TOPICS
TOPICS IN COGNITIVE SCIENCE

Topics in Cognitive Science 9 (2017) 51–68

ISSN:1756-8757 print / 1756-8765 online
DOI: 10.1111/tops.12246

This article is part of the topic "Best of Papers from the 2016 International Conference on Cognitive Modeling," David Reitter and Frank E. Ritter (Topic Editors). For a full listing of topic papers, see http://onlinelibrary.wiley.com/doi/10.1111/tops.2017.9.issue-1/issuetoc.

Encoding and Accessing Linguistic Representations in a Dynamically Structured Holographic Memory System

Dan Parker, Daniel Lantz

Department of English, Linguistics Program, College of William & Mary

Received 30 September 2016; received in revised form 9 November 2016; accepted 15 November 2016

Abstract

This paper presents a computational model that integrates a dynamically structured holographic memory system into the ACT-R cognitive architecture to explain how linguistic representations are encoded and accessed in memory. ACT-R currently serves as the most precise expression of the moment-by-moment working memory retrievals that support sentence comprehension. The ACT-R model of sentence comprehension is able to capture a range of linguistic phenomena, but there are cases where the model makes the wrong predictions, such as the over-prediction of retrieval interference effects during sentence comprehension. Here, we investigate one such case involving the processing of sentences with negative polarity items (NPIs) and consider how a dynamically structured holographic memory system might provide a cognitively plausible and principled explanation of some previously unexplained effects. Specifically, we show that by replacing ACT-R's declarative memory with a dynamically structured memory, we can explain a wider range of behavioral data involving reading times and judgments of grammaticality. We show that our integrated model provides a better fit to human error rates and response latencies than the original ACT-R model. These results provide proof-of-concept for the unification of two independent computational cognitive frameworks.

Keywords: Language processing; Negative polarity; Memory; Binding; Holographic reduced representations; ACT-R

Correspondence should be sent to Dan Parker, Department of English, Linguistics Program, College of William & Mary, Williamsburg, VA 23187. E-mail: dparker@wm.edu

1. Introduction

A hallmark of human cognition is the ability to encode, access, and manipulate compositional structures (Anderson, 1983; Fodor, 2001; Newell, 1990). A prominent example involves language processing. For instance, successful language comprehension requires the ability to relate words and phrases that can be separated by a potentially large amount of material, forming so-called linguistic dependencies. For instance, in Example 1a, the verb *discussed* must be related to its subject, *the candidates*, and in Example 1b, the anaphor *themselves* must be related to its referent, *the girls at the boarding school*. In order to construct linguistic dependencies like these, comprehenders rely on mental mechanisms for encoding and accessing linguistic structure in working memory. However, it remains an open question as to how these mechanisms are neuro-computationally instantiated.

Example 1
a. The candidates, in the face of public scrutiny, discussed the nation's economy.
b. The girls at the boarding school told stories about themselves.

One model that has received much attention in the psycholinguistics literature is the activation-based model of sentence processing developed by Lewis and Vasishth (2005; henceforth, LV05). The LV05 model characterizes the moment-by-moment working memory retrievals that support sentence comprehension, realized in the Adaptive Control of Thought—Rational (ACT-R) architecture (Anderson et al., 2004). In this model, the task of comprehending a sentence is construed as a series of associative, cue-based memory retrievals, subject to fluctuating activations and similarity-based interference. The main claim of the model is that a single, cue-based retrieval mechanism is used to access linguistic information in memory, and that this mechanism is engaged for the range of linguistic dependencies encountered in natural language, including those in Example 1. The model is considered to be the most precise expression of the cue-based memory retrieval theory, and it is frequently used to investigate the timing and accuracy of memory retrieval in sentence comprehension.

Previous work has shown that the LV05 ACT-R model of sentence processing achieves good quantitative fits to behavioral data. For instance, an initial success of the LV05 model was that it captured interference effects observed in the processing of linguistic dependencies, such as those involving negative polarity items (NPIs) (Vasishth, Brüssow, Lewis, & Drenhaus, 2008). NPIs are words like *ever* or *any*, which are generally acceptable only in sentences that contain a negative-like word in a syntactically higher position, such as ***No bills*** *that the senators supported will* ***ever*** *become law*. Several studies have shown, however, that syntactically irrelevant negative distractors can intrude on NPI licensing, in sentences like *The bills that* ***no senators*** *supported will* ***ever*** *become law* (e.g., Drenhaus, Saddy, & Frisch, 2005; Parker & Phillips, 2016; Vasishth et al., 2008; Xiang, Dillon, & Phillips, 2009).[1] This effect reflects a kind of similarity-based interference that manifests in human behavior as decreased accuracy in judgments of grammaticality and decreased reading time disruptions for sentences with a negative distractor, relative to sentences that lack negation.

