Finding *Event Structure in Time*: What Recurrent Neural Networks can tell us about *Event Structure in Mind*

Forrest Davis
Cornell University

Gerry T.M. Altmann
University of Connecticut

Under a theory of event representations that defines events as dynamic changes in objects across both time and space, as in the proposal of Intersecting Object Histories (Altmann & Ekves, 2019), the encoding of changes in state is a fundamental first step in building richer representations of events. In other words, there is an inherent dynamic that is captured by our knowledge of events. In the present study, we evaluated the degree to which this dynamic was inferable from just the linguistic signal, without access to visual, sensory, and embodied experience, using recurrent neural networks (RNNs). Recent literature exploring RNNs has largely focused on syntactic and semantic knowledge. We extend this domain of investigation to representations of events within RNNs. In three studies, we find preliminary evidence that RNNs capture, in their internal representations, the extent to which objects change states; for example, that chopping an onion changes the onion by more than just peeling the onion. Moreover, the temporal relationship between state changes is encoded to some extent. We found RNNs are sensitive to how chopping an onion and then weighing it, or first weighing it, entails the onion that is being weighed being in a different state depending on the adverb. Our final study explored what factors influence the propagation of these rudimentary event representations forward into subsequent sentences. We conclude that while there is much still to be learned about the abilities of RNNs (especially in respect of the extent to which they encode objects as specific tokens), we still do not know what are the equivalent representational dynamics in humans. That is, we take the perspective that the exploration of computational models points us to important questions about the nature of the human mind.

Keywords: Recurrent Neural Networks; event representation, discourse dependencies.

INTRODUCTION

Semantic space is all around us. Contemporary approaches to semantic memory, both its computer and human instantiations, have converged on the idea that semantic knowledge – the knowledge we have of the world around us and the things it contains – is organized in such a way as to encode similarity between concepts along multiple dimensions (e.g. Yee, Jones, & McRae, 2018). LSA (Landauer & Dumais, 1997) and HAL (Lund & Burgess, 1996) were conceptually simple approaches to generating such similarity spaces by computer. More recently, a number of additional approaches to generating semantic similarity spaces have evolved (see also Perconti & Plebe, 2020), including word2vec (Mikolov, Sutskever, Chen, Corrado, & Dean, 2018), ELMo (Peters, Neumann, Iyyer, Gardner, Clark, Lee, & Zettlemoyer, 2018), BERT (Devlin, Chang, Lee, & Toutanova, 2018), and ERNIE 2.0 (Sun, Wang, Li, Feng, Tian, Wu, & Wang, 2020). Each of these is based on the same underlying principle as govern LSA and HAL – their internal representation of a word, after learning, is a reflection of the contexts in which it occurred and the other words with which it co-occurred in those contexts (for now we gloss over the fact that some models reflect contextual co-occurrences as representations abstracted across individual co-occurrences, while no longer encoding those individual co-occurrences – e.g. LSA, while other kinds of model can reflect contextual co-occurrences not only as abstracted representations but also as representations that maintain those individual co-occurrences – e.g. BERT). These models underpin almost all practical AI (Artificial Intelligence) approaches to NLP (natural language processing). And while the implementations vary in respect of the (deep learning) technologies they require, they all capture that same underlying principle – words that are experienced in similar contexts will have similar meanings and will thus be "closer" in semantic space than words occurring in more dissimilar contexts. However, there are important differences between these models: After learning, HAL, LSA, and word2vec

The authors would like to thank the anonymous reviewers whose comments on prior versions of the manuscript significantly improved this paper. We thank also Marten van Schijndel, Eiling Yee, Yanina Prystauka, and Zac Ekves for their continuing and thoughtful discussion of this work. Finally, there is an unwritten convention that one does not thank the Editor. We therefore cannot thank Dick Aslin for his astute comments and insightful guidance that helped bring this paper into being.

return the same representation for a word regardless of the actual context in which that word might subsequently occur. Contemporary models of semantic memory, on the other hand, propose that concepts are *dynamic* – the knowledge we access about any given concept depends on the context in which we access that knowledge – as defined, for example, by task demands, the broader situation in which the knowledge is accessed, or the idiosyncratic experience of the individual accessing the concept (e.g. Yee & Thompson-Schill, 2016; Mirkovic & Altmann, 2019). BERT, ELMo, and ERNIE 2.0 do return different word representations (*word embeddings* or *vectors* that constitute a mapping from word form to semantic space) depending on the context, but while they might provide useful insights into the information that a semantic space might ideally (or in practice) encode, they are not intended as, and nor are they, psychologically plausible models of (human) natural language processing (and specifically, natural language *learning*).

Our focus here is not on semantic space *per se* but on *representation* (or its equivalent in a dynamical system – see below). Equally, our focus is not on *lexical* representation but on *event* representation. We shall describe a number of "simulations" with Recurrent Neural Networks (RNNs: similar to Elman's Simple Recurrent Network (SRN: Elman, 1990) but with more than one hidden layer) using Long Short-Term Memory units (LSTMS – see below). These essentially scale up the insights that Elman reported with respect to emergent representations of syntactic and semantic dependencies (Elman, 1990, 1993). Below, we apply some of these insights to learned representations of *discourse* and *event* dependencies, in networks with vocabularies in the tens of thousands. Our aim is not to build a better NLP device, but to better understand the possible (and probable) encoding of event structure in the human mind. We return in the General Discussion to the relationship between RNNs and models such as ELMo (essentially a bidirectional RNN) and BERT.

Below, we explore whether RNNs can develop sensitivity to the essential content of event representations (Studies 1 and 2 below), and the factors that may influence the propagation of both linguistically relevant and event relevant representations through time and context (Study 3). Our goal is not to evaluate whether RNNs or some other computational model are the model that best fits human behavior, but instead to try to understand how a possible mechanism for acquiring and processing event representations (recurrence) may relate to human behavior. For example, it may be the case that simple exposure to language and/or corresponding variation in the external world is sufficient to enable the emergence of event-relevant behaviors. At issue is: *how*? As outlined below, our results call for deeper understanding of how *humans* maintain (and change) linguistic representations while processing language, using similar techniques to the computational approach we outline below (in Study 3).

Event Representation

We follow Altmann & Ekves (2019) in assuming that an event occurs when, minimally, an object changes state across time. On this approach, event representations are grounded in representations of object *histories* – the distinct states of an object across time. If a knife cuts through an onion, both the knife and the onion change state (albeit in different ways) – both the knife and the onion can be considered as *trajectories* through space and time whose intersection defines the event in respect of its participants and the changes in state they endure. Altmann & Ekves (2019) referred to this as the Intersecting Object Histories account of event representation (the "IOH").

The IOH makes certain assumptions about the "computational substrate" that are based on properties of SRNs originally observed by Elman (and see Altmann & Mirkovic, 2009, for fuller discussion of these, and their relevance for sentence processing and event representation). The representation of a sentence – the linguistic realization of an event – is, in the SRN, a trajectory across time (afforded by recurrence through time), with different sentence types having different trajectories that nonetheless reflect similarity in structure. These trajectories also reflect the constraints acquired through learning on how, at each point in the trajectory, the trajectory may continue. We can therefore operationalize *representation* in Elman's networks not simply as activation patterns across the hidden layers, but as constraints on which patterns can follow which other patterns.

At issue for present purposes, is whether, and how, a system based on a recurrent architecture like the SRN might acquire event-relevant structure. At a minimum, and following the approach described above in relation to the IOH, we contend that it would need to track individual entities *across* sentences (similar in some respects to tracking entities across events or situations), and to track changes to those entities as a consequence of the events they participate in.

Keeping track of discourse entities across sentence boundaries

Early work on the influence of inter-sentential discourse dependencies on subsequent (linguistic) behaviors concerned the influence of context on ambiguity resolution (Tyler and Marslen-Wilson, 1977; Crain & Steedman, 1986; Altmann & Steedman, 1988). In one such study (Spivey-Knowlton, Trueswell, & Tanenhaus, 1993), the presence of more than one possible antecedent in an earlier sentence (e.g. "*Two knights were attacking a dragon … the dragon killed one of the knights*") impacted on behaviors in a later sentence (e.g. beginning "*The knight killed…*") – Altmann & Steedman's Principle of Referential Support (Altmann & Steedman, 1988) predicted that a simple noun phrase which failed to pick out the intended referent ("*The knight killed*") would be interpreted as the first noun phrase in a complex noun phrase construction ("*The knight killed by the dragon fell to the ground with a thud*"). Thus, depending on the referential context, syntactic ambiguities ("*killed*" as a main verb or as a past participle in a reduced relative clause) will be resolved one way or another. Crucially, this requires that information about entities introduced earlier on (the knights) is propagated forwards, from one sentence to the next (the basis for anaphoric dependencies across sentences). Could an RNN learn to use referential context to resolve syntactic ambiguities of this kind?

To explore this, Davis and Schijndel (2020) trained RNNs with LSTMs on an 80 million word subset of Wikipedia. SRNs have relatively short memories – long-distance dependencies are difficult for an SRN to learn because the hidden layer gets more and more information added to it on each successive time-step, and resolving information from the more distant past gets progressively harder (this is an over-simplification, but the intuition will suffice). LSTMs (Hochreiter & Schmidhuber, 1997) are units that overcome this by maintaining information across successive time-steps (they each have a dynamic memory that the network learns to modify, update, and draw from depending on the current word and the preceding context; they learn to balance the need to propagate information forward in time with the need to

modify or even forget that information). During training, the networks (each with different random initializations) had to predict each successive word in the corpus given the preceding words (Elman's 1990 prediction task). After training, each network was tested on the 16 actual stimuli from Spivey-Knowlton et al. (1993). *Surprisal* (Hale, 2001) was calculated at the phrase *verb+by* ("*killed by*" in the example above). Surprisal was lower for the networks when the preceding context contained two referents than when it contained one ("*A knight and his squire were attacking a dragon*"). Similar patterns were observed for main verb / reduced relative ambiguities when embedded not in referential contexts but in temporally supporting contexts (Trueswell & Tanenhaus, 1991). There were no such differences in surprisal for networks trained on versions of the corpus in which the order of sentences was scrambled (thereby breaking any inter-sentential dependencies). It would appear, then, that RNNs with LSTMs *can*, at least to an extent (see below), track discourse entities across sentence boundaries and if necessary resolve referential ambiguities in order to establish continuity of reference. Or at least, like humans, they *behave* as if this is what they do.

