there, goal states appear to be the primary cognitive representation, with actions (specifically, their manifestation as motor movement) merely an executable means to an end.

Can the same be said about the causal relations that define the roles of each event's protagonists (e.g. agent, patient and other roles such as instrument, etc.)? At issue is whether some explicit representation of individual participant roles and their causal relations across time is actually required, including likely but unstated participants (the assumed knife in the chopping example), or whether such information can remain implicit in the encoding across time of object-state changes. Of course, such roles/relationships are themselves abstractions across multiple experiences. These abstractions encode the likelihood that e.g. a chef will slice something rather than punch something, or will more likely chop onions than chop wood, and will more likely do that with a knife than with an axe (see Ferretti et al., 2001, and McRae et al. 2005, for empirical evidence concerning generation of such expectancies during sentence comprehension). In other words, these abstractions encode the *contingencies* between objects and their interactions with other objects, and through doing so *implicitly* encode the thematic, and indeed, causal relationships between the participants in the event. This raises the question of how young infants, who are necessarily limited in the nature of their experience, are able to recognize causal participant roles (c.f. Rochat, Striano, & Morgan, 2004). However, this sensitivity to participant roles need not mean that these are the *same* participant roles – supporting the same generalizations – that adults are sensitive to; in the absence of experiential honing, infants' notions of different causal roles may be broader than adults' (in much the same way that their earliest words' meanings often reflect over-generalizations relative to adults' interpretation of the same words). The issue, then, is what kind of representational framework will permit the encoding of the relevant contingencies across time (and space), whether in the infant or the adult?

In Altmann & Mirković (2009) we presented an emergentist account of thematic roles, and thematic role assignment, in the context of predictive encoding in language. In that account, the prediction at each moment in time of what input may come next encodes exactly the contingencies that are required to capture causal role information – anticipating onion (among others) after "*the chef will chop*" constitutes the representation, based on prior experience, of the spatiotemporal contingencies between chefs, chopping, and the things that are generally chopped in the context of chefs (hence wood being less likely). The encoding of those contingencies includes other objects that may have been experienced concurrently, such as instruments of the chopping. And regardless of how the input actually unfolds, the predictive (and dynamic) encoding that accompanies the unfolding of the input (through whatever sensory medium) reflects these previously experienced contingencies. That is, it reflects the participant roles that accompanied such experience, where participant roles are simply contingencies between one participant and others, and the spatiotemporally contiguous changes in state (physical or psychological) that they underwent or could undergo (this emergentist approach stands in contrast to the claim that roles such as agent and patient might be innately specified in certain ways; e.g. Pinker, 1984). As the input does unfold, this encoding of contingencies interacts with co-occurrences, in space and time, which are actually experienced during that unfolding; not all co-occurrences are equal, that is, some relationships in that moment-to-moment experience will be more salient than others, reflecting the greater informativity of those relationships in respect of constraining what may come next: If someone brings a coffee cup to their lips, the relationship between that person and the coffee cup is quite different from that between the person and, for example, a painting on the wall behind them; one is informative of how the world will unfold, and the other is not. However, we anticipate a gradient of such "relevance", modulated by existing knowledge (if we recognize that the painting is new to that room, and happens to be Hockney's "Portrait of an Artist (Pool with Two Figures)" – the most expensive painting sold at auction, at least at the time of writing – the co-occurrence of the person and the painting would now become more salient than the Starbucks coffee cup being raised to her lips. Thus, while certain participant roles may become more salient because, experientially, they prove more informative (which these might be is not relevant to the discussion) in respect of the subsequent unfolding of the world (real or mental) being experienced, others may become more salient because of their informativity in the moment.

3.2 Representational primitives of event encoding: Objects as trajectories through space and time

One way or another, events need to be encoded in a manner that captures (i) the participants in the events, (ii) the initial and end states of those participants, as well as intermediary states, (iii) the spatiotemporal contingencies between both individual and multiple participants' state changes (i.e. their intersecting *trajectories* through space and time), and related to this, the causal relationships between the event participants (including participants' *intentions*, insofar as they can be inferred) as well as the causal relationships between any sub-events. The latter requires that the encoding supports hierarchical event structure. A further requirement of the encoding of events is that a distinction must be supported between specific knowledge of the details of an actual event on the one hand (e.g. this particular chef chopped that particular onion), and on the other, generalized knowledge about typical events and the typical participants that participate in them (e.g. onions are often chopped; chefs often cook with onions). This latter distinction corresponds, loosely, to the distinction between episodic and semantic knowledge. In this section we consider the necessary ingredients of event encoding, and the related phenomena that require theoretical explanation.

Clearly, a fundamental primitive of events, and of cognition more generally, is the *object representation*. One of the puzzles concerning object representation has been to understand how such representations persist across time in the face of changes to the represented object (e.g. that the chopped onion is still an onion and, more specifically, the same onion as it had been before being chopped: for review, see e.g. Carey & Xu, 2001; Scholl, 2007). Much discussion of this has taken place in the context of visual cognition, and *object files* (Kahneman, Treisman, & Gibbs, 1992). Typically, object files are viewed as "mid-level" representations of physical objects that mediate between low-level visual features and high-level object recognition (e.g. Scholl, 2001). Object files reflect the perceptual experience of objects – their sensory features and changes to those features across time, but they do not reflect stored long-term knowledge of that object. Object persistence (the continuing identity of an object across time) arises primarily through spatiotemporal continuity. As Scholl (2007) points out, however, object persistence through such continuity seemingly breaks down in the face of significant property change: To borrow an example Scholl cites, from Hirsch (1982); a car that has been crushed into a cube of metal is no longer a car even if certain properties such as its history, remain. Carey and Xu (2001; see also Carey, 2011) review a number of studies showing that in infants aged around 10 months, object persistence seems to be predominantly tied to spatiotemporal continuity, while infants aged around 12 months are able to use what Carey & Xu (2001) refer to as "kind-based" information, corresponding loosely to conceptual knowledge encoded in semantic memory. The crushed cube-of-a-car is therefore no longer a car because it violates kind-based knowledge of what cars are. As such, this example demonstrates the potential dissociation between an object's identity (its enduring history) and that object's kind. But unlike crushed cars, crushed garlic is still garlic, so some changes in token-state appear to preserve both token identity and token type, whereas others (e.g. crushed cars) maintain token identity but change token type. For now, we shall assume that kind-based knowledge of garlic permits its crushed state as being of the same *kind*, while kind-based knowledge of cars does not (we return to this issue below when discussing semantically-mediated object persistence). Critically, while object persistence, and specifically object identity, might come about through spatiotemporal continuity (given the caveats just described), such continuity is not an *explanation* – to explain continuity of identity requires a mechanism for the memory trace at one instant in time to be bound to the memory trace at the previous instant in time and to the perceptual trace (if there is one) at the next. While Pylyshyn's (1989) FINST mechanism is capable of maintaining object identity through spatiotemporal continuity (imagine a finger tracing a moving object – the finger-as-index ensures continuity of identity), there is no account of how, when features of an object change in the *absence* of spatiotemporal continuity, object identity is maintained (beyond brief absences). FINSTS are primarily spatial, and hence discontinuities in spatial position relative to the physical context pose a challenge to the mechanism. As we shall propose below, semantic and contextual mediation are required in such cases, and FINSTS are "blind" to the content of the visual information they index. Moreover, the FINST mechanism was not intended to explain how object identity is maintained during language comprehension – where spatiotemporal continuity of referenced objects is lacking in the input. We return to an alternative mechanism, able to operate both in the visual and linguistic domains, in the next section where we discuss in more detail how, during event comprehension, we represent instantiated object tokens, and how we create such tokens *on-the-fly*.

While object files may provide a representational medium for physical objects accessible to the visual sense, something more akin to kind-based information is required to explain how we instantiate objects as tokens during event comprehension in the *absence* of the corresponding real-world, perceivable, objects – for example, when hearing or reading about an event. This representation of an instantiated object has to encode kind-based information, but something more is required to explain how it reflects an *individuated instance* of the object (as distinct from a generic), and how such objects can persist across change (albeit linguistically described, or in memory) in the absence of the same spatiotemporal contiguities that are afforded by real-world objects. We hypothesize that the same representations and principles that support object persistence in the observable real-world also support object persistence in the mental world (for example, as constructed in response to language input). As in the case of objects whose physical instantiation can be perceived, these principles must enable the maintenance of an object's history while also permitting significant changes in an object's properties. Scholl (2007) reviews a number of philosophical approaches to persistence in the face of property change. One class of theory, *perdurance* theories, postulates that objects exist across the three physical dimensions but they also extend through continuous time. As such, they contain their own history; distinct instantiations of an object in time are bound to one another through spatiotemporal continuity. A similar approach can be applied to object representation (as distinct from its application to real-world objects), and if object representations do encode an object's history, they encode their prior states also. However, encoding prior states is not enough; their spatiotemporal properties must be somehow encoded also; that is, the trajectory through time of those state changes. And even that is not sufficient to encode an event; event representations are not simply the equivalent of lists of objects and the changes they undergo through space and time – they are an ensemble of trajectories that encode not simply the spatiotemporal contingencies that obtain within a single object's history, but also those that obtain across multiple objects' histories – at a minimum, the objects participating in the event, but also including (perhaps only in the shorter term) incidental properties of the contexts within which those trajectories intersected.

To put this last 'ingredient' in concrete terms: On hearing that a particular chef chopped a particular onion, the prior state of the unchopped onion, its transition to a chopped state, and all the other concomitant object-state changes (including those of the chef, any knife she used, etc.) that accompany this transition, must be encoded relative to one another in some internal representation that maintains the temporal contingencies between each. This raises the puzzle that whereas objects persist in the real world in real-time, and events unfold across real-time, the internal temporal encoding of object-state change must somehow be divorced from real time. For now, we shall take the possibility of such encodings as a necessary given, but we return to this below, when discussing the encoding of an event's temporal dynamic. Importantly, and what makes something an "event", more than just the encoding of the participating objects and their intersecting trajectories, is their "grounding" in a spatiotemporally defined context – that is, their *incidental* intersection with other concomitant objects (or other events) in that context; the kitchen counter, the ticking of the clock on the wall, the noise of the roadworks outside the open window, and so on. The existence of these incidental intersections (giving rise to incidental *associations* – see below), no matter how transitory, poses the following question: What constitutes "taking part" in an event? Is the ticking clock a part of the event? To the extent that it may convey information (c.f. earlier discussion of Hockney), yes. To the extent that it does not, no. But it *does* take part in the *experience* of the event. We view the notion of participation as graded (entailing more, or less, overlap in representation): some things intersect in space and time (the chef's knife and the onion) and other things only in near-space and time (the clock) or only in near-space and near-time (the subsequent eating of the dish). Some things intersect causally (which, for now, we shall take to mean there is some experiential basis for assuming a contingency – e.g. the kitchen counter), and others do not (the clock, again). There may be (abstract) aspects of an event that are common across contexts, but such abstractions are what we refer to as generalized events; they are not the actual instantiation of the event in a particular space and time. The episodic instantiation of an event necessarily entails more than just those abstract components of the event that are common across all instances of that class of event. Below, in Section 3.5, we address the question of why we do not necessarily recall (or maintain) all the incidental associations that accompany our episodic experience of an event.