Representational content: Tracking object states across events

While impressive, the Davis and Schijndel (2020) demonstration does not tell us the nature of the information that was propagated across the context-sentence pairings. Surprisal tells us about lexical expectations, but it does not tell us about the internal representational *content* that the models propagated across each sentence and which presumably was the causal antecedent of these expectations. Encoding event representations requires the encoding not only of information about the entities themselves, but also about the states they pass through as events unfold. If a chef chops an onion and then weighs it, the thing being weighed is chopped. If the chef chops an onion *but first* weighs it, the thing being weighed is not chopped. In these cases, the object representation that propagates through the second clause (beginning "*and then*" or "*but first*") should reflect the event-related changes to the onion that occurred in the first clause, depending on the temporal adverbial. Importantly, the state that the onion is in depends on *which* onion is referred to in the second clause: Chopping an onion and then weighing *another* onion means that the thing being weighed is (most likely) not a chopped onion. These different scenarios (both the real-world scenarios and their linguistic equivalents) exemplify what we mean by propagating a representation forward in time and reflecting whatever changes in state it undergoes as it transitions from one event (chopping) to another (weighing). Could an RNN spontaneously develop the appropriate representational generalizations just through exposure to a large but otherwise arbitrarily chosen language corpus?

Within a semantic similarity space typical of a model such as LSA, different objects (concepts) occupy different parts of the space. But they do not occupy single points, they occupy *regions* of space, with different points within the region reflecting different contextual dependencies associated with the different states of the object. The concept corresponding to the object "onion" is a region of space that includes yellow, white, and red onions, peeled onions, chopped onions, and fried onions (for related discussion see Solomon, Medaglia, & Thompson-Schill, 2019). In principle, then, an RNN should be able to develop emergent categories (i.e. regions of space) that are structured in such a way as to capture the contextual dependencies between verbs such as "*chop*" and nouns such as "*onion*" that specify the distinct states (points in the space) that should be activated as a sentence such as "*chop the onion*" unfolds. More interesting is whether the network can then propagate the appropriate states (the appropriate representational content) across multiple sentences such that when "*weigh the onion*" is encountered, the onion is still at that same point in space (albeit displaced slightly by the weighing). Importantly, distance within the region of space, between one point and another, could in principle reflect the degree of change that the object undergoes as it is displaced in the space (as modulated by the verb) – chopping the onion might displace that particular onion within the onion region by more than peeling the onion, reflecting more movement along the different featural dimensions along which the onion changes: The more change, the more movement, and the greater the distance. But how could a network learn, from linguistic input alone, the relevant featural dimensions along which objects change as they participate in events? While featural dimensions may emerge as an abstraction over the network's hidden unit activations (c.f. Elman, 1993), what the network is *exposed* to is more akin to *affordances* (e.g., Gibson, 1979; Glenberg, 1997), albeit in the linguistic domain. A sentence describing an event in which an onion is chopped is unlikely to be followed by a sentence in which that same onion is then peeled. So given that the RNNs are in the business of predicting upcoming input, event descriptions constrain what upcoming descriptions are afforded by the current input (reflecting the real world affordances that accompany events across time). Whether RNNs can develop sensitivity to affordances of these kinds is the basis for the studies to which we now turn.

Preview of the studies and main results

In all three studies we adopted the same computational architecture as in Davis & Schijndel (2020). **Study 1** was motivated by Hindy, Altmann, Kalenik, & Thompson-Schill (2012), who collected degree-of-change ratings for the sentence pairs that they used in an fMRI study of object-state change. Participants had been instructed to read sentences such as "*The chef will chop the onion*" or "*The man will choose the bagel*" and to rate on a 7-point scale by how much the thing that was acted upon in the sentence changed relative to how it had been before it was acted upon. Inspired by Representational Similarity Analyses (RSA, Kriegeskorte, Mur, & Bandettini, 2008) we compare the activation patterns across the hidden layers at the end of each sentence to the activation patterns that resulted from presenting to the networks the indefinite form of the noun referenced at the end of the sentence (e.g. "*an onion*" or "*a bagel*" for the two examples above). We found that the network's unfolding representations (as indexed by the similarity to this baseline; see Supplemental Material A) correlated with the degree-of-change ratings in human participants asked to judge the exact same sentences.

In **Study 2** we asked whether the representation of the target object at the end of the sentence would propagate appropriately into a second sentence. We contrasted matched pairs of two-sentence sequences such as "*The chef will chop the onion. Then, she will weigh the onion*" and "*The chef will chop the onion. First, she will weigh the onion*". In the THEN condition, the onion at the end of the second sentence is chopped – it should be dissimilar to a prototypical onion (indexed by the activation pattern due to just the word "*onion*"). In the FIRST condition, the onion at the end of the second sentence is being referenced in its prior unchopped state, and so should be more similar to a prototypical onion

(relative to the observed similarity in the THEN condition). This is exactly what we found in our similarity analyses. And when we replaced "*the onion*" at the end of the second sentence in the THEN condition with "*another onion*", the networks, like people (Solomon, Hindy, Altmann, & Thompson-Schill, 2015) treated this onion as a more prototypical onion.

Whereas Study 2 explored how representations of the object undergoing change propagate through the sentences, **Study 3** explored how representations of the sentential *subject* likewise propagate. This more exploratory study used similarity through time to track the representation of the sentential subject through to the end of the first sentence and into both the second and a third. We found, unsurprisingly, that the sentential subject *does* propagate through the sentences, but that its representation changes dynamically as a function of other input and its perturbation of the network's activation state. We interpret these dynamics as reflecting the extent to which the linguistic input places constraints on what states the network, as a dynamical system, can move into next. We now turn to the studies in detail before discussing the implications of these results for our understanding of both network, and human, behavior.

STUDIES

Neural Networks

We followed the architectural details in Davis and van Schijndel (2020). Specifically, we trained RNNs with LSTM hidden units using a language modeling objective (i.e. predicting the next word; as in Elman, 1990).[1] The models had two LSTM layers with 400 hidden units each, 400-dimensional word embeddings, a dropout rate of 0.2 and batchsize 20. They were trained for 40 epochs (with early stopping) using PyTorch. To disassociate effects of training data, we trained two sets of models on different data. The first (Wikipedia models; N=25) was trained on approximately 103 million tokens of preprocessed Wikipedia text taken from verified higher quality articles (Wikitext-103; Merity, Xiong, Bradbury, & Socher, 2016). The other set of models (Web models; N=25) was trained on approximately 100 million tokens of web data taken from URL links in "higher quality" reddit posts, which crucially excluded all Wikipedia data (OpenWebTextCorpus; Gokaslan and Cohen, 2019).[2] Sentence length was similar across the two corpora (18 and 17, respectively) although there was greater variance in the Web corpus (standard deviations: 15 and 22 respectively). Each of the models was initialized with a different set of connection weights. The vocabularies of the models were constrained to the top 50K most frequent words in their respective training corpora. Words were represented using one-hot encodings (that is, 49,999 bits "off" and one bit "on") for the input, and the output at each time step was a probability distribution for the next word ranging over the vocabulary.

Study 1: RNN encoding of object-state change

The first study evaluates the internal representations of RNN language models while processing stimuli that describe a change in state of an object. "Knowledge" of events under IOH requires "knowledge" of object trajectories -- a network that builds event representations should represent object affordances under different contexts, corresponding to the objects in a real-world event undergoing a change in state. Blended mangoes afford different interactions than do whole mangoes, with the blending causing changes in state that are accompanied by different sets of affordances. In principle, the consequences of changes in state on the affordances of the object (in its new state) should manifest in the language used to describe events and their consequences (changes in state would be accompanied by changes in what may unfold next).

Using stimuli rated for degree-of-change (the amount an object was changed by an action), we evaluated whether RNNs encoded degree of change in a way that mapped onto human judgments. Sentences with corresponding human ratings were taken from Hindy et al. (2012) and pooled together with an additional set of stimuli and ratings developed by Prystauka, Ekves, and Altmann (in preparation). Across all our studies we selected the maximum number of stimuli from this original pool of 326 stimuli that satisfied the constraints of the study (e.g. that all words were known to the networks, that no verb + object combination appeared more than once, and depending on the study, that the stimuli either were paired ("minimum" vs "substantial" change) or were drawn from the same category (e.g. as in Study 2 below). The resulting 145 stimulus pairs had the following structure:

The chef will weigh the mango [minimal change implied by the verb]
The chef will blend the mango [substantial change implied by the verb]

Hindy et al. (2012) had used such pairs to show (among other effects) that the fMRI BOLD response elicited by such sentences correlated with degree-of-change ratings supplied by a separate group of participants. We excluded any stimulus pair that had any word in either sentence of the pair that was not contained within the models' vocabularies. This left 136 sentence pairs for the Wikipedia models and 140 pairs for the Web models. Ratings had previously been collected online with each stimulus rated by a minimum of 25 participants. Because the Hindy et al. ratings were collected in 2011, and the Prystauka et al. ratings in 2018/19, we recently collected new ratings for the entire set of stimuli (containing 326 stimuli from which the sentence pairs for this study were drawn) and calculated interrater reliability across the two sets of ratings. Reliability was extremely high (Pearson's r = .95). For the data described below it did not matter whether we used the original ratings, the new ratings, or the average (we report statistics based on the average). Each pair of stimuli constituted a

[1] The mean perplexity for the Wikipedia models on the validation data for Wikitext-103 was 40.6 with a standard deviation of 2.05, and for the Web models using a held out validation set of 10 million tokens of web data the mean perplexity was 64.53 with a standard deviation of 0.73.

[2] This is an open source version of the training data in the popular language model GPT-2 (Radford et al., 2019). The code with which to further explore the models and recreate the results in the present study is available at (and the models linked from): https://github.com/forrestdavis/ExperimentNorming.

minimal pair that differed only in the verb, and consequently in the manner and degree of change that the entity in object position would undergo. The stimuli were designed so that one member of the pair would entail a substantial change to that entity and the other a minimal change (the degree-of-change ratings confirmed the minimal/substantial change designation). We followed Hindy et al., (2012) in using this same paired stimulus structure, using the minimal change sentence as a baseline for assessing degree-of-change effects (both for the human ratings and for the model-derived measure which we describe next).

To assess the internal representations of the RNNs, we calculated the similarity between the hidden representation of the final word in each sentence (taken from the final hidden layer of the RNN) and the hidden representation of a baseline (see Supplemental Material A). The baseline for each sentence was the indefinite form of the relevant object (e.g., "*a mango*" given "*the chef will blend the mango*"); the model's hidden representation after "*mango*" thus corresponded to the encoding of the whole phrase (e.g., the model's hidden representation of "*mango*" following "*a*"). For complex nouns such as "*swimming pool*" both nouns were included. To quantify similarity, we took the normalized cosine similarity (the Pearson correlation coefficient)[3] of the two vectors corresponding to (i) the hidden representation after each word and (ii) the hidden representation of the baseline. We used this correlation coefficient as a measure of distance (i.e. similarity) in activation space. We predicted that RNNs with at least some knowledge of event structure (i.e. the consequences of an event for changes in the affordances/states of objects affected by that event) would have a graded degree of similarity between the baseline and the object. The baseline reflects the broadest set of affordances (i.e. the broadest prediction space: "*A mango*" licenses future washing, peeling, chopping, blending, freezing, etc) while changes in the trajectory of an object necessarily restrict possible affordances ("*blending a mango*" makes future washing, peeling, or chopping unlikely). Similarity between the baseline object and the object embedded in a particular event, then, corresponds to the extent to which that particular event restricts the affordances of the object, with lower similarity corresponding to greater restriction and higher similarity to less restriction (cf. "*the chef blends the mango*" vs. "*the chef weighs the mango*").

There are a number of different ways to probe the internal informational content of a network, including diagnostic classifiers (e.g. Giulanelli, Harding, Mohnert, et al., 2018) and *minimal description length* probing (Voita & Titov, 2020). We chose to use a more direct analog of the question that was asked of the human participants: By how much is an object *changed* after an event relative to how it was before the event? Our measure of representational similarity asked by how much the hidden unit activation changes after a linguistic description of an event relative to how it would be after a generic descriptor. Classifiers can give a measure of change if interpreted probabilistically (i.e. the change in likelihood that a pattern will be classified as belonging to a particular category), but the distance between two representations is necessarily influenced by the distances between the other representations on which the classifier has been trained. Cosine similarity is instead an absolute measure of similarity/change.

Statistical analyses of the relationship between model degree of change and human ratings were performed by calculating a difference score for each item pair ("substantial" change minus "minimal" change) and for each variable (human ratings and model degree of change, averaged across the 25 models). This allowed us to maintain the pairwise structure in the stimuli. We then computed the Pearson correlation coefficient r to quantify the relationship between the difference scores for the two variables, calculating the upper and lower 95% confidence intervals using the Monte Carlo method from Preacher (2012). For the Wikipedia models, $r = -.20$, $p = .012$ (95% CI: LL -.36, UL -.03); For the Web models, $r = -.20$, $p = .018$ (95% CI: LL -.36, UL -.03). We also used linear mixed effects models which confirmed the relationship between the network similarity measure and the human ratings (see Supplemental Materials B).

Greater similarity between the object and the baseline correlated with lower human degree-of-change ratings. In other words, the model representations seem to encode information about the magnitude of change the object is undergoing. There is significant variance left unexplained by the models internal representations, however. That is to be expected given that these models are trained only on text. The networks provide a unidimensional measure of representational change (based only on experiential knowledge of the language). The human raters, on the other hand, likely provided a multidimensional measure of such change, grounded in experiential knowledge of both linguistic *and* non-linguistic origin, with the latter spread across multiple sensorimotoric dimensions. Thus, neither of the corpora we used could encode object-state changes to the degree that humans experience them in their daily lives, but some corpora may encode state change more explicitly than others (cookery books may be a better source of object-state change information in respect of e.g. chopping, peeling, or blending, for example). Nonetheless, given the experiential limits imposed on our networks – being exposed *only* to linguistic input and that input being impoverished in respect of conveying the full (real world) range of object-state change– it is all the more remarkable that RNNs encode event-relevant structure to the extent that they do. We return to the challenges that impoverished experience presents, both for the networks and for understanding the nature of the network's internal representations, in the general discussion below.

Study 2: Propagating event participants forwards, and backwards, in time

Study 1 demonstrated that our RNNs' encoding of discourse entities was modulated by the verb preceding that entity. This modulation correlated with human ratings of the degree to which those objects, in real life settings, would be judged to change state as a consequence of the event described by the sentence. Having demonstrated this sensitivity to event-relevant content, in this second study we ask whether RNNs can propagate the appropriate content into a subsequent sentence. Specifically, a sentence that refers to a second event *following* or *preceding*, in event time, the event described in the first sentence. Consider the following

[3] The Pearson correlation is equivalent to normalized cosine similarity, making the measure invariant to the addition of a constant. Qualitatively similar results hold when using unnormalized cosine similarity (which lacks this property). The particular implementation we used was corrcoef from numpy: https://numpy.org/doc/stable/reference/generated/numpy.corrcoef.html.

The chef will chop the onion. Then, she will weigh the onion [same token, future event]
The chef will chop the onion. First, she will weigh the onion [same token, past event]

In principle, the onion that is weighed in the "Then" condition should be less similar to a generic onion (it has been chopped) than the onion that is weighed in the "First" condition, which should be more similar to a generic onion (it has not yet been chopped). We followed Solomon et al. (2015) in adding another condition:

The chef will chop the onion. Then, she will weigh another onion [different token, future event]
The chef will chop the onion. First, she will weigh another onion [different token, past event]

Solomon et al. (2015) found in an fMRI experiment that sentences with "another onion" patterned as if this other onion had not undergone any change (i.e. it was a newly instantiated generic onion). While we anticipate that "*another onion*" should be more like a generic onion than "*the onion*" after "*Then,…*". Less clear is how "*another onion*" will pattern (in respect of its similarity to "*an onion*") after "*First,…*". In both cases, we would expect some representation of the original chopped onion to propagate forwards, because the pragmatics of such constructions (manifested in their usage) suggests that the chopped onion will be referred to again in the future (otherwise it would not have been mentioned at all). Most likely, we would see greater *dissimilarity* in the "Then" condition simply because, as event time moves forward, there is probably a greater likelihood that the chopped onion will come back into (linguistic) play, in which case the RNN may increase its activation, thereby decreasing the similarity of its internal representations to "*an onion*". We return below to discussion of the network's encoding of distinct tokens.

We selected 150 two-sentence stimuli from the original set described above. All stimuli included verbs that had previously been designated by human participants as causing "substantial change" (see Study 1). We selected "substantial change" items for this study so as to better explore the effects of the "Then/First" alternation (for verbs entailing minimal change, the representation of the changed entity would be little changed from before or after the event that changed it, and using such verbs would have lacked sensitivity). As with Study 1, we calculated the similarity between the hidden representation of the final word (this time at the end of the second sentence) and the baseline for each sentence –the indefinite form of the relevant object (e.g., "*an onion*" for the examples above).

We performed a 2 (temporal adverb) x 2 (determiner) within-subjects ANOVA (every network was given every item in all 4 conditions), followed by planned comparisons of the contrast between "then…the" and "first…the". For each analysis, we treated networks as participants and report both by-network (F1) and by-item (F2) analyses. Table 1 shows the similarity values (Pearson's r) in each of the 4 conditions for both the Wikipedia and Web models.

For the Wikipedia models there was a main effect of adverb ("then" vs. "first": $F1(1,24) = 22.4$, $p < .0001$; $F2(1,149) = 43.0$, $p < .0001$) and of determiner "the" vs. "another": $F1(1,24) = 1379.6$, $p < .0001$; $F2(1,149) = 1333.5$, $p=.000$ but no interaction between the two ($F1(1,24 = 1.8$, $p = .187$; $F2(1,149) < 1$). Similarly for the Web models ("then" vs. "first": $F1(1,24) = 49.1$, $p < .0001$; $F2(1,149) = 20.1$, $p < .0001$; "the" vs. "another": $F1(1,24) = 1652.3$, $p < .0001$; $F2(1,149) = 1025.4$, $p < .0001$) although for these models there was a marginal interaction between adverb and determiner ($F1(1,24) = 7.3$, $p = .012$; $F2(1,149) = 1.0$, $p = .316$). For both sets of models, the "then…the" condition was less similar to baseline than the "first…the" condition (Wikipedia: $F1(1,24) = 22.0$, $p < .0001$; $F2(1,149) = 306.4$, $p < .0001$; Web: $F1(1,24) = 48.1$, $p < .0001$; $F2(1,149) = 86.9$, $p < .0001$).