If we accept that object representations contain their own histories, the ensemble of such representations and their relative spatiotemporal properties (relative also to incidental properties of the context) necessarily encodes all the information that constitutes an event, in much the same way as the unfolding of such histories and their spatiotemporal contingencies in the real world "defines" an event. But this raises, again, the issue of whether something more is required to *interpret* the event in terms of object identities, participant roles, causality, and possibly, intentionality. We find it useful to operationalize interpretation in terms of the behaviors that are consequent on, and reflect, interpretation. In this regard, if such behaviors include being able to predict how our world may unfold, or may have unfolded to reach the current state, and being able to predict how this unfolding may constrain our own actions and perceptions, there is not much left for "interpretation" to do, as a theoretical construct. Participant roles then become subsumed into the spatiotemporal contingencies that constrain the unfolding of that world (similarly, for causality and intentionality, to which we return in the final section). We would claim that the puzzle is not to explain participant roles (or causality, or intentionality), but is instead (or at least, first) to explain the ingredient that is even more primitive: the encoding of object identity, and the manner in which an object's existence at successive moments in its trajectory through space and time come to be *bound* to that same identity. In the next section, we describe a conceptual model (partly described in Altmann, 2017) which supports individuated object histories and event ensembles.

3.3 Constructing object tokens on-the-fly

A principle challenge for any account of event representation (indeed, for any account of human cognition) is to explain how it is that we can create *on-the-fly* representations of newly experienced entities (whether experienced directly or indirectly via language); entities that can take on new histories of their own. The chef, in the examples above, has an identity, created through such *tokenization*, that can accumulate its own unique history, perhaps about the cut on her finger, or the unusual leather apron she wears – once an entity is tokenized, we can add to its history, or we can retrieve contextually relevant parts of that history. In our chef example, we embellish the tokenized entity with long term knowledge of chefs and their typical attributes (i.e. semantic memory for the class of entity of which this particular chef is an individual instance), but these embellishments are added to the episodic knowledge we have accumulated also (e.g. about that cut). Object histories are critical to our account of event representation. But their primacy in IOH raises several challenges: If event representations are ensembles of intersecting object histories, an account is needed to explain how such histories are *encoded*, and how fragments of history from one time are *bound* to those fragments of history from another that pertain to the same object. Equally, the account must explain how, in the representational medium, object histories *intersect*. And how the nature of an object's representation, when that representation is reactivated (such as when we refer to that object's identity), causes relevant parts of its history to be reactivated also (leading to the simultaneous activation of distinct states as is required for event understanding). And finally (at least for this discussion), the account must specify how individual episodic experience interacts with semantic knowledge, itself abstracted over multiple instances of individual experience (or learned as a fact conveyed through language or a single episodic experience). And not just semantic knowledge about classes of objects (the conceptual knowledge typical of contemporary accounts of semantic memory), but also semantic knowledge of classes of situations and events (c.f. schema and scripts). In this section, we describe a theoretical framework, based in part on insights from the neurobiology of memory, as well as on insights from computational modeling, which offers a mechanistic account of these necessary ingredients of event representation. Our starting point is not with the theoretical properties of the architecture – these will be described as they are required, but rather with the nature of the problem itself – the nature of *experience*.

The experience of something, whether of an object or an event, is integrally bound up with the context of that experience. The claim we shall develop in this section is that there is a very tight theoretical relationship between the context of experience and what it means to tokenize an entity. We experience tokens, not types, even if we might, in the moment of that experience, make generalizations based on type. But while our actual experience is of tokens, experiences are not isolated encounters with isolated objects. Instead, we encounter objects in the presence of others, and across time. Each such encounter is unique, such that our experience (and to an extent our subsequent recollection) of such encounters will include incidental properties of the encounter – the color of the tablecloth on which we ate that remarkable tiramisù, or the sudden noise outside the window, or the taxi ride to the restaurant. We may forget parts of the context, but they are uniquely defining, nonetheless, of the actual experience (unless we forget them all, in which case there can be no episodic recollection). These incidental components of the context in which we experience something are the stuff of episodic memories – the encoding of objects and events *in their spatiotemporal contexts* (Tulving, 1983). And crucially, these incidental co-occurrences must become associated with one another and with the object (or event) of our experience in order for that experience to become encoded as an episodic memory. The *mechanism* by which we associate the incidental co-occurrences within a context has been described in terms of *relational binding* (Cohen & Eichenbaum, 1993).

In Altmann (2017) we pointed out the central role that relational binding plays in statistical learning. There, the learning process requires the extraction of systematic regularities in the face of, initially, having no basis to distinguish between the systematic and non-systematic (incidental) co-occurrences that define the individual episodes of experience: We cannot learn which dependencies matter (i.e. which are informative in respect of predicting concurrent or subsequent input) unless we initially encode all potential dependencies (constrained by factors such as attention, saliency, and previously extracted/abstracted knowledge – see below). The predictive encoding exhibited in the simple recurrent networks (SRN) of Elman (1990) worked in exactly this way: Non-systematic co-occurrences in the input would result in connectivity within the network's structure which would gradually become less influential as the network gained more experience of those co-occurrences in its input that enabled the network to better predict subsequent input (i.e. those co-occurrences across time that were systematic rather than arbitrary). The SRN is relevant here because, aside from its predictive encoding, it exhibits two other crucial properties that underpin our account of tokenization. The first is this ability to, in essence, form arbitrary associations between an input and its context (c.f. relational binding). The second is its ability to do this across time. And while Elman's SRNs were restricted to the domain of language, the theoretical principles that underpin their workings are domain-independent (c.f. Altmann & Mirković, 2009). And just as time and context are critical to the SRN's ability to learn contingencies across time, so they are critical to the process of tokenization.

So what does it mean to define something as an *instance* (a token) of a particular type? Indeed, what does it mean to *experience* such an instance? The following is one possibility. We start with the case where we experience an object (or event) directly, and for the sake of the discussion, we shall focus on visual experience, but the logic does not require it to be visual. We shall use the same cartoon stills from Figure 1 to help visualize the mechanism of tokenization we are proposing here.

When we observe Garfield in the Garfield & Friends cartoon, his corresponding perceptual features are bound both to each other and to the representations in semantic memory that are activated by those features. Critically, they are bound also to other incidental features of the context (again, limited by factors such as salience and attention). Similarly for the pie he is holding, and for the other elements in the scene (we use "elements" to refer to any hierarchically organized set of features, passed up by the visual system, that in principle could be labelled or segmented into a meaningful group – e.g. the features corresponding to the pie, or some part of the background – where a "meaningful" grouping of features is defined as one that affords constraints on what other features may co-occur either in the present or through time; the account is agnostic with respect to the transformation of visual input into meaningful units). Thus, there is indiscriminate association of perceptual features with other elements in those

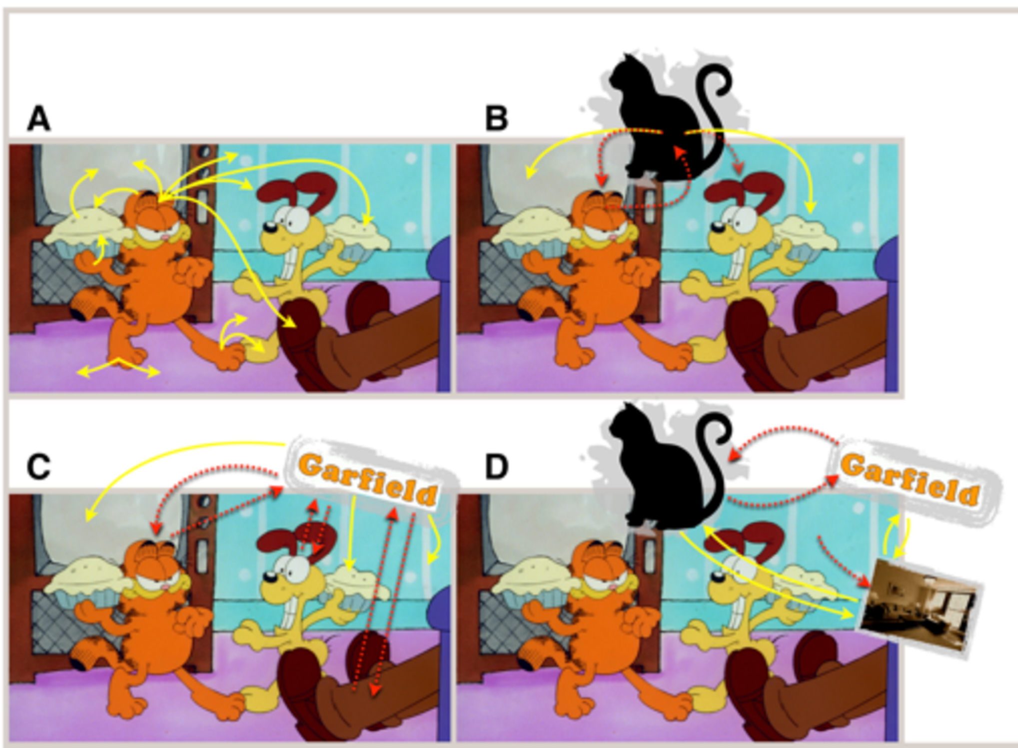

Figure 2. Relational binding during event perception: Each panel shows just a fraction of the associations that form during the same momentary snapshot of the unfolding event. Solid yellow arrows indicate indiscriminate associations; dotted red arrows indicate systematic associations (i.e. associations with prior knowledge). The panels are separated out for expository purposes only – the associations shown across all four panels apply at the same single moment in time. **A**: Elements within the scene are associated indiscriminately with one another. Some of these associations are arbitrary (Garfield's foot associated with the floor) and others systematic (Garfield associated with Odie). Some are systematic but uninformative: That Garfield is standing on a surface rather than floating is systematic, but his standing on a surface is always the case and hence *uninformative*, unless the surface is itself less predictable. The mechanism of relational binding is "blind" to which associations are systematic or arbitrary. **B**: Elements within the scene activate semantic knowledge (e.g. seeing Garfield activates semantic knowledge of cats) and these in turn reinforce the elements that activate them as well as other elements with which they are associated in semantic memory (dogs are associated with cats, and hence the semantic knowledge of cats activated by Garfield activates, in turn, Odie). There is also indiscriminate relational binding of the activated knowledge to the elements in the context (e.g. association of the activated representation of cats to the television, pie, etc.) **C**: Elements in the scene also activate schema (semantic knowledge of situations and the typical events they entail and participants that take part). Hence the activation of knowledge of the protagonists and kinds of relationship typical in Garfield cartoons. This schema knowledge is also relationally bound to other elements in the context. **D**: The semantic representations activated by the scene are relationally bound to one another also, with some associations being systematic (e.g. the association between cats and the Garfield schema) and others arbitrary (e.g. the association here between the Garfield schema and living rooms). Original Still Image © Paws Inc. Reproduced with permission. Living room: GFDL (http://www.gnu.org/copyleft/fdl.html)], from Wikimedia Commons.

features' sensorimotoric contexts (Figure 2a), as well as systematic association between those features and the representations in semantic memory that they activate (Figures 2b and 2c). The conjunction of these representations grounds Garfield (i.e. this token cat) in a specific context at a particular time. But just as Garfield is grounded in the concurrent context, by virtue of those associations to elements of that context, so are the representations activated from semantic memory by Garfield's perceptual features (as are other representations in semantic memory that become activated by other elements of the scene; Figure 2d). Thus, we construe relational binding broadly, to also include the binding of activated semantic representations to the episodic contexts that lead to their activation, as well as to other (perhaps arbitrary) semantic representations also activated within those contexts. Equally, we construe relational binding to be a purely bottom-up process that is not mediated by semantic (long term) knowledge – the indiscriminate association of features to one another and with the context is not mediated by such knowledge (Colzato, Raffone, & Hommel, 2006), although the activation of those lower-level features is reinforced by the semantic knowledge they activate (c.f. Hommel & Colzato, 2009).

Relational binding has been studied extensively within the context of the neurobiology of memory: The binding of perceptual features to one another, and to semantic memory, is thought to result from interactions between hippocampal and neocortical brain regions, with hippocampal regions primarily responsible for relational binding, and neocortical regions for encoding of (semantic) knowledge abstracted across accumulated experience (for review see e.g. Konkel & Cohen, 2009; Moscovitch et al., 2016; and Preston & Eichenbaum, 2013). We return below to the theoretical relevance of what is known about this neural circuitry.

While this grounding of tokens in the current context provides a unique representational signature for each such token, there is one critical ingredient that is still missing to explain tokenization. The distinct sets of associations, both arbitrary and systematic, must also be grounded in time. We posit that this grounding in time is a consequence of recurrence through time. In the SRN or other recurrent architectures, the input at one moment in time feeds into hidden layers within the network, changing their activation profile. But a copy of their activation profile from the previous time-step is also fed into the hidden layer, meaning that the current activation profile is a product not just of the input at that one moment in time but also of the inputs (and their impact on the hidden layer) at successively prior moments in time. Essentially, the idea here is that the input, and its concurrent (episodic) context, become combined, or associated, not simply at each single moment, but also across time with successively prior inputs and prior episodic contexts; each "episodic snapshot" is thus accompanied by echoes of the past, as afforded by recurrence through time – relational binding does not just occur within a timeframe (cf. a single frame of the Garfield cartoon) but occurs *across* timeframes also (see Figure 3) – it is itself a dynamical process. Recurrence is not simply a defining feature of computational models such as the SRN (Elman, 1990) but also of the brain structures implicated in episodic encoding and recall (Kumaran & McClelland, 2012; Schapiro, Turk-Browne, Botvinick, & Norman, 2017). Its role is critical here because it ensures that object tokens are associated with their past selves through relational binding across time. That is, object tokens essentially carry their history with them and are in fact unique *trajectories* through representational space and time.

Our discussion thus far of tokenization has been based on the case where we can directly perceive, or have directly perceived, the tokens in question. But what of language? Does the account we have developed here apply to the chef we periodically refer to and the properties she has taken on (her weighing, chopping, and smelling of the onion; the cut on her finger; her unusual leather apron)? In fact, the account holds in just the same way: Our Garfield example describes how visual features are relationally bound, through time, to one another and to their context. But if instead of *seeing* Garfield, we see the word GARFIELD or hear the word "Garfield", those are still perceptual features that, like seeing Garfield, activate semantic memory, or can be relationally bound to the context in which they are experienced. The perceptual features associated with the phrase "the chef" activate semantic knowledge corresponding to the meaning of the phrase; and both they and this activated knowledge associate also with the incidental features of the accompanying linguistic and non-linguistic context—the location, time and other incidental features co-occurring with the experience of that phrase (whether spoken, written, or signed). When we first introduced the chef, she took on a unique episodic signature that allowed us, subsequently, to refer back to "her", and to use linguistic expressions (such as "our chef") which refer to that unique token. There is thus little difference in the *process* of tokenization, except in perceptual detail, between directly experiencing an object and indirectly experiencing it through language. That is, seeing the chef will not only instantiate perceptual details about her that are lacking when hearing "the chef", but will also instantiate perceptual details of her physical surroundings (a kitchen, or some other location) that are also lacking when hearing "the chef". On the other hand, hearing "the chef" instantiates perceptual (acoustic) details of the uttered phrase, and does so in the context of the physical surroundings of the hearer. These perceptual details offer a different grounding of the chef when hearing "the chef" than when seeing the chef (i.e. different representational content), but in each case the grounding provides a unique episodic signature through the same mechanisms of relational (and semantic) binding through time. This same mechanism, that binds an auditory word or phrase to its context (both linguistic and extra-linguistic) enables tokenization of other kinds of auditory or transient events (beeps or flashes, or less transient events such as thunderclaps and so on that are not changes in external *object* states but which constitute changes to the experiencer's perceptual state). That is, they become episodically grounded in their context, and in our experience, through relational binding. Surprisingly, few neuroscientific studies of language have explored the role in language processing of those same (hippocampal) brain regions that are implicated in relational binding and episodic encoding (but see e.g. Duff & Brown-Schmidt, 2012, for a notable exception).

3.4 Keeping track of, and keeping apart, object trajectories

One further challenge: how are object tokens "resolved" against the representational backdrop in which they are embedded through the relational binding and recurrent processes we have just described? How can the different representations, whether of an object, its context, or their pasts, be kept apart and remain accessible as individuated representations? At first glance it might appear that the process of laying down successive representations across the representational substrate, each one superimposed on each other, is a little like painting on water – the paint will wash into all the other images that were previously painted into the water, with no possibility of resolving one from the other. This is, however, where recurrent networks and equivalent dynamical

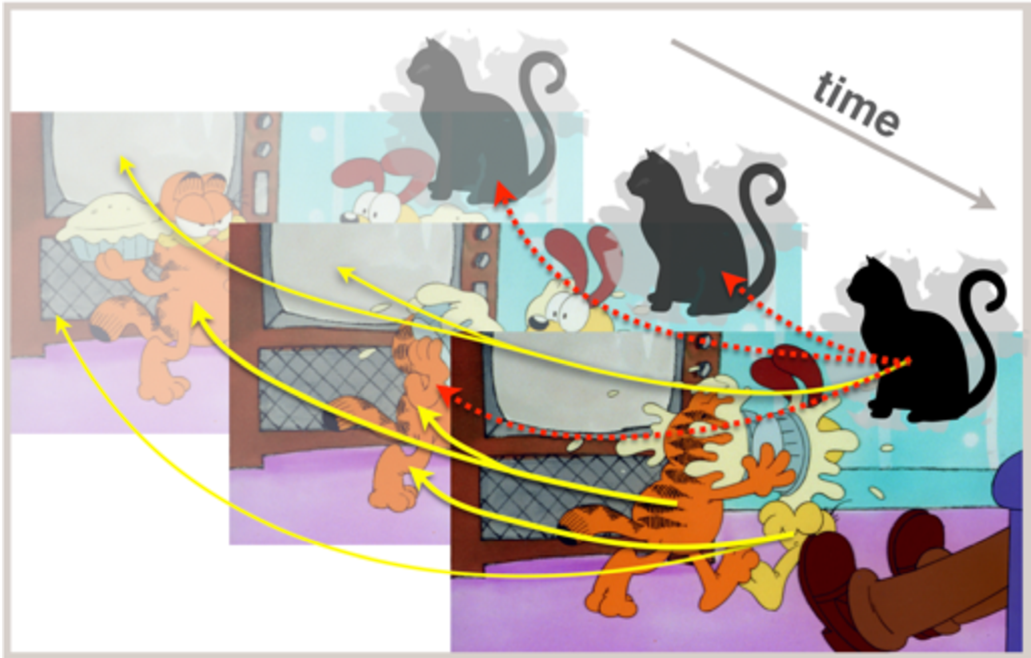

Figure 3. Relational binding through time. Elements at each moment in time, including semantic knowledge activated by elements in the scene, are associated through relational binding not just with other elements at that same moment, but with other elements, and with themselves, at previous moments in time. Some of these associations are arbitrary, and due to incidental co-occurrences (solid yellow), and others are systematic and due to prior knowledge (dotted red). The association between Garfield at one time and himself at another time can be considered systematic, but is formed through relational binding that is "blind" to this kind of spatiotemporal systematicity; for present purposes we define as relational any association that arises through co-occurrence (in or across time), regardless of whether there is any systematicity to that association. Only a fraction of the possible associations are shown. Original Still Images © Paws Inc. Reproduced with permission.

architectures excel – each "image" has a unique computationally instantiable signature, reflecting the unique properties of the concurrent context, as well as the unique properties of higher-level representations corresponding to those we referred to earlier as the semantic knowledge activated by the current sensorimotoric input. An intuition for how such representations can be individuated, given the recurrent and relational processes described above, can be conveyed as follows: For objects that we have experienced before, or which are instances of types we have experienced before, the semantic memory activated by the object's sensorimotoric features reinforces just those features in the perceptual experience that map onto that memory – the semantic type corresponding to cats, activated through experiencing Garfield, becomes a part of the experience of Garfield, distinguishing that experience from the experience of Odie, or of the pie, or of Jon Arbuckle's legs. These higher-level representations that are co-activated with the sensorimotoric features from the episodic experience are to those features as the puppeteer's strings are to the marionette – they individuate the body of one marionette from the body of another (the relationship between the position of the marionettes and the configuration of those strings changes dynamically through time, in much the same way as conceptual activation, and the relationship between semantic memory and the sensorimotoric input activating that input, is itself dynamic; see Yee, 2017 for review). For objects we have not experienced before and which we encounter for the first time (i.e. for which we have no higher-level semantic or other memory that is co-activated with the perceptual experience), features that travel together through space and time are self-reinforcing – for example, Garfield's paws (if we had never seen an animal before) travel with his arms across our experience of Garfield more than they do with pies. Thus, spatiotemporal continuity allows objects to become individuated from other incidental properties of their surroundings which may come and go in the face of that object's continuity.

How, or even, whether, representations of individual tokens and their histories are actually individuated one from the other is less important than the consequence for behavior in respect of such histories enabling the system to anticipate likely outcomes given particular (trajectories of) input. For example, Elman (1993) demonstrated how an SRN could distinguish the different instantiations of the same lexical item at different positions in a sentence, with the specific trajectory leading up to each of those instantiations appropriately constraining the network's predictions of what may come next – an example of tokenization of the kind described above: lexical items were grounded in the dynamics of both their contexts and the emergent representations activated by that input (corresponding to the network's equivalent of semantic memory), with identical lexical items distinguished through the network's encoding of what could be called each item's episodic context. In respect of visual, rather than linguistic, objects, when the chef sees two identical spoons with which to taste her sauce, each is individuated from the other by their relationship to their respective local contexts.

3.5 Event representation, tokenization, and relational binding through time

Relational binding through time is critical to the notion of episodic grounding and tokenization. Individual object tokens exist as episodically bound representations within a larger representational space – within the context of an individual frame from the Garfield cartoon, the individual objects there are not represented independently of the other objects or associated semantic knowledge with which they co-occur. And across successive frames, they do not exist independently of the other objects, and changes to those objects, that occur through time. Thus, it is the ensemble of overlapping representations (overlapping in space and time) that dynamically reflects the intersecting object histories that, if those representations change through time, define an event. Within contemporary theories of semantic memory, concepts that overlap in their representational content also overlap across their physical embodiment in the brain (cf. Allport, 1985). By the same logic, the trajectories described above that share spatiotemporal context will overlap both representationally (elements of their representation will be the same, reflecting those same elements of the co-occurring contexts) and physically, across the substrate supporting those representations. Analyses of the internal activation profile of recurrent networks (e.g. Elman, 1990, 1993) reveal similar overlap across the (artificial) neural substrate. One consequence of this overlap is that activation of one representation will (re)activate overlapping representations, as observed in studies of semantic priming (e.g. Meyer & Schvaneveldt, 1971), but also as observed when recall of an item cues reactivation of its episodic context (see e.g. Mack & Preston, 2016, and Zeithamova, Dominick, & Preston, 2012, for neuroimaging evidence on the reinstatement of episodic content).