	Wikipedia		Web	
	Then…	First…	Then…	First…
the…	.530	.535	.545	.550
another…	.700	.706	.670	.675

Table 1. Representational similarity analysis comparing hidden unit activations at the end of the second sentence to the baseline. Values are Pearson's r (higher value means more similar), and for networks (n=25) trained either on the Wikipedia or the Web corpus.

If an onion was chopped but first it was weighed, reference to the onion that was weighed engendered a representation that was more similar to a generic onion than if the onion had been chopped *and then* it had been weighed. The networks, regardless of which corpus they had been trained on, were sensitive to the temporal ordering of the events and the consequences of this ordering for the state of the onion (although most likely the distinct states of the onion are encoded in respect of the likely consequences of an event for what events can follow – see above). Equally, if an onion was chopped but then *another* onion was weighed, that onion was again more similar to a generic onion (i.e. more similar to the baseline "*an onion*") than if it was *the* onion that was weighed. We return in the general discussion for the implications of such a result for whether, or how, the RNNs can be considered to have encoded the onion that was chopped as a specific *token* onion with "*another onion*" encoded as a different token.

Perhaps surprisingly, the same "then"/"first" pattern was observed for "*another onion*" as was observed for "*the onion*". Why should "*then…another onion*" be more dissimilar to the (presumed) generic than is "*first…another onion*"? They are both new tokens, and in some sense should be identical regardless of the temporal context in which they are introduced. However, and as we shall discuss further below, the representations we are probing at the end of the second sentence are not *just* those associated with reference to the onion – they reflect the entire representational state of the network (operationalized as the second hidden layer). This state will include representational content pertaining to the onion at the end of this second sentence but also pertaining to the chef from the first sentence (we explore this further in Study 3 below), *and* the onion from the first sentence. So if there are two instances of onion – i.e. two onion tokens – the representation at the end of the second sentence will contain information about the new token (introduced by "*another onion*") *and* the original token (introduced in that first sentence, and whose state/affordances reflect having been chopped). We speculate that in the case of "*first…*" the state of the original token is "suppressed" (the affordances of a chopped onion no longer apply and are less active) meaning that the composite pattern will be more similar to a generic onion than after "*then…*" when the affordances of a chopped onion do still apply. Hence the same effect of temporal adverb on "*another onion*" as on "*the onion*".

Again, we return below to this issue of how and in what way the RNN encodes tokens.

If "another onion" puts the network into a state where it represents both this new token onion and the original token, would the network be able to distinguish between these two tokens? We believe that this may be a limitation of the networks as currently trained. It has been observed that there is a general recency bias in RNN language models (e.g., Ravfogel et al., 2019; Davis and van Schijndel, 2020). In our own testing, we have noted a recency bias for stimuli like "*The chef has a small onion and a big onion. He chopped the small onion. Then, he chopped the ...*", where rather than predicting "*big*" (as pragmatic reasoning would suggest) both the Wikipedia and Web-trained models had a greater preference for "*small*". But their preferences were modulated by training corpus: The Wikipedia networks preferred "*big*" over the pragmatically anomalous continuation "*banana*", whereas the Web networks surprisingly preferred "*banana*" over "*big*". In the real world, of course, where language meets visual experience, that experience is not subject to the same recency biases that are typical of language. For example, as our eyes move around a scene, we tend not to revisit the most recently viewed entities. And when navigating somewhere and back again, we revisit the earlier location, not the more recent location. We thus believe that there are attentional factors in our experience of the external world which essentially work against the recency biases that pervade our experience of the linguistic world. We cannot, at this time, tell whether the recency bias we find in our RNNs is due to their specific training (i.e. reflecting a general bias in the language they are exposed to) or due to an architectural limitation that could be overcome with an attention component (Bahdanau, Cho, & Bengio, 2014; Vaswani, Shazeer, Parmar, et al., 2017) operating either over the language or over a different but parallel domain of experience - c.f. the relationship between linguistic and non-linguistic domains of variation envisaged in Altmann & Mirkovic (2009). It is noteworthy that the language model GPT-2 (Radford, Wu, Child, et al., 2019), a transformer model with an attention mechanism, does not display a recency preference with these kinds of stimuli, but predicts the pragmatically expected continuations. On the other hand, it appears to fail with "*The chef has a small onion and a big onion. He chopped the small orange. Then, he chopped the ...*", where it exhibits a substantial preference for the continuation "*big*" over "*small*" (suggesting a structural preference over content). The RNNs do not do any better – they prefer the more recent "*small*" over "*big*", regardless of the corpus on which they were trained, although the web-trained networks continue to prefer "*banana*" over "*big*".

These last (informal) data, contrasting RNNs with GPT-2, highlight an issue that is central to the current series of studies: RNNs exhibit representational properties that we believe *a priori* to be necessary precursors to the behaviors we are targeting. But *representational space* is not the same as *word space*. The representational similarity analyses reported for Studies 1 and 2 operate over representational space, whereas the behaviors just described (with big and small onions) reflect operations over word space. GPT-2 exhibits the right behaviors in word space (insofar as we have started to explore them) but their correlates in representational space, at least in respect of object state affordances and trajectories through time, are relatively opaque. We return to GPT-2 in the General Discussion.

A final word, in this section, on the distinction between *representation* and *behavior*. We can think of representations in RNNs as corresponding to the regions of an abstract multi-dimensional (similarity) space that the system can move into as a function of where it has come from (c.f. our earlier description of representation in an SRN as constraints on which activation patterns can follow which other activation patterns). Behavior is what the system does when it actually *traverses* that space. Thus, we refer to network behavior not simply when, for example, describing its predictions in word space, but also when using representational similarity analyses to probe where the network is, in or after, its trajectory through that representational space. Representation and behavior are thus intimately intertwined inasmuch as tracing a trajectory – traversing the space – entails passing through different representational states.

Study 3: Propagation dynamics

In the previous studies, we investigated the degree to which RNNs encode object affordances in their representations both broadly and for specific tokens. In Study 3, we explored what effect emergent event representations in RNNs have on other participants in the event. In particular, we ask what happens to the representation of the sentential subject as the network encodes events across multiple sentences? Some representation of the subject must propagate forwards (a consequence of recurrence, and a desirable property of any model of human sentence processing), but what modulates the strength of that propagation? According to the IOH, the history of an object includes its intersections with other objects – in effect, that object becomes dynamically associated with the objects with which it has intersected (meaning that those associations are context-specific, depending on where in time and space the intersection occurred). But such associations can form only to the extent that the representation of one propagates strongly enough onto the representation of the other. If prior input has very little effect on the state of the system (i.e. it perturbs it less), the "trace" of that input will be weaker than that of an input that has greater effect on the state of the system. But similarly, if something subsequent to that input perturbs the system more, it may "mask" the impact of that earlier input. We might conceptualize this idea with the following analogy: a bigger splash will cause ripples to travel further. But the ripples due to a smaller splash may be overwhelmed by those caused by a subsequent bigger splash. In this more exploratory study, we use *entropy* to quantify the splash.

Entropy is operationalized in our RNNs as the amount of order or disorder in the predictions that the RNN makes at each point in time. If only a small number of words are predicted at the next time-step, the system is in a state of low entropy compared to one in which many words are predicted. And if many words are predicted but one is predicted by very much more than the others, then that too reflects a state of low entropy. The more constraining the context, the more the system is perturbed (the less random the activation state of the system becomes), and the lower the entropy. In Study 3 we probed the entropy at the offset of the subject+verb sequence in the first sentence of a 3-sentence sequence such as

The farmer will shear the sheep. Then, he will feed the sheep. Then, he will think about the sheep.

The entropy at the offset of "*shear*" reflects the impact of the combination of that verb with its subject on what the network will predict might come next (henceforth, although we shall be referring to the entropy at that first verb, we shall simply refer to it as "the entropy"). A verb like "*shear*" is more constraining than,

for example, "*select*", with a corresponding reduction in entropy at its offset. But while farmers might constrain the kinds of event that might be referred to subsequently, a verb like "*shear*" restricts the lexical space much more considerably. Thus, low entropy at the offset of "*shear*" most likely reflects the "ripples" of "*shear*" more than it does the ripples of "*farmer*" (although of course, it reflects the combination of the farmer and the shearing). But does the combined predictive strength of "*the farmer will shear*" (indicated by low entropy at the offset of this sequence) aid in the propagation of some representation of the farmer downstream and into the subsequent sentences, or will it hinder it? In other words, how does entropy – a proxy for the state of perturbation of the network – impact, if at all, on the extent to which the farmer is reflected in the hidden state representations at each instance of "*the sheep*"?

We selected 232 two-sentence stimuli from the original set described above with half drawn from the "substantial change" category and half from the "minimal change" category. We added the same third sentence frame to all the stimuli: "*Then, <pronoun> will think about <object from 1st sentence>*. This was in part to ensure all additional words were known to the networks and in part to ensure that any effects at the end of this 3rd sentence could not be due to variability across items at this 3rd sentence. All words contained within the entire stimulus set were "known" to all the networks (the 25 Wikipedia and the 25 Web networks). There was no difference in entropy across the minimal and substantial change verbs (means: 5.95 and 5.94 respectively, $F < 1.0$). We included both substantial and minimal change verbs so as to include a spread of degree-of-change (entropy did not correlate with degree-of-change; $r = .05$, $t < 1.0$).