As outlined earlier, relational binding through time not only binds objects to their dynamically changing episodic contexts, it also binds objects to their past selves. Thus, activating the representation of an object at one moment in time will, through that associative binding, activate prior (or successive) representations also. This activation across time of different parts of an object's trajectory will be modulated by the episodic contexts which, common to those different representations along that trajectory, will reinforce the activation of those different representations that share (aspects of) those contexts. Returning to our chef example, activating the representation of the onion as she fries it will re-activate the representation of that same onion as she first chopped it and, even before that, the representation of that same onion in its pre-chopped state (cf. Hindy et al. 2012; Solomon et al., 2015; see Joanisse & Seidenberg, 2003, for an example of a recurrent network able to correctly retrieve the antecedent presented earlier in a sequence to a subsequent pronoun – loosely equivalent in some respects to the activation of an earlier episodic form by a later episodic instantiation). Activation of an earlier part of an object's trajectory will in part depend on how useful it was to maintain that part of the trajectory: The more useful a memory across multiple experiences, the more likely it will be maintained; the same is presumably true of the earlier parts of an object's trajectory. For a sentence such as "The chef will peel the onion, chop it, smell it, pour it in a pan, put the pan over a low heat, and caramelize the onion" – it remains an empirical issue whether, at the final "onion", its unpeeled state is a part of the trajectory we remain sensitive to, and whether there are individual differences in such sensitivities.

This binding of objects to their past selves ensures continuity in object representation across time. When we see the chef chop and then fry the onion, spatiotemporal continuity across the different states of the onion, combined with relational binding across time, ensures that the initial episodic experience of the intact onion is bound through spatiotemporal continuity to the episodic experience of the chopped onion and, subsequently, to that of the onion being fried. But while spatiotemporal continuity is sufficient to explain object persistence across change, it is not required to explain it: If the transition from the onion being intact to it being chopped is occluded, the onion in its chopped state will activate semantic knowledge of onions in general, which will re-activate the episodic memory of the previously seen intact onion (its recency gives it pre-potency in respect of its activation state). This latter representation will, by virtue of its co-activation with the currently seen chopped onion, become bound through time with the chopped onion. This form of semantically mediated associative/relational binding is sufficient to support the experience of object persistence across changes so long as the distinct states of the object are each recognizable as belonging to the same semantic type (c.f. Carey & Xu, 2001). This said, there are cases where the distinct states may belong to different semantic types (which, depending on experience, may be related hierarchically or may be unrelated), as in the case of the crushed car discussed above or a butterfly and its prior state as a chrysalis and, before that, a caterpillar. Here, the semantic mediation that binds one state to a re-activated prior state is more complex, requiring associations between semantic types (e.g. seeing the butterfly activates semantic knowledge of butterflies which activates semantic knowledge of chrysalides which in turn re-activates the episodic memory of the chrysalis). In the absence of strong type-to-type associations, such mediation is less straightforward, which is why for some, the crushed car is not so recognizable as a car, whereas for others, it is.

Importantly, semantic mediation in the absence of spatiotemporal continuity is just one of the mediators of object persistence – while there may be discontinuities in the direct perception of an object, there may be continuities in the episodic context which, independently of any semantic mediation, also support that object's persistence: When the car goes through the crusher, it disappears from view, but there is spatiotemporal continuity in respect of the crusher itself and other elements of the context with which the two states of the car – the before and the after – are associated. The crushed cube of a car is associated in current time with the crusher which is itself associated across time with the uncrushed car, and this creates an association (through re-activation and relational binding) between the two versions of the car. This is the process we referred to earlier as modulation by the episodic context of the activation of different parts of an object's trajectory. Such contextual mediation, coupled with the semantic mediation described earlier, is what permits the illusion of continuity in language: If we read "the onion that was chopped and then fried…" there is a very strong tendency to assume that the chopping and the frying was done by (a) the same person and (b) the same chef as had been introduced earlier. Altmann (1999) demonstrated this tendency in the context of anticipatory processing at verbs, prior to the postverbal referring expression (c.f. "The chef chopped an onion, then she *fried*…"). Altmann & Mirković (2009) account for this preference to anticipate alreadyintroduced entities rather than to anticipate as yet unintroduced entities using the same mechanism of semantic mediation outlined here. Language is a paradigmatic case in which there is no spatiotemporal continuity in the perceptual input of the objects being referred to. Instead, there is a form of "representational continuity" afforded by the association of those objects with, and reactivation by, the higher-level semantic structures (pertaining both to the objects and their contexts) that form and/or are activated during the comprehension of the language.

A final puzzle regarding relational binding through time: If objects within a scene, or indeed, the perceptual features that constitute those objects, do not exist independently of others in the scene, but are relationally bound indiscriminately to those others, why do we not recall all the irrelevant minutiae of the episodic contexts in which events occur? One reason might be that the truly incidental ones are not reinforced through mutual activation of systematically related experiential (semantic) knowledge – as associations go, they are "weaker" and therefore more prone to interference and forgetting than are the associations that are reinforced (i.e. strengthened) by semantically mediated knowledge. Another reason (they are not mutually exclusive) is suggested by the observed interplay between different brain regions supporting episodic and semantic memory, and specifically, the interplay between hippocampal regions (more broadly, medial temporal lobe) and prefrontal cortical areas: van Kesteren and colleagues (van Kesteren, Ruiter, Fernández, & Henson, 2012; van Kesteren, Beul, Takashima, Henson, Ruiter, & Fernández, 2013) describe a complementary relationship between hippocampus and medial prefrontal cortex in which hippocampus is responsible for fast relational binding, and medial prefrontal cortex (mPFC) is responsible for integration with schema-based knowledge (in the next section we address how schema are learned within the IOH framework). Crucially, they observe that when input is highly congruent with existing schema, mPFC is more highly activated than it is when the input is incongruent. Conversely, hippocampus is more highly activated when the input is incongruent with any existing schema, and less activated when it is congruent with existing schema. They propose that the greater the integration with existing schema, the greater the inhibition between mPFC and hippocampus. Functionally, this has the effect of inhibiting incidental (i.e. non-systematic) associations and promoting associations between elements that can be better integrated with schema-based semantic knowledge. Thus, and returning to Figure 2, the relationships between Garfield and Odie, and between them and the instruments (the pies) they use against one another, are more relevant than between, for example, the living room wall and other elements within the scene – here, "relevance" is conditional on whatever schema are activated. To use a different example: The schema we have for restaurants makes the floor of a restaurant relatively irrelevant whereas the schema we have for ice rinks makes the floor of the ice rink more relevant. This is not inviolable, of course – relational binding is mediated by attention (e.g. Craik, Govoni, Naveh-Benjamin, & Anderson, 1996) and salience (Fine & Minnery, 2009), and thus the inhibitory effects of schema-based knowledge on incidental relations in an episodic context are just that; inhibitory, not nullifying.

Our account assumes the relational binding mechanisms that are typically associated with hippocampal function. And yet, hippocampus appears organized more in terms of spatial and temporal codes than object codes (higher-level visual cortex, such as the lateral occipital complex, is more usually associated with the encoding of object information; e.g. Grill-Spector, Kourtzi, & Kanwisher, 2001; Grill-Spector & Malach, 2004). But while much of our discussion above is couched in terms of objects, our account is not about object codes *per se* (however these might be construed); rather it is about changes in spatiotemporal configuration and the formation of relational associations to arbitrary contexts. These are defining representational functions of the hippocampus (e.g. Moscovitch et al., 2016). Thus, we couch our account in terms of featural elements within a scene, assuming that scenes are interpreted at many levels simultaneously, with low-level visual features interpreted (via experientially-derived cortically-represented knowledge) as higher-level cognitive constructs (e.g. objects, people, etc), and relational binding occurring at all these levels simultaneously. The hippocampus appears organized to support such multiple levels of representation, taking its inputs from two distinct processing streams – the perirhinal cortex and lateral entorhinal area for the encoding of objects, and the parahippocampal cortex and medial entorhinal area for the encoding of their contexts (e.g. Preston & Eichenbaum, 2013).

3.6 Events, schemas, and episodic (re)construction

The process of tokenization described above applies as much to individual events as it does to object tokens. Events, as intersecting object histories, are as grounded in their context of occurrence as are individual tokens. While object tokens and events are not the same (one can, for example, take physical hold of a token, unlike an event), there are important commonalities. For example, both afford the opportunity for emergent abstraction, from individual exemplars of objects and events, to types of objects and types of events (i.e. generalized events). The mechanisms of abstraction and emergence that give rise to these higher-level abstractions are assumed to be the same in the IOH as the emergentist principles identified by Elman and colleagues (Elman, 1990; Elman, Bates, Johnson, Karmiloff-Smith, Parisi, & Plunkett, 1996; see also Altmann, 2017). These principles explain how the accumulated experience of individual tokens, e.g. of individual dogs and cats, leads to the emergence of semantic memory for types, such as the class of dogs vs. the class of cats; this semantic knowledge of types includes knowledge of the sensory features their instantiations (i.e. specific tokens) likely comprise (see e.g. Rogers & McClelland, 2004). The same principles explain how accumulated experience of individual events can lead to the emergence of semantic memory for types of event, such as hitting vs. eating, and the participants they likely involve (c.f. Ferretti et al., 2001). In terms of the recurrent architecture we have described above, events as trajectories can be abstracted across to form, in effect, classes of trajectory as well as classes of participant and participant role (cf. Elman, 1993). This latter case explains how we can understand the roles played by novel objects interacting in novel contexts – so long as the spatiotemporal properties of that interaction activate abstract higher-level (i.e. "emerged") representations corresponding to abstract participant roles (abstract in the sense that they may map onto, for example, agent or patient, as may happen if a change of state in one precedes a spatiotemporally contiguous change of state in the other) we may understand, albeit crudely, the roles each novel object plays in the interaction.

Just as events are comprised of spatiotemporally contingent object trajectories, so spatiotemporal contingencies between events (and event types) can lead, in principle at least, to the emergence of higher-level contingencies across events that constitute schemas (collections of events typical for a given situation or context) or scripts (sequences of events typical for a given situation or context). The representational overlap between object tokens, object types, events, event types, and schema, ensures bidirectional activation of one given the other (as determined by the probabilistic contingencies abstracted across an individual's experience – "bidirectional" need not mean "symmetrical" nor, across individuals, "identical"). Thus, an individual waiter will activate generalized event knowledge about likely interactions with that waiter, as well as schema knowledge about likely chains of events in a given (e.g. restaurant) context. Similarly, a given context will activate schema knowledge about likely events and likely participants associated with that context.

This last point, that a given context, or contextually-relevant cue, will activate knowledge about events associated with that context or cue explains a further commonality between object tokens and events: both can be constructed on-the-fly, in the absence of episodic experience, as might happen when learning through language about some object or some event (as we did with the chef and her chopping of the onion), or as might happen during "scene construction" (Hassabis & Maguire, 2007) when imagining a future or fictitious experience (imagining that onion, perhaps, or that chef peeling that onion). Earlier (Section 3.3) we explained how there is little difference (within limits) between directly experiencing an object and indirectly experiencing it through language. The process of tokenization ensures that we can construct the appropriate representation and ground it in the context in which we experience either the language or the object itself (depending on how we experience the object). But the same is also true of an event – there is little difference between directly experiencing it and learning of it through language; yes, there are differences in detail (and goals), but by "little difference" we mean in respect of the mechanism by which the tokenized representations come about. When recalling an event we recall a trajectory of intersecting object histories (as instantiated via relational binding through time – see below for an alternative conceptualization of such trajectories) – an appropriate cue will reactivate those histories, and their intersection, with which it is associated. When imagining an event, we tokenize the entities based on our semantic memory for the class of each entity and we either tokenize established event types or schemas (as would happen when imagining the event corresponding to "the server came to our table"), or we tokenize fragments of schema, event types, or event tokens when the imagined event is not from a single class ("the server swam to our table with a mask and snorkel"). In this latter case, we activate associated fragments of memory and/or schemas (in effect, activating partial trajectories). IOH is thus compatible with accounts in which episodic memory is seen as a constructive process operating over fragments rather than as "picture-perfect" recall operating over veridical memory (see Schacter & Addis, 2007, for review).