Figure 1 shows the similarity of the activation pattern at each word to the activation pattern due to the baseline "*the farmer*" – it illustrates the degree to which some representation of the farmer is "contained" within the hidden state representation at each point in the sentences, and shows how a representation of that first sentential subject propagates forward from the beginning of the first sentence to the end of the last. We computed the Pearson correlation between entropy and similarity of the hidden unit activations to "*the farmer*" at the offset of each mention of "*the sheep*". There was a statistically significant and positive correlation at all three mentions. For the Wikipedia networks – Sentence 1: $r = .22$, $p = .001$ (95% CI: LL .09, UL .34); Sentence 2: $r = .15$, $p = .043$ (95% CI: LL .002, UL .26); Sentence 3: $r = .15$, $p=.021$ (95% CI: LL .02, UL .27). For the Web networks – Sentence 1: $r = .18$, $p = .006$ (95% CI: LL .05, UL .30); Sentence 2: $r = .15$, $p = .020$ (95% CI: LL .03, UL .27); Sentence 3: $r = .18$, $p=.007$ (95% CI: LL .05, UL .30). See Supplemental Material C for confirmation with linear mixed effects models. In contrast, we found no correlations between entropy and subject similarity at the pronouns, suggesting that our data are not simply a reflection of high subject similarity. We did find similar influences of the entropy when the object in the second sentence was changed from "*the sheep*" to "*another sheep*", although similarity to "*the farmer*" was significantly reduced.

The data from this study suggest that dynamics matter – that is, that the perturbation of the network, indexed here by the state of entropy after the combination of the subject and verb in the first sentence, does impact on the activation profile of representations in subsequent sentences. The lower the entropy, the less similar was the representation at "*the sheep*" to "*the farmer*". That is, the lower the entropy was, the less clearly the sentential subject propagated forwards through to both the end of that first sentence and subsequently into the following sentences. One interpretation of this result is that we are seeing the effects of a bigger splash on an earlier splash – the perturbation due to a strongly constraining verb masking the lesser perturbation due to the sentential subject. This may in part reflect the centrality of the predicate, in English, in respect of constraining the participants' roles in the sentence (i.e. in defining the "intersections" as described in the IOH). In languages such as Japanese, where the verb typically comes at the end of the sentence, we might expect to find equivalent effects at points within the sentence that are similarly constraining (i.e. at certain post-nominal particles that function as case markers; see Kamide, Altmann, & Haywood, 2003, for the behavioral manifestation of such constraints on incremental processing and prediction in Japanese).

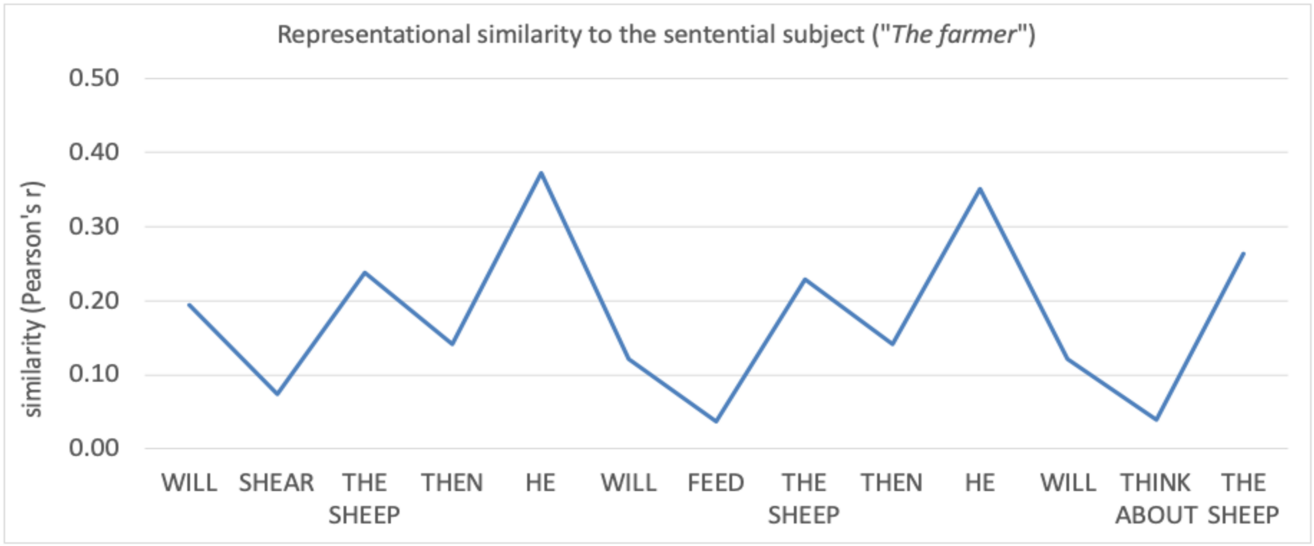

Figure 1. Similarity of the activation state of the hidden units at each word position to the activation state of those units after experiencing just the subject of the sentence ("*the farmer*")

Regardless of interpretation, the actual significance of these data is not in respect of what we might hope to know about networks, but in respect of what we *do not* know about the human brain: *We do not know what the equivalent dynamic is in the human brain*. Might recall of earlier material in a sentence similarly depend on entropy? Would participants better recall the farmer in a subsequent cued-recall task (cued with "*the sheep*") if he had selected a sheep rather than sheared a sheep? Could we use RSA in a neuroimaging task to generate a continuous measure of "representational integrity" as we did, in computational terms, for Figure 1? Here, we are equating representational similarity with representational *integrity* – the more similar the representation after "*sheep*" to the representation after "*farmer*", the greater the integrity of the representation (i.e. the less interference during propagation, or the stronger the association that formed between "*farmer*" and "*sheep*" when they first co-occurred in that first sentence). These are all questions for future studies. The point, simply, is that consideration of network dynamics lends itself naturally to consideration of brain dynamics. And a theme that will recur below is that, when it comes to such dynamics, we do not know even what the target behavior is in the human brain that we should be hoping to model.

GENERAL DISCUSSION

According to the IOH (Altmann & Ekves, 2019), a hallmark of event representation is the encoding of object-state change across time. In Study 1, we demonstrated that item-wise differences in similarity computed from the RNNs' internal representations correlated with item-wise differences in human ratings of the degree to which an object is changed by an event it participates in. Study 1 thus showed that the RNNs developed an emergent state space that is similar, at least along some limited dimensions, to the representational space encoded by human participants. Study 2 extended these state change findings to show that the representation of onion introduced in the first sentence ("*The chef chopped an onion*") propagated into the second sentence ("*Then/First, she smelled the onion*"). However, this propagation was modulated by the temporal adverb; when the onion was referenced at a point in time *after* the chopping, it was *less* similar to the representation engendered by the phrase "*an onion*" (the generic baseline) than when it was referenced at a point in time *before* the chopping.

A related hallmark of event representation – related that is, to the encoding of object-state change – is the distinction encoded in such representations of object *tokens* versus object *types*; it is not just *any* onion that is being smelled, it is the *same individual* onion (the same token) as had been chopped. The RNNs in Study 2 were sensitive to this distinction between the same onion and another onion: After the chopping, smelling "*another onion*" engendered a representation that was more similar to a generic onion than did smelling "*the onion*". The networks thus appear to distinguish between cases when reference is to the same token and cases when reference is to a different token of the same type. Nonetheless, we do not have direct access to the networks' *actual* representations (as distinct from the raw activation values across the networks' hidden units) – we are no more able to determine whether the network in fact encodes objects as tokens than we are able to determine whether a *human participant* encodes objects as tokens – we return below to why there is no such direct access, and why, nonetheless, we believe that the RNNs *did* instantiate tokens.

The behaviors observed in Studies 1 and 2 address just one aspect of an event representation; namely, object-state change. Study 3 was more exploratory, examining factors that might mediate the extent to which the object acted upon (e.g. the onion) becomes representationally associated with the object that acted upon it (the chef). Our interest here was in how properties of the verb (e.g., how *constraining* the verb was in respect of its predictive informativeness – how much it *perturbed* the system) might impact on the network's ability to propagate and re-activate those representations as appropriate. We found that the greater the perturbation at the verb, the harder it was for the representations associated with the subject to propagate forwards and make contact with the representations due to the object (the onion).

Our longer-term goal in running this study was to raise an issue not about *network* dynamics but about *brain* dynamics: Would we see the same propagation dynamics if we were to probe the equivalent in human participants? For example, perhaps the effects we observed at the final word in the 3-sentence sequence were unrelated to the content of the word "onion" in that position – i.e. unrelated to that object's history. Would this be a "good" thing, or a "bad" thing? We cannot know until equivalent analyses of the equivalent dynamic are carried out in human participants (e.g. using neuroimaging data and RSA through time – see e.g. Choi, Marslen-Wilson, Lyu, Randall, & Tyler, 2020). Perhaps they would show the same dependence on the entropy of the verb in the first sentence that we observed in our networks. Without knowing what the human equivalent dynamic is, we cannot know which is the target behavior we should hope to explain. If nothing else, our RNNs have opened the door to asking such questions and to probing human behavior in new ways that would inform the nature of the dynamical properties of the brain's encoding of the unfolding language.

Event Representation and the Sparsity of the Input

These data are by no means exhaustive – they are just a first step in our understanding of how RNNs can or might encode event representations and their corresponding dependencies, and as just mentioned, they open the door to future investigations of propagation dynamics in the human case. But while much further investigation is warranted, the current data, limited as they are, do nonetheless beg the question: What was the basis for our RNNs' abilities? Surprisingly, this is not so straightforward a question. Even knowing *what* the RRN can do is far from straightforward; on what basis do we evaluate *how* the network works? Can we even evaluate what, *in the corpus*, led to the networks' behaviors? One very significant challenge is captured by the following statistic: The word sequence "*chop the onion*" appears just once in the whole of Wikipedia. And "*weigh the onion*" appears … not at all. In fact, for the Wikipedia corpus, 94% of the 290 verb+object combinations did not appear in the corpus on which the networks were trained; for the WEB corpus this figure was 89%. The extent of this *sparsity* within the corpus poses a major challenge for understanding the causal mechanisms through which the networks acquired, encoded, and deployed the knowledge that contributed to their event-relevant performance. In the context of such sparsity, how *could* the networks, even in principle, learn that chopping an onion changes that onion by more than weighing it?