4. Time matters

We have defined a basic representational principle of event representation: events are ensembles of intersecting object histories in which one or more objects undergo a change in state or location. And schemas are abstractions across ensembles of intersecting events; they are knowledge about the spatiotemporal dynamics of intersecting (abstract – i.e. probabilistically likely) instances of event types. But glaciers change location (and the landscape), and yet when we observe one, we would find it hard to describe what we are observing as "an event". And a child hopping ends up in the same state that she started (ignoring for the moment her breathlessness). How do we reconcile these two apparently different kinds of observation with IOH? In fact, hopping and glacial movement are not so different, in that the issue here concerns the temporal resolution with which one views these distinct kinds of event. In the hopping case, if one were to merely "sample the world" precisely at those points when the child is on the ground prior to lifting off, or after landing, there would be no apparent change in state. But if instead one sampled more frequently, it would become apparent that there would be changes in state/location (body posture, height from the ground, etc) *during* the hopping. Similarly with the glacier moving – in these cases, we have to sample more slowly (or evaluate change across a longer timeframe). At issue here is the rate at which we sample the external world. Is it fixed? Or variable, depending on attentional state and/or the exigencies of whatever task we are currently engaged in?

4.1 Sampling the external world across time

Perception is not continuous, but discretized into samples whose frequency depends on attention and task (e.g. Dehaene, 1993; vanRullen, Reddy, & Koch, 2005), as well as on expectations regarding the rate at which information accrues (Akyürek, Toffanin, & Hommel, 2008). In Altmann & Mirković (2009) we pointed out how even with a single sample rate, a dynamical system with recurrence (e.g. an SRN) could develop sensitivities to contingencies across variable-width temporal windows. An SRN can learn by having to predict successive inputs; the manner in which it reduces the error between its predictions and the actual successive inputs causes it to learn just those contextual dependencies (through time) that decrease prediction errors. The internal representation of prior states of the system, as afforded by recurrence, enables the predictive process to operate across multiple time frames and multiple levels of "representational abstraction" (essentially, reflecting a hierarchical representation of contingencies, as occurs in language between phrases, words, and phonemes, for example, or in the realm of category structure between animals, cats, tabbys, and Garfield). Thus a single sample rate can in principle support prediction at varying sized units of temporal incrementation.

Reynolds et al. (2007) implemented Event Segmentation Theory (EST, Zacks et al., 2007) in a recurrent architecture in which, as we assume here, the perceptual system is continuously making predictions about the upcoming perceptual input. The network was better able to predict successive input within events than across events, and the increased prediction error between events – that is, at event boundaries – allowed the network to segment events at the appropriate boundaries. Thus, while in Section 3.3 we have described the dynamic in IOH as if it were continuous across time, neither the apprehension of sensory information nor perception are continuous; rather, the input stream is perceived as discontinuous, as a function of the waxing and waning of prediction error. While there are similarities between EST and IOH in respect of continuous prediction, prediction error, and discontinuous perception, one crucial difference concerns the status, or consequence, of these perceptual discontinuities. In EST, increases in prediction error trigger the updating of the current event model. In the extreme, if the prediction were perfect, there would be nothing to update. But this raises the question: What if the prediction is only slightly imperfect – how big an increase is required before the current event model is updated? Would it not be useful to update the current event model as the event is still unfolding so as to better anticipate what may come next? In EST, event segmentation occurs simultaneously at multiple timescales, resulting in hierarchically structured event representations (Zacks et al., 2007). At issue, is how deep is the hierarchy and whether there is ever the case where prediction error, no matter how small, does not result in updating of a part of the event representation at some perhaps deeper (more finely segmented) level in the hierarchy (c.f. studies showing integration of information within perceptual events lasting a few hundred milliseconds; e.g. Akyürek et al., 2008). In many respects, the IOH is complementary to EST in respect of its goals (to explain *content* rather than segmentation *per se*). But in regard to segmentation, there is a subtle divergence: In the IOH, an "event model" corresponds, roughly, to a set of intersecting trajectories. But updating is continuous. If each prediction were perfect, the anticipated trajectory would not change at each successive input. But if there were any error to that prediction, that error would drive updating of the internal states that in turn drive the prediction – that is, they would modulate the trajectory, and hence, in EST terms, the event model. Thus, updating is continuous in the IOH, or as continuous as the system's sampling permits. We return to the relationship between the IOH and EST in the final section below.

In respect of the actual sampling rate of the perceptual system, it is unlikely that the encoding of episodic experience is based on a single sample rate. The hippocampus exhibits a "gradient of abstraction", with more perceptually grounded, spatiotemporally fine-grained, representational properties in posterior parts and less grounded, spatiotemporally coarser-grained, and more abstract representation in more anterior parts (for review, see Poppenk, Evensmoen, Moscovitch, & Nadel, 2013; for interpretation of this gradient in terms of spatiotemporal scale, see Collin, Milivojevic, & Doeller, 2015; and also Long, Bunce, & Chrobak, 2015). Indeed, Collin et al. (2015) showed that multi-event sequences (visual narratives) are represented simultaneously at multiple scales along this hippocampal gradient, from smaller-scale properties in posterior hippocampus to larger-scale properties (the complete narrative, including all the associations between the sub-events) in anterior hippocampus. This spatiotemporal scaling is related to the hierarchical principle described by Altmann & Mirković (2009) – hierarchical representations of contingencies through time necessarily confound abstraction and temporal scale, with more abstract "higher-level" contingencies (i.e. dependencies among dependencies) necessarily occurring across longer timescales than the lower-level, less abstract, contingencies on which they are based.

4.2 Encoding trajectories *through* time

That we sample the world is just one side of the equation – it is the mechanism by which we apprehend the continuously changing state of the external world and internalize it. But while sampling the external input is something we do in real time, our recall of an event does not, like the real world event itself, unfold in corresponding real time. And yet we do have knowledge about the time that typical events take and about the time that specific (even atypical) events took. This knowledge appears, even, to mediate the time it takes to process descriptions of events of different durations (e.g. Coll-Florit & Gennari, 2011; Zwaan, 1996) or to recall from memory directly perceived events (e.g. Faber & Gennari, 2015). While it may seem that IOH does not have anything to say about the encoding of time, as it applies to knowledge about event-time, the properties of dynamical systems do constrain how we might conceptualize its encoding. We shall focus here on just one such constraint, which results in how the encoding and recall of individual events (or classes of events) can occur over a timeframe that is compressed relative to that at which the event was apprehended.

As observed by Elman (1990), the internal state of an SRN at any one moment in time constrains what states it can move into at the next moment in time, and at the next moment after that (generally, the likelihood of it *actually* moving into a particular state diminishes the further away in time it is from the current state). Thus, a snapshot of the internal state of the SRN taken at one time will include within it these constraints on how its future states may unfold. Crucially, it also includes within it the trajectory that led to the current state of the system (the present state constrains both how future states will unfold and how past states must have unfolded). Consequently, re-activating a single state of the system (or a part of a state) will re-activate its history – the trajectory of past states that led to that one. Much like the individual pieces of a broken hologram each contain a whole image, so an individual snapshot in time of a dynamical system can itself contain an entire "image" of the trajectory of states that led to that state, and the trajectory of states that can follow it. Of course there are limits on the temporal resolution afforded by any given recurrent system: The more distant those echoes of past states are in time, the harder they are to resolve, and the harder it will be to recover the temporal dynamics associated with those past states (hence, the temporal dynamics of more distant episodic experiences will be less accessible). This trajectory-within-a-snapshot approach to temporal encoding has the consequence that event-time can be divorced from real time – recall of an event activates, during the moments of recall, the entire temporal dynamic of the event (subject to those limits on temporal resolution for distant events, as well as for more recent, but very short, events; Akyürek et al., 2008), allowing the observer or recaller to retrieve that temporal dynamic in a timeframe that is divorced from the actual time it took for the event itself to play out. This is quite different from the case of a single frame of a cartoon (c.f. Figure 1) which does not encode the frames that led to it, let alone their temporal relationship to the current frame. Crucially, and returning to the prior discussion of recalling events (in Section 3.6), this means that recalling an event does not entail reconstructing an entire sequence of distinct states of the memory system (equivalent to recalling the three successive time points in Figure 3), but can simply entail the recall (the partial re-instatement) of a single state which itself encodes that trajectory of states. This is not to say that such recall is instantaneous – the process of recall is itself a dynamical process as activity spreads through the memory system.

One way to conceptualize the encoding of entire histories within a single state is to recall the multiflash photography of Harold Edgerton (1903–1990), in which a tennis player's serve would appear as superimposed images of the player's arm and racket at different positions during the serve – each image overlaid on the previous one as the stroboscope flashed at successive moments in time. This is (very loosely) a visual equivalent to the recurrent network overlaying its current input over prior inputs (see above, Section 3.3). The entire sequence of movements is captured in a single frame. To recall the sequence requires only recalling that single frame. If, instead, each stroboscopic flash were caught on successive frames of a film, recalling the entire sequence would require recalling multiple frames of the film. However, and returning to our trajectory-within-a-snapshot encoding, this discussion raises the question: Which state of the system do we retrieve (and partially re-instantiate) when recalling an event or episode? We assume that as with any cued-retrieval, whatever cue causes retrieval will cause retrieval of just those states (potentially more than one) with which the cue is most strongly associated. Consider having watched a friend cook dinner – the end of that event (as signaled by a discontinuity in the probability profile regarding what may come next; Zacks et al., 2007, and see above) contains, in principle, the entire trajectory of states of the person/objects that took part in the cooking. But if asked to recall what your friend was doing, during the cooking, when the delivery man called (supposing that that had indeed happened), the delivery man calling would be a cue that would most strongly activate the encoding of whatever states of the world had been experienced concurrently with, and episodically encoded during, that period.

Our emphasis in this discussion of a single state of the system encoding a trajectory through time – from an object's prior history (c.f. causal history) to its potential future states – has focused on *retrieval* of that single state re-activating that encoded trajectory. But recent evidence suggests that the equivalent occurs not just during retrieval but during the real-time experience of an object. Chen and Scholl (2016) showed participants two 2D geometric objects (e.g. two black squares), with the second suddenly replacing the first. This second object looked as if some other object (e.g. a circle or star-shape) had been pushed into it – part of the square was "missing" due to the intruded shape. On seeing this transition observers perceive something akin to apparent motion – as if the actual intrusion motion is projected onto the transition between perceiving the first square and perceiving the second (see Spröte, Schmidt, & Fleming, 2016, for demonstration of a similar visual inference using a different paradigm). Thus, even seeing an object (the second square with its intruded contour) can activate, in the right circumstances (i.e. when the context constrains the space of possible trajectories), a trajectory through time.

4.3 Estimating time

Divorcing how we represent the temporal dynamic of an event, and its recall, from the actual time that an event takes to play out leads to the following challenge: If that temporal dynamic is encoded in the reactivation of a single state of the (dynamical) memory system, how is it that we are able to generate explicit estimates of the time-course of that dynamic – that is, of the *duration* of an event or a part of an event? Here, we assume the approach taken by Gennari and colleagues (e.g. Coll-Florit & Gennari, 2011; Faber & Gennari, 2015). They make two broad claims: First, that knowledge about the time it typically takes for an event to unfold is encyclopedic, or factual – a part of the semantic knowledge about generalized events (e.g. that caramelizing onions takes around 30 minutes, but getting tenure takes about 7 years). This is not knowledge that is learned through experiencing the passage of clock time as events unfold (although on occasion it is, but that would still be factual knowledge derived by subtracting one time from the other). The second claim is that estimates of the duration of an event are modulated by the relationship between that event and other associated events or sub-events; the more associated events there are, and the more dissimilar they are to one another (i.e. the greater their diversity), the greater the estimated duration of the event. This is as true for typical events indexed through language (with longer reading times for events having more diverse associations, Coll-Florit & Gennari, 2011) as it is for novel events which are directly experienced for the first time (Faber & Gennari, 2015).

That an event is associated with its sub-events and with other related events is part-and-parcel of the IOH; activating an event representation (i.e. an ensemble of intersecting object histories) necessarily activates related events and sub-events (see above, Section 3.5), and each of these has associated with it semantic knowledge about its typical duration. The nature and diversity of these associated events, coupled with knowledge of their durations, and perhaps even the time it takes to reactivate them (with longer times for more, and more diverse, associations), somehow map onto a numerical estimate of an event's typical duration. But exactly how we *calibrate* our duration estimates (i.e. how we learn that mapping) remains to be established – our aim here is not to provide an exhaustive account of how we extract timing information from events. Rather, it is to highlight how the conceptual architecture we have described for both tokenization and event representation (i.e. as ensembles of intersecting object histories) is compatible with what is known about the cognitive machinery that underpins contemporary accounts of how we estimate the duration of events. Similarly, the manner in which the encoding of intersecting object histories maintains important aspects of their temporal dynamic enables distinctions to be made that define the aspectual qualities of events (e.g. Dowty, 1979), as well as the distinction between events (in which objects change state) and states (in which they do not).

5. Action matters

One of the central tenets of IOH is that whereas we typically think of events as entailing change, and as change being (generally) contingent on action, actions are not a representational primitive of event representation (see Section 3.1). Having begun our account of IOH with discussion of this issue, we end our account by returning again to this same issue. Here, we shall focus on the relationship between action and perception, and the theory of event coding (TEC) by Hommel et al. (2001). They distinguish between the "cognitive antecedents of actions that stand for, or represent, certain features of events that are to be generated in the environment … and the motor processes that subserve their realization (i.e. the control and coordination of movements)" (Hommel et al., 2001; p.849). TEC, as an account of action representation and perception is in fact an account of the cognitive antecedents of motor actions (equivalent to representations of the consequences of those actions) and of the "cognitive products" of perception, and how these share a common (event) code (c.f. Pickering and Garrod, 2013, who argue for a parallel with language production and comprehension).

The dynamical framework we have adopted above is entirely compatible with this distinction. We have assumed here (and in Altmann & Mirković, 2009) that the representational products of language comprehension and of event comprehension also share a common code, or rather, a common representational substrate[4]. A challenge that we have not addressed here is how the same theoretical framework we propose as underpinning event comprehension could also underpin action planning. While we cannot attempt here as mechanistic an account of such integration as we have provided for event representation, there are a number of ingredients for such integration that are contained within IOH:

First, we note that accounts of action representation such as TEC assume that actions are driven by intended goal states – actions are not the primitives of action planning, states (and more precisely, changes in state) are the primitives: "action control deals with the intended outcome of an action, not with the particularities of the movement or the sensorimotor interplay producing that outcome" (Hommel et al., 2001; p. 862). Critically, the "particularities of the movement" are not a part of the cognitive representation; they are a response to that cognitive representation. Second, we note that in the action understanding literature, and specifically the developmental literature on action imitation, children tend to interpret actions not as motor sequences (commonly considered the "action") but as goal-directed behaviors in which the goal is the guiding organizational principle (Bekkering, Wohlschlager, & Gattis, 2000); children will commonly reproduce the goal in an imitative task rather than the motor actions that produced that goal (Loucks & Meltzoff, 2013; Meltzoff, 1995). These two points suggest, as outlined earlier, that how the world is, or how it will be (i.e. its state now or in the future) are the building blocks of action representation, at least from the standpoint of the *cognitive* representation of action (as distinct from representations of the effectors, anticipation of their future state, feedback regarding their actual states, etc. during motor preparation and execution). The representational scheme we outlined to explain the encoding of episodic states across time, and hence the encoding of events and object histories, will necessarily encode goal states of the kind described in TEC (IOH is designed to do just that). But there is also the intriguing possibility that it could encode the contingent trajectories through space and time that would lead from the current state to the goal state via motoric action: In Section 3.6 we described the emergence through experiential learning of higher-level contingencies across events – schemas and scripts. Essentially, similar event trajectories (i.e. sequences of events and their concomitant changes in state through time) are abstracted across to give rise to emergent representations corresponding to schema. This is not so different from the concept of emergent action schema, in which the contingencies between body movements and changes in states of the objects in the environment are encoded through experiential learning, leading to the emergence of generalized (motoric) action-effect schema (c.f. one of the basic tenets of ideomotor theory to which Hommel et al., 2001, subscribe). In Section 4 we described how the current state of the system constrains movements into subsequent states (c.f. Elman, 1990); in effect, then, the current state of the system (and the external environment it encodes) constrains selection of the action schemas afforded by that state (Cooper & Glasspool, 2001; Norman & Shallice, 1986). Thus, while IOH may appear radical in eschewing action representations as primitives of event representation (where we now construe "event" in the broader terms of Hommel et al.'s TEC and "event files" – representations that serve both perception and action; Hommel, 2004), it is in fact fully consistent with contemporary accounts of action planning and action control.

6. Summary, and challenges

Events, as conceived here, occur when one or more objects change state as they intersect in time and space. This approach to what constitutes an event is markedly different from that proposed by Zacks and Tversky (2001), for whom events are conceived by the observer. Thus, whereas the IOH maintains that events occupy the physical world, EST maintains that they are psychological phenomena. Beyond this difference, there are many commonalities between the two accounts. Where the IOH differs from EST is primarily in respect of its focus on the *content* of event representations, and *how* this content is encoded. The approach we have taken to event representation, and specifically to the content of such representations, is predicated on objects having trajectories across time, or *histories*, through which they change state. The different roles that objects (participants) play in an event are determined primarily by the spatiotemporal contingencies among these changes (although how these contingencies are interpreted is also constrained by e.g. prior knowledge, attention, salience of individual participants in the event, and so on). These changes in state may be physical (Joan cut her finger with a kitchen knife) or locational (Joan moved to California with her family), but they may also be in non-physical domains: Psychological or emotional/affective (Joan conceived the idea; she realized it was time to leave), functional (Joan's coffee machine broke), regarding kinship (Joan became a mother), and so on, although in many of these cases there may be physical correlates (as in becoming a mother, inheriting a house, or no longer having a functional coffee machine). Further, changes in state may be absolute and permanent (the onion changing from intact to chopped) or more continuous and transient (the chef's changes in posture as she chops the onion). When there is only a single participant in an event, there is just a single trajectory/history of change through time, as in "*the tree fell*". But when there are multiple participants/objects, events are defined by different objects coming together in space and time, with changes in the state of one typically having some particular spatiotemporal relationship with changes in the state(s) of other(s). Thus, events are ensembles of such histories insofar as they intersect within a particular spatiotemporal frame. Describing events in this way raises a number of challenges, not least the burden of explaining how objects can have histories in the first place, how such histories are encoded such that objects persist across change, and what kind of theoretical mechanism might support such encoding. We addressed these challenges in Section 3.3, where we described a mechanism for constructing individuated object tokens on-the-fly (tokenization) and related, the mechanism by which such tokens could accumulate history, with the different fragments of an object's history being bound to that same specific token. We also described how different object trajectories can be construed as intersecting in time, space, and representational (and neural) substrate. Essentially, the mechanism relies on a recurrent architecture in which relational binding through time ensures that features/elements/objects within the

---

[4] The term "code" is too easily associated with explicitly symbolic representational systems. We prefer the more neutral term "representational substrate".

context are bound (i.e. associated) with one another regardless of whether such associations are arbitrary (e.g. a pie and a television) or systematic (e.g. Garfield and Odie). It also ensures binding of activated semantic representations, whether of objects, generalized events, or schemas, with the episodic context. Indeed, the context in which objects find themselves, and to which they are bound, is fundamental to their individuation as tokens. Crucially, we argued that the same mechanisms responsible for tokenization and event representation when directly experiencing events can account for tokenization and event representation when events are instead described through language.

The mechanism we postulated to explain how events are encoded – as trajectories through time – is in fact agnostic to whether there is any change at all. That is, regardless of whether anything actually happens, objects and their spatiotemporal relationships to other objects are encoded in the same way; the picture hanging on the wall opposite is still represented as a trajectory through time even though the object does not change state. Thus, the representational mechanism is the same regardless of whether it is an event being represented or a state. This does lead, however, to a potential point of divergence relative to EST: Any period of time conceived to have a beginning and an end is an event in EST, including, in principle, periods of time in which there is no change in state or location of the protagonists. Thus, if a robot arm places a ball-bearing on a surface, stops still for a few seconds before then moving the ball-bearing elsewhere, that period in which nothing happens would in principle be an event by virtue of having a perceived beginning and end. However, what happens within that period of time (i.e. nothing) reflects a *static state* of the world, albeit sandwiched between two events. We do not designate such states as events precisely because there is no change (if the ball rolled along the surface, we *would* designate it an event). While EST and the IOH might differ on what they call an event or a state, the IOH encodes each in the same way; our theoretical focus is on the *content* of that encoding rather than on the label used to classify it.

There is still much to be explained. Intentionality and causality, for example, are lacking from our account. On the one hand, the assumed encoding of spatiotemporal contingencies within the perceptual stream should be sufficient to capture whatever information drives the phenomenology of intentionality and causality. On the other hand, there are specific properties of this encoding that may be relevant to understanding the ontology of these experiential phenomena. Intentionality reflects the recognition of an anticipated goal state given other possible states; i.e. the recognition of a likely consequence. Anticipating goal states is one of the hallmarks of the IOH. A bias to selectively attend to goals is apparent, even, in infancy (e.g. Lakusta & Carey, 2015). Crucially this bias requires an animate agent (Lakusta & Carey, 2015; see also Lakusta & Landau, 2012, Woodward, 1998). The conjunction of a goal bias and an animate causal agent may be all that is required to drive the perception of intentionality (although Gergely, Nadasdy, & Csibra, 1995, permit intentionality as rational goal-directed behavior in the absence of an animate agent). Animate agency and intentionality go hand-in-hand (cf. early observations by Heider & Simmel, 1944), and the attribution of animacy to a moving object may in part be due to low-level visual processes in brain areas sensitive to biological motion (e.g. Grossman at al., 2000). Thus, the perception of intentionality may have its roots in a combination of perceptual (animacy) and cognitive (privileged status of goals) factors that are compatible with the encodings assumed in the IOH.

The anticipation of goal states in the IOH reflects experiential knowledge of contingencies between the current state of the world and the possible future states into which the world may transition. In developing the IOH, we have focused more on *contingency* than on *causality*, although the two are related: As we repeatedly tell our students, passing an exam is contingent on them studying, and if they do not study they will likely fail their exams – but while we can identify a causal relationship between studying and exam performance, we do not *perceive* causality in the same way here as we do when observing a launching event, for example (in which A moves towards a stationary B, stops at the contact point, at which moment B then launches; Michotte, 1963). Our earlier discussion of causal participant roles, for example, requires the encoding of contingencies, but does not entail the perception of causality in this same launch-like way. Perceiving causality in the case of a launch event is a sensation associated with the apprehension of (certain kinds of) events, and likely due to, or interacting with, low-level visual processes that are sensitive to spatiotemporal coincidence and which allow two spatiotemporally continuous events to be merged into a single event (Rolfs, Dambacher, & Cavanagh, 2013). However, this is a quite different case from the "inference" of causality that accompanies studying and passing exams, or that accompanies the comprehension of a sentence such as "*Joan dropped the jug. It broke*". The perception of causality when viewing a launch event reflects a metacognitive awareness which may not directly contribute to the encoding of the spatiotemporal characteristics of the event; that sensation may be epiphenomenal on those characteristics (see relevant discussion by Scholl & Gao, 2013, on the relationship between perception and higher-level cognitive attribution).