The answer to this last question is related to the question "*What do categories, as encoded in semantic memory, offer the cognitive system?*" The traditional answer is: "*generalization*". In the present context this means that it should not matter that "*chop*

the onion" is effectively absent from the corpus. What matters is that "*chop*" and "*onion*" appear separately many thousands of times and, perhaps critically, that "*onion*" frequently co-occurs with "*garlic*", "*carrot*", "*mushroom*", and other choppable things. So long as nearby semantic space encodes *something* as affording chopping, and so long as that space, or the semantic space associated with chopping[4], encodes the class of state change that constitutes being chopped (e.g. the class of change that is common across the chopping of onions, carrots, logs, text, etc), or encodes a space of consequent actions, the novel combination of chopping and onions can be interpreted. Thus, "*onion*" would, in lieu of actual experience, inherit properties of other objects in nearby semantic space. This inheritance is due to constraints on where (in state space) the system can move next as a function of were (in state space) it has come from. These constraints do not reflect simple *context-independent* co-occurrence statistics (c.f. LSA) but rather reflect accumulated experience of *context-dependent* trajectories through state space (c.f. SRNs). Hence, if sparsity in the corpus is accompanied by an appropriate category structure across the semantic space (defined through proximity in the similarity space), novel combinations of verbs and objects, or in real-world terms, of actions and participants in those actions, can be interpreted through such inheritance.

Trajectories and their Propagation through Time

Novel combinations of verbs and the discourse entities that participate in the actions denoted by those verbs constitute novel trajectories through state space. However, an unintended interpretation of such a statement is that these trajectories are *independently* realized within the networks' internal states, like veins running through the network's body, albeit across time. However, in the recurrent architecture we envisaged in Altmann & Ekves (2019), and certainly within the RNNs employed here, there are no such independently realizable trajectories (beyond some theoretical abstraction). Rather, the entire state of the representational substrate (which may or may not coincide with the entire network) is in flux; an individual trajectory is the manifestation in that substrate of information that evolves through time, distributed across the entire representational substrate both in network space (hidden unit activation space) and time (c.f. "neural manifolds", although these are generally associated with subsets of the entire neural substrate; e.g. Gallego, Perich, Naufel, et al., 2018). These are not veins that can be stripped from the network's body. And this makes a causal interpretation of the network's behavior (i.e. what internal "representations" drive those behaviors, and what from their experience drove the emergence of those representations) particularly challenging (see e.g. Tabor, Cho, & Szkudlarek, 2013, and references contained therein, for related discussion). How, for example, can we possibly know if the network has a representation corresponding to a *specific token object*? But equally, and in the scientifically-mandated absence of intuition, how can we possibly know if a *human participant* has a representation corresponding to a specific token object? What *behavior* would we expect to observe under what conditions? And imagine that our RNNs exhibited the equivalent behavior… should we interpret the RNN's behavior differently from how we interpret the human participant's? The answer to this last question is, of course, "no". Or rather, "no" is the answer to the related question "should we interpret the human participant's behavior differently from how we interpret the network's?"

The key behavior that we believe underlies our RNNs' ability to capture key aspects of event representation is the propagation and modification of object representations forwards in time (that is, forwards through the sentence – we established in Study 2 that the networks exhibited some sensitivity to the linguistic time travel afforded by temporal adverbs). This directionality matters. It is common to assume that, in the case of referential dependencies, a subsequent anaphor or referring expression refers back in time to some specific token discourse entity introduced previously. Equally, it is common to assume, in the terms of a recurrent architecture, that the current state of the network contains echoes of its past states, and that the current input can cue retrieval of information from those past states (c.f. cue-based retrieval approaches to sentence processing; e.g. Lewis, Vasishth, & Van Dyke, 2006). An alternative assumption is that in cases of anaphora or other referential dependency, the antecedents (the knights from the Davis & Schijndel (2020) study, or the chef/farmer from our own studies reported here) are propagated forward across the sentences such that the antecedent to a subsequent expression such as "*the chef*" or "*she*" is not an antecedent at all (in its literal sense), but a *concurrent* component of the network's internal representation. In discussing the likely workings of our networks, we use the concept of propagation forward in time, rather than retrieval from backwards in time, as this more accurately reflects the underlying computational mechanism (as instantiated in the LSTMs). It is not the case that, for example, a "representation" is put in a metaphorical box where it remains, static, until retrieved at some later time, or that the representation is carried forward in time on the crest of a predictive wave, remaining unchanged for the duration of the wave on which it travels (c.f. models of human memory based on cue-based retrieval, which argue that what is retrieved is reconstructed from the context at the time of retrieval and that, in essence, it is impossible to access/retrieve the same representation twice; e.g. Roediger, 2001). As representations carry forward, they change with the network as the network itself changes state dynamically through time. Whatever representation is initially activated on-the-fly changes as more of the sentence accrues. The chopping, the onion, the smelling… these each impact on the chef as each sentence unfolds word-by-word.

While it is an inherent property of recurrence in the RNN that information can propagate forward in time, how did our RNNs learn to propagate the *right* information forward (right in the sense of enabling the observed behaviors), modulating it to reflect the exigencies of the (described) event? RNNs are constrained to be forward looking – they predict upcoming input on the basis of prior input, with no access to the *right context* (i.e. the input that would come *after* the target item to be predicted). This is distinct from models such as word2vec (Mikolov et al., 2018), BERT (Devlin et al., 2018) or ELMo (Peters et al., 2018). For many instances of syntactic or sense disambiguation, the right context is completely disambiguating (cf. "*The knight killed by the dragon fell to the*

[4] According to the IOH, representations of *actions* are emergent properties of the representational system; to the extent that classes of objects change states in analogous ways, the analogy can emerge as a category across those changed states. Use of the same label, e.g. "chop", to refer to these analogous changes would encourage such emergence.

ground" vs. "*The knight killed the dragon which fell to the ground*" or "*I went to the bank to get my money*" vs. "*I went to the bank of the river*"). But without access to the right context, learning to propagate from the left, using e.g. referential dependencies to inform the resolution of ambiguities to the right, can contribute to reducing the prediction error through correctly predicting how an ambiguity should resolve, or through correctly predicting what kinds of actions might be referred to next given the new state of a propagated object. We conjecture that, to the extent that the left context can contribute to reducing error during training, the propagation of object representations as trajectories through time and object-state space is an emergent feature of forwards prediction (left-to-right predictive contingencies) in a recurrent or equivalent architecture.

Although this has still to be systematically tested (see Ettinger, 2020, for evidence suggesting that BERT lacks event knowledge; and Tran, Vusazza, & Monz, 2018, and Abnar, Dehghani, & Zuidema, 2020, for further elucidation of the role of recurrence within NLP), we did briefly explore whether the results reported here (Study 1) are unique to the RNN's architecture. Models such as word2vec, which return the same word embedding regardless of context, will not be able to model the contextual dependencies on which our data rest. But what of BERT, ELMo, or the more recent GPT-2 (Radford et al., 2019)? We in fact tested all three of these models (different pre-trained and open-sourced instantiations that differed in training set and parameters; see Supplemental Materials D) and found that each could model the data from Study 1 – that is, they had developed hidden-layer representations that, across the range of sentences used in that study, predicted human ratings of change in state. We used seven variants of BERT, each with 12 hidden layers. Treating each as a participant (i.e. for each item, averaging across all seven models – equivalent to our analytic procedure in Study 1 above), the first hidden layer was sensitive to degree-of-change (i.e. a statistically significant correlation to the human ratings; $r = -.16$, $p = .049$, 95% CI: LL -.32, UL -.001). We had just a single instantiation of GPT-2 and therefore analyzed each of its 48 hidden layers separately. Six of these were sensitive to degree-of-change (i.e. we found statistically significant correlations to the human ratings); $-.21 < r < -.17$). We note, however, that these statistical analyses of BERT and GPT-2 would not reach statistical significance if corrected for multiple comparisons (reflecting multiple correlations, at each of their 12 and 48 layers respectively). The four different instantiations of ELMo, treated as participants, were also sensitive to degree-of-change ($r = -.18$, $p = .033$, 95% CI: LL -.33, UL -.01). None of this is surprising, given our original premise that degree-of-change manifests in the language models as differences in linguistic affordances – i.e. differences in the contexts that can follow the critical event descriptions. It is noteworthy that both BERT and ELMo take into account the context following a word/sentence when developing their internal embeddings – it would be surprising if these models were *not* sensitive to rightwards contextual contingencies.

This last observation begs the question: Why invest all this (theoretical and practical) effort in RNNs rather than these more powerful and widely-used models? Our emphasis throughout this work has been on the propagation of representations, updated as they travel from left to right through a sentence or series of sentences to reflect changes afforded by the events described in those sentences. Models such as BERT and ELMo are bidirectional – they simultaneously apply left *and* right context to the processing of each word, and it is not possible to assess their performance on left-to-right word-by-word incremental changes in representation without fundamentally deviating from how they are trained. Whereas left-to-right incremental processing is a given for human speech processing, NLP models operating over text (and even over speech) have the luxury during training (and after) of not being limited to left-to-right incrementation. Study 3, for example, is beyond the reach of BERT and ELMo because, except for the very final instantiation of "*the sheep*" at the end of the third sentence, the representation of each word is given by both its left *and* right context. These are not models of incremental processing. GPT-2 *does* permit incremental representational propagation and updating. However, we observed earlier that as representations propagate forward through and across sentences, they change with each incremental step – representations are not put into a metaphorical box where they remain unchanged until retrieved some time later. GPT-2 would need to learn the dynamic that causes such continuous change – it is not built into the architecture of GPT-2 as it is in the architecture of an RNN. That is, the use of attention in GPT-2 affords the model the ability to query past time steps while ignoring intervening words (and representations). This may be a key distinguishing feature between models such as GPT-2 and recurrent architectures when applied to the task of modeling incremental left-to-right processing, language acquisition, or even human memory. And while GPT-2 has met with considerable success in respect of modeling prediction, and its neural correlates, during human sentence processing (e.g. Goldstein, Zada, Buchnik, et al., 2020; Heilbron, Armeni, Schoffelen, et al., 2020) such studies do not (yet) track the representational content that changes in lockstep with the unfolding language and that underpins those behaviors. This is a further reason to understand better the nature of the brain's own propagation dynamics (c.f. Study 3 above).