Our intention here has not been to explain all of cognition, even if at times it seems like that is what is required (and in some respects that *is* what is required). Our focus has been on the representational content that distinguishes one event from another, and the cognitive mechanisms that would enable such content. The IOH is necessarily far-reaching because distinguishing one event from another, and even recognizing that an individual event has occurred, requires a representational mechanism able to individuate objects and bind them to their histories (and by "history", we mean trajectory of states through space and time, whether backwards in time or, when anticipated, forwards in time). Several components of the IOH, while explaining existing data, are amenable to further empirical exploration: For example, the role of spatiotemporal continuity of the episodic context in respect of enabling the binding together of temporally separate experiences of an object across changes in its state; the role of long-axis specialization in the hippocampus in respect of the encoding of object states and object identities through time; and the ability to recognize objects that have undergone a change in state as being the same object before and after (i.e. bound to the same object identity) in cases of hippocampal damage (indeed, the ability to fully comprehend that an event has occurred) – these are all cases where further empirical research will refine, or compromise, the IOH. For example, in Section 3.5 we pointed out the role that spatiotemporal continuity of the episodic context may play in binding the representation of an object at one moment in time with a distinct representation of that same object (perhaps in another state) at another moment in time; were we to find that binding of object representations was not impaired by discontinuity of context, our theoretical framework would be compromised. Similarly, if patients with complete damage to the hippocampus were able to create tokens on-the-fly as described in Section 3.3

(and especially parahippocampal cortices, given their role in contextual representation; e.g. Eichenbaum, Sauvage, Fortin, Komorowski, & Lipton, 2012; Wang, Yonelinas, & Ranganath, 2013), we would have to revise how our framework maps onto the underlying neurobiology. A more complete neurobiological model of event cognition will also require further exploration of the interplay between different memory systems and the different neurobiological substrates implicated in semantic memory (for individual concepts and, separately, for schema and situation-based knowledge), episodic memory, and short-term non-hippocampal memory systems, as well as exploration of how these interactions change as a function of e.g. direct observation of an event as compared with processing the equivalent event communicated through language. And while we cannot explain every single aspect of cognition as it pertains to event comprehension and event representation (hence the intended and occasionally unintended omission of pertinent literature), our account is offered as a starting point from which to proceed. It is as much a theory as it is an agenda of items that merit discussion and future investigation.

References

Akyürek, E.G., Toffanin, P., & Hommel, B. (2008). Adaptive control of event integration. *Journal of Experimental Psychology: Human Perception and Performance*, *34*, 569-577.

Allport, D. A. (1985). Distributed memory, modular subsystems and dysphasia. In S. K. Newman & R. Epstein (Eds.), *Current perspectives in dysphasia* (pp. 207–244). Edinburgh, Scotland: Churchill Livingstone.

Altmann, G. T. M. (1997). The Ascent of Babel: An Exploration of Language, Mind, and Understanding. Oxford: Oxford University Press.

Altmann, G. T. M. (1999). Thematic role assignment in context. *Journal of Memory and Language*, *41*(1), 124-145.

Altmann, G. T. M. (2017). Abstraction and generalization in statistical learning: implications for the relationship between semantic types and episodic tokens. *Philosophical Transactions of the Royal Society B*, *372*(1711), 20160060.

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

Altmann, G. T. M., & Mirković, J. (2009). Incrementality and prediction in human sentence processing. *Cognitive Science*, *33*(4), 583-609.

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences*, *22*(04), 577-660.

Barsalou, L. W., Simmons, W. K., Barbey, A. K., & Wilson, C. D. (2003). Grounding conceptual knowledge in modality-specific systems. *Trends in Cognitive Sciences*, *7*(2), 84-91.

Bartlett, F. C. (1932). Remembering: A study in experimental and social psychology. Cambridge, England: Cambridge University Press.

Barwise, J., & Perry, J. Situations and Attitudes. MIT Press; Cambridge, MA: 1983.

Bekkering, H., Wohlschlager, A., & Gattis, M. (2000). Imitation of gestures in children is goal-directed. *The Quarterly Journal of Experimental Psychology: Section A*, *53*(1), 153-164.

Carey, S., & Xu, F. (2001). Infants' knowledge of objects: Beyond object files and object tracking. *Cognition*, *80*(1), 179-213.

Carey, S. (2011). Précis of the origin of concepts. *Behavioral and Brain Sciences*, *34*(03), 113-124.

Chen, Y. C., & Scholl, B. J. (2016). The perception of history: Seeing causal history in static shapes induces illusory motion perception. *Psychological Science*, *27*(6), 923-930.

Cohen, N. J., & Eichenbaum, H. (1993). *Memory, amnesia, and the hippocampal system*. Cambridge, MA: MIT Press.

Coll-Florit, M., & Gennari, S. P. (2011). Time in language: Event duration in language comprehension. *Cognitive Psychology*, *62*(1), 41-79.

Collin, S. H., Milivojevic, B., & Doeller, C. F. (2015). Memory hierarchies map onto the hippocampal long axis in humans. *Nature Neuroscience*, *18*(11), 1562-1564.

Colzato, L.S., Raffone, A., & Hommel, B. (2006). What do we learn from binding features? Evidence for multilevel feature integration. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 705-716.

Cooper, R., & Glasspool, D. (2001). Learning action affordances and action schemas. In *Connectionist models of learning, development and evolution* (pp. 133-142). Springer, London.

Craik, F. I., Govoni, R., Naveh-Benjamin, M., & Anderson, N. D. (1996). The effects of divided attention on encoding and retrieval processes in human memory. *Journal of Experimental Psychology: General*, *125*(2), 159-180.

Dehaene, S. (1993). Temporal oscillations in human perception. *Psychological Science*, *4*(4), 264-270.

Dowty, D. (1979). Word meaning and Montague grammar. Dordrecht, The Netherlands: Kluwer Academic.

Duff, M. C., & Brown-Schmidt, S. (2012). The hippocampus and the flexible use and processing of language. *Frontiers in Human Neuroscience*, *6*.

Eichenbaum, H., Sauvage, M., Fortin, N., Komorowski, R., & Lipton, P. (2012). Towards a functional organization of episodic memory in the medial temporal lobe. *Neuroscience & Biobehavioral Reviews*, *36*(7), 1597-1608.

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

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, 48, 71–99.

Elman, J. L., Bates, E. A., Johnson, M. H., Karmiloff-Smith, A., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A Connectionist Perspective on Development*. Cambridge, MA: MIT Press.

Elman, J. L., & McRae, K. (2019). A model of event knowledge. *Psychological Review*, *126*(2), 252-291.

Faber, M., & Gennari, S. P. (2015). Representing time in language and memory: The role of similarity structure. *Acta Psychologica*, *156*, 156-161.

Ferretti, T. R., McRae, K., & Hatherell, A.(2001). Integrating verbs, situation schemas, and thematic role concepts. *Journal of Memory and Language*, *44*, 516–547.

Fine, M. S., & Minnery, B. S. (2009). Visual salience affects performance in a working memory task. *Journal of Neuroscience*, *29*(25), 8016-8021.

Gergely, G., Nfidasdy, Z., & Gergely Csibra, S. B. (1995). Taking the intentional stance at 12 months of age. *Cognition*, *56*, 165-193.

Gibson, J. J. (1979). The ecological approach to visual perception. Houghton- Mifflin.

Glenberg, A. M. (1997). What memory is for: Creating meaning in the service of action. *Behavioral and Brain Sciences*, *20*(01), 41-50.

Grill-Spector K, Kourtzi Z, & Kanwisher N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*, 1409–1422.

Grill-Spector K, & Malach R. (2004). The human visual cortex. *Annual Reviews Neuroscience*, *27*, 649–677.

Grossman, E., Donnelly, M., Price, R., Pickens, D., Morgan, V., Neighbor, G., & Blake, R. (2000). Brain areas involved in perception of biological motion. *Journal of Cognitive Neuroscience*, *12*(5), 711-720.

Harrison, S. A., & Tong, F. (2009). Decoding reveals the contents of visual working memory in early visual areas. *Nature*, *458*(7238), 632-635.

Hassabis, D., & Maguire, E. A. (2007). Deconstructing episodic memory with construction. *Trends in Cognitive Sciences*, *11*(7), 299-306.

Heider, F., & Simmel, M. (1944). An experimental study of apparent behavior. *The American Journal of Psychology*, *57*(2), 243-259.

Hindy, N. C., Altmann, G. T. M, Kalenik, E., & Thompson-Schill, S. L. (2012). The effect of object state-changes on event processing: Do objects compete with themselves? *The Journal of Neuroscience*, *32*(17), 5795-5803.

Hindy, N. C., Hamilton, R., Houghtling, A. S., Coslett, H. B., & Thompson-Schill, S. L. (2009). Computer-mouse tracking reveals TMS disruptions of prefrontal function during semantic retrieval. *Journal of Neurophysiology*, *102*(6), 3405-3413.

Hindy, N. C., Solomon, S. H., Altmann, G. T. M, & Thompson-Schill, S. L. (2015). A cortical network for the encoding of object change. *Cerebral Cortex*, *25*(4), 884-894.

Hirsch, E. (1992). The concept of identity. Oxford University Press.

Hommel, B. (2004). Event files: Feature binding in and across perception and action. *Trends in Cognitive Sciences*, *8*(11), 494-500.

Hommel, B., & Colzato, L.S. (2009). When an object is more than a binding of its features: Evidence for two mechanisms of visual feature integration. *Visual Cognition*, *17*, 120-140.

Hommel, B., Müsseler, J., Aschersleben, G., & Prinz, W. (2001). The theory of event coding (TEC): A framework for perception and action planning. *Behavioral and Brain Sciences 24* (5), 849-878.

January, D., Trueswell, J. C., & Thompson-Schill, S. L. (2009). Co-localization of Stroop and syntactic ambiguity resolution in Broca's area: Implications for the neural basis of sentence processing. *Journal of Cognitive Neuroscience*, 21(12), 2434-2444.

Joanisse, M. F., & Seidenberg, M. S. (2003). Phonology and syntax in specific language impairment: Evidence from a connectionist model. *Brain and Language*, *86*(1), 40-56.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201-211.

Johnson-Laird, P. N. (1983). Mental models: Towards a cognitive science of language, inference, and consciousness. Cambridge, MA: Harvard University Press.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175-219.

Kalénine, S., Mirman, D., Middleton, E. L., & Buxbaum, L. J. (2012). Temporal dynamics of activation of thematic and functional knowledge during conceptual processing of manipulable artifacts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*(5), 1274.

Konkel, A., & Cohen, N. J. (2009). Relational memory and the hippocampus: representations and methods. *Frontiers in Neuroscience*, *3*(2), 166-174.

Kratzer, A. (1995). Stage-level and Individual-level Predicates. In G.N. Carlson & F. J. Pelletier (Eds.) *The Generic Book*. Chicago: The University of Chicago Press.

Kumaran D., & McClelland J.L., 2012. Generalization through the recurrent interaction of episodic memories: a model of the hippocampal system. *Psychological Review*, 119, 573 – 616.