With respect to the representations that our RNNs propagated forwards in time, we cannot with any certainty claim that these were object *tokens*, although their behavior (probed in representational space using similarity) suggests that, functionally at least, they were doing something close. But how close? Elman's SRNs (Elman, 1993) operationalized tokenization as the distinction between different exemplars of the same lexical item occurring at different positions in a sentence (as in e.g. "boys who boys *chase chase* boy"). The trajectory associated with each instantiation constrained the network's prediction of which words might plausibly come next. In his examples, lexical items were grounded in an interaction between their contexts across time and the (emergent) representations activated at each point of that time. In essence, each token lexical item was distinguished from each other on the basis of its unique trajectory through the network's hidden state space (see Altmann & Ekves, 2019, for further discussion of tokenization). The onion in our examples propagated forwards from one sentence to another in a different representational form depending on the subject and verb with which it was associated in the first sentence (and, in Study 2, modulated by whether the onion being referred to in that second sentence was marked as the version after the chopping or before). That is, the onion had a trajectory across time that encoded both the specific, dynamically changing contexts in which it had occurred (c.f. episodic memory) and the different regions of semantic space associated with those contexts and its own representational affordances (c.f. semantic memory). And just as we cannot "see" in a human brain distinct representations for distinct token objects, so we cannot see them in the RNN – we are forced in both cases to infer their existence from analyses of these systems' behaviors in different contexts. We do not know whether

the RNN individuates representations as tokens that accrue attributes (with each successive experience of the token) that are bound to that "specific" token (e.g. that specific onion as unique from all others), or whether it experiences each instance of a token as unique, with each attribute modifying that instance without a commitment to all instances of the token having the same identity. It may be impossible to distinguish between these two possibilities, in networks and indeed, even in humans (for discussion of the continuity of representational existence of tokens across discontinuities in perceptual experience, see Altmann & Ekves, 2019). To the extent that the RNN encodes objects as trajectories, and to the extent that each trajectory is unique and has continuity of representational existence (through forward propagation), the manifestation of a word in a sentence is the manifestation of a token that, functionally, has a unique identity.

CONCLUSIONS

What have we learned from the studies we have reported here – that a "black box" that is relatively opaque to representational analysis can mimic human behavior (itself the behavior of a "black box")? In fact, it is only opaque to a classical analysis that assumes bounded representations that can be teased apart one from the other. It is only opaque to an analysis that assumes a combinatorial semantics predicated on discrete combinations of discrete elements. We would claim that the propagation of "representations" (in quotes to reflect their non-discrete realization within a dynamical system) within and across sentences in our RNNs is combinatorial semantics (perhaps not in the sense of mapping onto formal semantic structures, but certainly in the sense of driving, and predicting, behavior – c.f. Glenberg, 1997, and certainly in the sense of the dynamic combination of representations through time to create new representations that are more than just the conjunction of the original). Much further work is required to understand the nature of the semantic space that our networks acquired, and to understand how that space changed dynamically as each sentence we gave it unfolded through time. But the purpose of this first set of studies was to explore whether RNNs can encode even the most basic properties of event structure, and a prerequisite for that was to explore whether they could predict the same behaviors that indicate that we humans encode event structure. In demonstrating that RNNs can indeed do that (for an admittedly limited set of event-relevant behaviors), we have identified a need to further investigate *human* processing: For example, more recent testing of the Davis and Schijndel (2020) networks found that for the contexts "*Two knights were attacking a dragon*" or "*A knight and his squire were attacking a dragon"* and the continuation "*the dragon killed one of the knights",* the networks showed lower surprisal for a subsequent sentence starting *"The knight tickled by..*" when it was two knights than when it was a knight and his squire. The networks anticipated a particular structure rather than particular *content* (that is, even though the verb "*tickled*" is contextually anomalous, unlike "*killed*", the networks still preferred the participle interpretation over the main verb interpretation). In fact, Altmann and Steedman (1988; fn 5 p.202) predicted that a relative clause modifier, *regardless of content*, should indeed be preferred in a two-referent context. And yet, to our knowledge this has never been tested – a prerequisite to evaluating *the models'* performance on such cases (we also successfully modelled the influence of situational context on syntactic ambiguity resolution reported by Tyler & Marslen-Wilson, 1977, using the exact same stimuli. But again, we were able to show that the models predicted the human behavior on the basis of structural (cataphoric) dependencies across clauses, and we do not know whether the reported human behaviors were similarly based on structural cues). And with respect to Study 3, we know of no study that has explicitly considered how the representation of the subject of a sentence (or the object, or any other discourse entity) propagates forward, moment-by-moment, into successive sentences that maintain discourse cohesion. We found that the less constraining the verb in that first sentence, the greater the integrity of the representation of the subject that propagated forward into successive sentences. We interpreted this result in terms of "network perturbation" – a kind of computational salience. The propagation dynamics we observed in that study may be a fundamental property of dynamical systems, or of the brain, or of both. Recent advances in using RSA across time in neuroimaging (e.g. Choi et al., 2020) suggest that equivalent studies with human neuroimaging may be possible – allowing researchers to identify if equivalent patterns emerge in the brain, and where.

The answer, therefore, to where this leaves us, is that, at worst, consideration of the event-representational abilities of RNNs has opened up novel avenues of research into the human mind that have not, hitherto, been considered. At best, we have a computational tool whose analysis may enable us to ground basic properties of event representation in the dynamics of a computational machinery that acquires, encodes, and deploys experiential knowledge across the senses, and which most likely encodes events as the encoding of their *consequences* for how the language, or corresponding world, can unfold. Our claim in this respect is that the RNNs, once trained, are more than just a model of the language – the knowledge they encode is a product of the input *and* of the computational dynamics of the system. Those dynamics constrain the model to acquiring certain kinds of knowledge in certain kinds of ways, and they constrain the model to subsequently deploying that knowledge in particular ways. It is undoubtedly the case that these networks would, with further testing, fail more than they would succeed. But their successes thus far suggest avenues of research, on representational content and its propagation, in the computational, behavioral and neuroscientific domains that in fact render the future success or failure of these particular networks moot.

DEDICATION

Thirty years ago, Jacques Mehler asked GTMA: What have we learned about sentence processing in the past 10 years? The provocation was explicit in his prosody, a domain of language that was foremost on his mind at that time, having moved away from sentence processing research some time before (at least 10 years before, one would assume from that prosody). But Jacques' provocation was a method. It taught those of us around him to think, and to identify our passions, and to use those passions to create our science. Jacques was a mentor whose impact undoubtedly contributed to the collaboration that led to the current work. He is missed. GTMA, October 2020.

References

Abnar, S., Dehghani, M., & Zuidema, W. (2020). Transferring Inductive Biases through Knowledge Distillation. *arXiv preprint arXiv:2006.00555*.

Altmann, G., & Steedman, M. (1988). Interaction with context during human sentence processing. *Cognition*, ***30***(3), 191-238.

Altmann, G.T.M. & Ekves, Z., (2019). Events as intersecting object histories: A new theory of event representation. *Psychological Review,* **126**(6), 817-840. doi: 10.1037/rev0000154

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

Bahdanau, D., Cho, K., & Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Choi, H. S., Marslen-Wilson, W. D., Lyu, B., Randall, B., & Tyler, L. K. (2020). Decoding the Real-Time Neurobiological Properties of Incremental Semantic Interpretation. *Cerebral Cortex*. Doi: 10.1093/cercor/bhaa222

Crain, S., & Steedman, M. J. (1985). On not being led up the garden path: the use of context by the psychological parser. In D. Dowty, L. Karttunen, & A. Zwicky (Eds.), *Natural language parsing: Psychological, computational, and theoretical perspectives*. (pp. 320–358). Cambridge, UK: Cambridge University Press.

Davis, F., & van Schijndel, M. (2020). Interaction with Context During Recurrent Neural Network Sentence Processing.

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Ettinger, A. (2020). What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, **8**, 34-48.

Gallego, J.A., Perich, M.G., Naufel, S.N., Ethier, C., Solla, S.A., & Miller, L.E. (2018). Cortical population activity within a preserved neural manifold underlies multiple motor behaviors. *Nature Communications* **9**(4233). https://doi.org/10.1038/s41467-018-06560-z

Gardner, M., Grus, J., Neumann, M., Tafjord, O., Dasigi, P., Liu, N.F., Peters, M., Schmitz, M. & Zettlemoyer, L. (2018). AllenNLP: A Deep Semantic Natural Language Processing Platform. In *Proceedings of Workshop for NLP Open Source Software,* 1-6.

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

Giulianelli, M., Harding, J., Mohnert, F., Hupkes, D., & Zuidema, W. (2018). Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *arXiv preprint arXiv:1808.08079*.

Glenberg, A. M. (1997). What memory is for: Creating meaning in the service of action. *Behavioral and Brain Sciences, **20**,* 41–50. http://dx .doi.org/10.1017/S0140525X97470012

Gokaslan, A. & Cohen, V. (2019). OpenWebTextCorpus: https://skylion007.github.io/OpenWebTextCorpus

Goldstein, A., Zada, Z., Buchnik, E., Schain, M., Price, A., Aubrey, B., ... & Jansen, A. (2020). Thinking ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*. doi: 10.1101/2020.12.02.403477.

Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

Heilbron, M., Armeni, K., Schoffelen, J. M., Hagoort, P., & de Lange, F. P. (2020). A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*. doi: 10.1101/2020.12.03.410399.

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

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, ***9***(8), 1735-1780.

Kamide, Y., Altmann, G. T., & Haywood, S. L. (2003). The time-course of prediction in incremental sentence processing: Evidence from anticipatory eye movements. *Journal of Memory and language*, *49*(1), 133-156.

Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, ***2***, 4. DOI=10.3389/ neuro.06.004.2008

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, ***104***(2), 211.

Lewis, R. L., Vasishth, S., & Van Dyke, J. A. (2006). Computational principles of working memory in sentence comprehension. *Trends in Cognitive Sciences*, ***10***(10), 447-454.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. and Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lund, K., & Burgess, C. (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, ***28***(2), 203-208.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. (2016). Wikitext-103. Technical report, Salesforce.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (pp. 3111-3119).

Mirković, J., & Altmann, G. T. (2019). Unfolding meaning in context: The dynamics of conceptual similarity. *Cognition*, ***183***, 19-43.

Nakagawa, S., Johnson, P. C., & Schielzeth, H. (2017). The coefficient of determination R^2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of the Royal Society Interface*, *14*(134), 20170213.

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition*, ***203***, doi: 10.1016/j.cognition.2020.104365

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.

Preacher, K. J. (2012, November). Monte Carlo method for assessing correlations: An interactive tool for creating confidence intervals for correlation coefficients [Computer software]. Available from http://quantpsy.org/.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1**(8), 9.

Ravfogel, S., Goldberg, Y., & Linzen, T. (2019). Studying the inductive biases of RNNs with synthetic variations of natural languages. *arXiv preprint arXiv:1903.06400*.

Roediger, H. L. (2001). Psychology of reconstructive memory. *International Encyclopedia of the Social & Behavioral Sciences*, *12844-12849*.

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.

Solomon, S. H., Medaglia, J. D., & Thompson-Schill, S. L. (2019). Implementing a concept network model. *Behavior research methods*, ***51***(4), 1717-1736.

Spivey-Knowlton, M. J., Trueswell, J. C., & Tanenhaus, M. K. (1993). Context effects in syntactic ambiguity resolution: Discourse and semantic influences in parsing reduced relative clauses. *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Experimentale*, ***47***(2), 276.

Sun, Y., Wang, S., Li, Y. K., Feng, S., Tian, H., Wu, H., & Wang, H. (2020). ERNIE 2.0: A Continual Pre-Training Framework for Language Understanding. In *AAAI* (pp. 8968-8975).

Tabor, W., Cho, P. W., & Szkudlarek, E. (2013). Fractal analysis illuminates the form of connectionist structural gradualness. Topics in Cognitive Science, 5(3), 634-667.

Tran, K., Bisazza, A., & Monz, C. (2018). The importance of being recurrent for modeling hierarchical structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4731-4736.

Trueswell, J. C., & Tanenhaus, M. K. (1991). Tense, temporal context and syntactic ambiguity resolution. *Language and Cognitive Processes*, ***6***(4), 303-338.

Tyler, L. K., & Marslen-Wilson, W. D. (1977). The on-line effects of semantic context on syntactic processing. *Journal of Verbal Learning and Verbal Behavior*, ***16***(6), 683-692.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems* (pp. 5998-6008).

Voita, E., & Titov, I. (2020). Information-Theoretic Probing with Minimum Description Length. *arXiv preprint arXiv:2003.12298*.

Warstadt, A., Zhang, Y., Li, H.S., Liu, H. and Bowman, S.R. (2020). Learning which features matter: Roberta acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing,* 217-235.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., Cistac, P., Funtowicz, M., Davison, J., Shleifer, S. and Louf, R. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations,* 38-45.

Yee, E., & Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin & Review*, ***23***(4), 1015-1027.

Yee, E., Jones, M. N., & McRae, K. (2018). Semantic memory. *Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience*, ***3***, 1-38.

SUPPLEMENTAL MATERIALS

A. Similarity Analyses

For Study 1, we correlated human ratings and a measure of model similarity. The human ratings were gathered from existing works (Hindy et al., 2012; Prystauka, et al., in preparation). We additionally conducted another rating study to confirm the consistency of these ratings (detailed in the main text). Participants were given sentences drawn from pairs of events, which either described a minimal change event (e.g., "*the chef will weigh the mango*") or a substantial change (e.g., "*the chef will blend the mango*"). They were asked to rate the degree to which the object (in this case mango) changed. This resulted in low scores for the minimal change stimuli (e.g., 1.83 for weighing the mango) and large scores for the substantial change stimuli (e.g., 5.92 for blending the mango). We used the minimal change stimulus from a pair as a baseline, deriving a difference score that accounts for the pairwise structure (e.g., the pair of weighing and blending a mango would have a difference score of 4.09). For the models, we also generated a difference score accounting for this pairwise structure: For each sentence we had a baseline (the indefinite form of the relevant object; e.g., "*a mango*"). We gathered the internal representation (the activation pattern at the final hidden layer of the LSTM) for this indefinite baseline and calculated its similarity (normalized cosine similarity) to the internal representation at the end of the stimulus (again, the activation pattern at the final hidden layer of the LSTM, after processing e.g., "*the chef will weigh the mango*"). The similarity should be greater for the minimal change events (e.g., for "*the chef will weigh the mango*" the similarity to the baseline was 0.60) than the substantial change events (e.g., for "*the chef will blend the mango*" the similarity to the baseline was 0.56). Objects may have different similarities, so to enforce the pairwise structure of the stimuli we took the difference in similarity (here 0.04). Thus, for each pair we had a human change of state score and a model change of state score. We then found the correlation between the human and model scores across the pairs of stimuli.

B. Study 1 linear mixed effects models

We used a linear mixed effects model in R (R Core Team, 2020) to confirm the patterns from Study 1. Similarity between model representations at the object and the baseline (ObjSim) was included as a fixed effect factor predicting human degree-of-change ratings. Random by-item and by-model effects were included with ObjSim by-model random slopes. We simplified the model only if it failed to converge. Across the set of human degree-of-change ratings, we found significant main effects of ObjSim for both sets of models. For the Wikipedia models: $\beta = -2.11$, SE = 0.22, $z = -9.41$, $p < 0.001$. For the Web models: $\beta = -1.69$, SE = 0.45, $z = -3.75$, $p < 0.001$. To estimate the amount of variance explained by these models we used the method detailed in Nakagawa et al. (2017), using the implementation in the R package performance (https://github.com/easystats/performance) which returns a conditional R^2 value giving the amount of variance explained by both the fixed and random effects and a marginal R^2 value giving the amount of variance explained by just the fixed effects. For the Wikipedia models, the conditional R^2 was 0.20 and the marginal R^2 was 0.03. For the Web models, the conditional R^2 was 0.10 and the marginal R^2 was 0.01. These estimates of variance explained are lower than those calculated from Pearson's r ($r^2 = .04$ for both the Wikipedia and Web models).

C. Study 3 linear mixed effects models

We used linear mixed effects models as in Study 1 to confirm the patterns observed in Study 3. Verb entropy was included as a fixed effect predicting degree of subject-object similarity across the 3 sentences. Random by-item and by-model effects were included with verb entropy by-model random slopes. We simplified the model only if it failed to converge. Across the three sentences we found a significant main effect of verb entropy for both the Wikipedia models and the Web models. The one exception was that the Wikipedia models found no effect at the object in Sentence 2. Wikipedia models: First object $\beta = 0.01$, SE = 0.001, $z = 6.80$, $p < 0.001$; second object ($\beta = 0.003$, SE = 0.002, $z = 1.56$, $p = 0.12$), third object ($\beta = 0.004$, SE = 0.002, $z = 2.08$, $p = 0.04$). Web models: First object $\beta = 0.01$, SE = 0.001, $z = 2.62$, $p = 0.009$); second object $\beta = 0.01$, SE = 0.01, $z = 2.15$, $p = 0.03$; third object $\beta = 0.01$, SE = 0.01, $z = 3.34$, $p < 0.001$. The conditional R^2 ranged from 0.70-0.79 while the marginal R^2 value was quite small at 0.01. The equivalent estimates of variance explained from the Pearson's correlations ranged from .02 to .05.

D. Testing additional computational models

Training instantiations of large scale models in NLP is extremely computationally costly (especially with regards to BERT and GPT-2 where large numbers of dedicated GPUs, and even TPUs, are used for training). We therefore used existing models. GPT-2 XL and BERT base (uncased) were used via Hugging Face's API (Wolf, Chaumond, Debut, et al., 2020). We additionally made use of 6 RoBERTa models from Warstadt, Zhang, Li, et al. (2020). RoBERTa is an optimized version of BERT which included some hyperparameter tweaks, more data, and removed the next-sentence prediction objective in BERT (Liu, Ott, Goyal, et al., 2019). These tweaks improved on BERT's performance, although the overall architecture is similar (same number of layers, use of attention, etc.). Warstadt et al. (2020) trained RoBERTa models on varied amounts of the data from BERT. We made use of three of their models trained on 100M tokens and three trained on 1B tokens via HuggingFace (https://github.com/nyu-mll/msgs). Finally, we used the four ELMo English models provided by AllenNLP (Gardner, Grus, Neumann, et al., 2018): a small, medium, and original model trained on the same 1B tokens (they differ in number of parameters, with the original as specified in Peters et al., 2018), and a large model trained on 5.5B tokens. We caution explicit generalizations from the results of these models. Given our compute limitations, we are unable to tease apart the conflicting influences of number of parameters and data size.