Kurby, C. A., & Zacks, J. M. (2008). Segmentation in the perception and memory of events. *Trends in Cognitive Sciences*, 12(2), 72-79.

Kutas, M., and Hillyard, S. A., Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science* 207.4427 (1980): 203-205.

Lakusta, L., & Carey, S. (2015). Twelve-month-old infants' encoding of goal and source paths in agentive and non-agentive motion events. *Language Learning and Development*, *11*(2), 152-175.

Lakusta, L., & Landau, B. (2012). Language and memory for motion events: Origins of the asymmetry between source and goal paths. *Cognitive Science*, *36*(3), 517-544.

Long, L. L., Bunce, J. G., & Chrobak, J. J. (2015). Theta variation and spatiotemporal scaling along the septotemporal axis of the hippocampus. *Frontiers in Systems Neuroscience*, *9*. doi:10.3389/fnsys.2015.00037.

Loucks, J., & Meltzoff, A. N. (2013). Goals influence memory and imitation for dynamic human action in 36-month-old children. *Scandinavian Journal of Psychology*, *54*(1), 41-50.

Mack, M. L., & Preston, A. R. (2016). Decisions about the past are guided by reinstatement of specific memories in the hippocampus and perirhinal cortex. *NeuroImage*, *127*, 144-157.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163.

McCarthy, J. (1963). Situations, Actions and Causal Laws. Technical Report, Stanford University Department of Computer Science .

McRae, K., Hare, M., Elman, J. L., & Ferretti, T. (2005). A basis for generating expectancies for verbs from nouns. *Memory & Cognition*, *33*(7), 1174-1184.

Metusalem, R., Kutas, M., Urbach, T. P., Hare, M., McRae, K., & Elman, J. L. (2012). Generalized event knowledge activation during online sentence comprehension. *Journal of Memory and Language*, *66*(4), 545-567.

Meltzoff, A. (1995). Understanding the Intentions of Others: Re-Enactment of Intended Acts by 18-Month-Old Children. *Developmental Psychology, 31*(5), 838-850.

Metzler, C. (2001). Effects of left frontal lesions on the selection of context-appropriate meanings. *Neuropsychology*, *15*(3), 315.

Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227.

Michotte, A .(1963) *The perception of causality*. New York: Basic Books.

Moens, M., & Steedman, M. (1988). Temporal ontology and temporal reference. *Computational Linguistics*, *14*(2), 15-28.

Moscovitch, M., Cabeza, R., Winocur, G., & Nadel, L. (2016). Episodic memory and beyond: the hippocampus and neocortex in transformation. *Annual Review of Psychology*, 67, 105-134.

Moss, H. E., Ostrin, R. K., Tyler, L. K., & Marslen-Wilson, W. D. (1995). Accessing different types of lexical semantic information: Evidence from priming. *Journal of Experimental Psychology: Learning, memory, and cognition*, *21*(4), 863.

Norman, D. A., & Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In R. J. Davidson & G. E. Schwartz & D. Shapiro (Eds.), *Consciousness and self regulation: Advances in research, Vol. IV* (Vol. IV). New York: Plenum Press.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition*, *107*(1), 179-217.

Pettijohn, K.A., Thompson, A.N., Tamplin, A.K., Krawietz, S.A, Radvansky, G.A. (2016). Event boundaries and memory improvement. *Cognition*, *148,* 136-144.

Pickering, M.J., and Garrod, S. (2013) An integrated theory of language production and comprehension. *Behavioral and Brain Sciences*, *36*(4), pp. 329-347.

Pinker, S. (1984). Language learnability and language development. Cambridge, MA Harvard University Press.

Poppenk, J., Evensmoen, H. R., Moscovitch, M., & Nadel, L. (2013). Long-axis specialization of the human hippocampus. *Trends in Cognitive Sciences*, *17*(5), 230-240.

Preston, A. R., & Eichenbaum, H. (2013). Interplay of hippocampus and prefrontal cortex in memory. *Current Biology*, *23*(17), R764-R773.

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65-97.

Radvansky, G. A. (2012). Across the event horizon. *Current Directions in Psychological Science*, *21*(4), 269-272.

Radvansky, G. A., & Zacks, R. T. (1991). Mental models and the fan effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *17*(5), 940.

Radvansky, G. A., & Zacks, J. M. (2011). Event perception. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*(6), 608-620.

Radvansky, G. A., & Zacks, J. M. (2014). Event Cognition. New York: Oxford University Press.

Radvansky, G. A., & Zacks, J. M. (2017). Event boundaries in memory and cognition. *Current Opinion in Behavioral Sciences*, *17*, 133-140.

Rappaport Hovav, M., & Levin, B. (1998). Building verb meanings. In M. Butt & W. Geuder (Eds.) *The projection of arguments: Lexical and compositional factors*. CSLI Publications.

Reynolds, J.R., Zacks, J.M., & Braver, T.S. (2007). A computational model of event segmentation from perceptual prediction. *Cognitive Science*, 31, 613-643.

Robinson, G., Blair, J., & Cipolotti, L. (1998). Dynamic aphasia: an inability to select between competing verbal responses? *Brain*, *121*(1), 77-89.

Robinson, G., Shallice, T., & Cipolotti, L. (2005). A failure of high level verbal response selection in progressive dynamic aphasia. *Cognitive Neuropsychology*, *22*(6), 661-694.

Rochat, P., Striano, T., & Morgan, R. (2004). Who is doing what to whom? Young infants' developing sense of social causality in animated displays. *Perception*, *33*(3), 355-369.

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

Rolfs, M., Dambacher, M., & Cavanagh, P. (2013). Visual adaptation of the perception of causality. *Current Biology*, *23*(3), 250-254.

Saigusa, T., Tero, A., Nakagaki, T., & Kuramoto, Y. (2008). Amoebae anticipate periodic events. *Physical Review Letters*, *100*(1), 018101.

Sakarias, M., & Flecken, M. (2019). Keeping the Result in Sight and Mind: General Cognitive Principles and Language-Specific Influences in the Perception and Memory of Resultative Events. *Cognitive Science*, 43(1). doi: 10.1111/cogs.12708

Schacter, D. L., & Addis, D. R. (2007). The cognitive neuroscience of constructive memory: remembering the past and imagining the future. *Philosophical Transactions of the Royal Society B*, *362*(1481), 773-786.

Schank, R. C., & Abelson, R. P. (1977). Scripts, plans, goals, and understanding: An inquiry into human knowledge structures. Lawrence Erlbaum Associates.

Schapiro A.C., Turk-Browne N.B., Botvinick M.M., Norman K.A. (2017). Complementary learning systems within the hippocampus: a neural network modelling approach to reconciling episodic memory with statistical learning. *Philosophical Transactions of the Royal Society B*, 372, 20160049.

Scholl, B. J. (2001). Objects and attention: The state of the art. *Cognition*, *80*(1), 1-46.

Scholl, B. J. (2007). Object persistence in philosophy and psychology. *Mind & Language*, *22*(5), 563-591.

Scholl, B. J., & Gao, T. (2013). Perceiving animacy and intentionality: Visual processing or higher-level judgment. In M.D. Rutherford and V.A. Kuhlmeier (Eds.), *Social perception: Detection and interpretation of animacy, agency, and intention*. The MIT Press: Cambridge, MA. pp. 197-230.

Smith L., & Yu C., (2008). Infants rapidly learn word-referent mappings via cross-situational statistics. *Cognition*, 106, 1558 – 1568.

Solomon, S. H., Hindy, N. C., Altmann, G. T., & Thompson-Schill, S. L. (2015). Competition between mutually exclusive object states in event comprehension. *Journal of Cognitive Neuroscience*. 27(12), 2324-2338.

Spröte, P., Schmidt, F., & Fleming, R. W. (2016). Visual perception of shape altered by inferred causal history. *Scientific Reports*, *6*, 36245.

Tabor, W., & Tanenhaus, M. K. (1999). Dynamical models of sentence processing. *Cognitive Science*, *23*(4), 491-515.

Tamplin, A. K., Krawietz, S. A., Radvansky, G. A., & Copeland, D. E. (2013). Event memory and moving in a well-known environment. *Memory & Cognition*, *41*(8), 1109-1121.

Thompson-Schill, S. L., Bedny, M., & Goldberg, R. F. (2005). The frontal lobes and the regulation of mental activity. *Current Opinion in Neurobiology*, *15*(2), 219-224.

Tulving E. 1983. Elements of Episodic Memory. Oxford, UK: Clarendon

van Dijk, T. A., & Kintsch, W. (1983). Strategies in discourse comprehension. New York: Academic Press

van Kesteren, M. T., Beul, S. F., Takashima, A., Henson, R. N., Ruiter, D. J., & Fernández, G. (2013). Differential roles for medial prefrontal and medial temporal cortices in schema-dependent encoding: from congruent to incongruent. *Neuropsychologia*, *51*(12), 2352-2359.

van Kesteren, M. T., Rijpkema, M., Ruiter, D. J., & Fernández, G. (2010). Retrieval of associative information congruent with prior knowledge is related to increased medial prefrontal activity and connectivity. *Journal of Neuroscience*, 30(47), 15888-15894.

van Kesteren, M. T., Ruiter, D. J., Fernández, G., & Henson, R. N. (2012). How schema and novelty augment memory formation. *Trends in Neurosciences*, *35*(4), 211-219.

VanRullen, R., Reddy, L., & Koch, C. (2005). Attention-driven discrete sampling of motion perception. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(14), 5291-5296.

Vendler, Z. (1957). Verbs and time. *Philosophical Review*, 56, 143–160.

Warglien, M., Gärdenfors, P., & Westera, M. (2012). Event structure, conceptual spaces and the semantics of verbs. *Theoretical Linguistics*, 38(3-4), 159-193.

Wang, W. C., Yonelinas, A. P., & Ranganath, C. (2013). Dissociable neural correlates of item and context retrieval in the medial temporal lobes. *Behavioural Brain Research*, *254*, 102-107.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, 9(4), 625-636.

Winograd, T. (1972). Understanding natural language. *Cognitive Psychology*, *3*(1), 1-191.

Woodward, A. L. (1998). Infants selectively encode the goal object of an actor's reach. *Cognition*, *69*(1), 1-34.

Yee, E. (2017). Fluid semantics: Semantic knowledge is experience-based and dynamic. In A. Lahiri & S. Kotzor (Eds), *The Speech Processing Lexicon: neurocognitive and behavioural approaches* (pp. 235-262). Berlin/Boston: Mouton

Yee, E., Drucker, D.M., & Thompson-Schill, S.L. (2010). fMRI-adaptation evidence of overlapping neural representations for objects related in function or manipulation. *NeuroImage*, 50, 753-763.

Yee, E., & Sedivy, J. C. (2006). Eye movements to pictures reveal transient semantic activation during spoken word recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*(1), 1-14.

Zacks, J. M., Speer, N. K., Swallow, K. M., Braver, T. S., & Reynolds, J. R. (2007). Event perception: a mind-brain perspective. *Psychological Bulletin*, *133*(2), 273.

Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. *Psychological Bulletin*, *127*(1), 3-21.

Zeithamova, D., Dominick, A. L., & Preston, A. R. (2012). Hippocampal and ventral medial prefrontal activation during retrieval-mediated learning supports novel inference. *Neuron*, *75*(1), 168-179.

Zwaan, R. A. (1996). Processing narrative times shifts. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 22, 1196-1207.

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

Zwaan, R. A., Langston, M. C., & Graesser, A. C. (1995). The construction of situation models in narrative comprehension: An event-indexing model. *Psychological Science*, 292-297.