Vasishth et al. (2008) argued that interference effects in NPI licensing are a natural consequence of the error-prone memory retrieval mechanisms embodied in ACT-R. Under this view, encountering an NPI triggers a memory retrieval for a negative licensor from the set of previously encountered items. Interference arises when retrieval is misled by the lure of a negative item in a syntactically irrelevant position. This effect can give rise to an "illusion of acceptability," where comprehenders are fooled into thinking that an ill-formed sentence is actually acceptable (Phillips, Wagers, & Lau, 2011). An important prediction of the retrieval-based account of NPI interference is that interference effects should generalize across syntactic and semantic environments, since the effect is attributed to error-prone retrieval mechanisms that are engaged whenever an NPI is encountered.[2]

The LV05 model provides good quantitative fits to previous behavioral data, but there are cases where the model makes the wrong predictions. For instance, Parker and Phillips (2016) showed that NPI interference effects can be reliably switched on and off, depending on when the NPI is encountered in the sentence. Parker and Phillips (2016) tested sentences such as *The journalist that no editors recommended (ever) thought that the readers would (ever) understand the complicated situation*, where the NPI *ever* appeared either early, in a main clause position, or later, in an embedded clause position. Interference was observed when the NPI appeared in the main clause, replicating previous findings, but the effect disappeared when the NPI appeared later in the embedded clause. These findings are unexpected under the ACT-R account, since the model predicts that NPI interference effects should generalize across environments.

Parker and Phillips (2016) argued that the contrasting profiles observed for NPIs reflect untested assumptions about how sentences are encoded in memory. The LV05 ACT-R account argues that interference effects are the product of error-prone memory retrieval processes, with the additional assumption that the encoding of the sentence remains fixed over time. However, the finding that NPI interference effects can be switched on/off depending on when the encoding is accessed for NPI licensing suggests that the encoding is not fixed, as previously assumed, but rather changes over time, such that the internal items become opaque as candidates for causing interference as the parse unfolds.

1.1. The present study

This paper presents a computational model that integrates a holographic memory system (e.g., Plate, 2003) into the ACT-R framework to capture the contrasting NPI profiles. Holographic memory systems assume that the atomic components of a compositional structure are dynamically bound together at various points throughout processing to create a single, integrated encoding that feeds interpretation. If the format of the encoding changes with the passage of time, as assumed in holographic memory systems, we might expect different behaviors at different points in time depending on when the encoding of the licensing context is accessed. Thus, a key prediction of our model is that NPI interference effects should be selective, depending on when the encoding is accessed. Modeling results show good quantitative fits to the behavioral data from Parker and Phillips (2016),

providing proof-of-concept for the unification of two computational cognitive frameworks.

Previous work in cognitive science has argued that vector symbolic architectures, including holographic memory systems, might play an important role in describing human linguistic behavior (e.g., see "Open Peer Commentary" in Van der Velde & de Kamps, 2006). The research reported in this paper unites this work with recent efforts in cognitive psychology to integrate holographic memory into the ACT-R framework. For instance, Rutledge-Taylor, Kelly, West, and Pyke (2014) and Kelly, Kwock, and West (2015) have shown that a holographic declarative memory system, similar to the one proposed here, can be integrated into ACT-R to capture decision-making tasks, the fan effect, and delayed learning. Our model demonstrates that a unified framework can also capture specialized cognitive abilities involving language comprehension.

2. The ACT-R model of sentence processing

ACT-R is a cognitive architecture based on independently motivated principles of memory and general cognition, and it has been used to study a wide range of cognitive phenomena involving memory access and retrieval, attention, executive control, and learning (Anderson et al., 2004). The LV05 ACT-R model applies the cognitive principles embodied in the ACT-R architecture to the task of sentence processing.

In the LV05 ACT-R model, linguistic items are encoded as "chunks" in a content-addressable memory, and the syntactic representation of a sentence arises as the consequence of pointers that index the hierarchical relations between chunks. Chunks are encoded as bundles of feature-value pairs, inspired by the attribute-value matrices described in head-driven phrase structure grammars (Pollard & Sag, 1994). Features include lexical content (e.g., morpho-syntactic and semantic features), syntactic information (e.g., category, case), and local hierarchical relations (e.g., sister, parent). Values for features include symbols (e.g., ±singular, ±animate) or pointers to other chunks (e.g., NP_1, VP_2).

Linguistic dependencies, such as the relation between an NPI and its licensor, are formed using a domain-general, cue-based retrieval mechanism that accesses all task-relevant chunks in parallel to locate the left part of the dependency (the target/licensor) using a set of retrieval cues. Retrieval cues are derived from the current word, the linguistic context, and grammatical knowledge, and correspond to a subset of the features of the target (Lewis, Vasishth, & Van Dyke, 2006). Chunks are differentially activated based on their match to the retrieval cues, and the probability of retrieving a chunk is proportional to the chunk's overall activation at the time of retrieval, modulated by decay and similarity-based interference from other items that match the retrieval cues.

The activation of a chunked item i (A_i) is defined as in Eq. 1.[3] Eq. 1 makes explicit four fundamental principles that are known to impact memory dynamics: (a) an item's resting, baseline activation B_i, (b) the match between the item and each of the j retrieval cues in the retrieval probe S_{ji}, (c) the penalty for partial matches PM between the cues of the retrieval probe and the item's feature values, and (d) stochastic noise.

$$A_i = B_i + \sum_{j=1}^{m} W_j S_{ji} - \sum_{k=1}^{p} PM_{ki} + \epsilon \tag{1}$$

The first term of Eq. 1 describes the baseline activation of chunk i, which is calculated according to Eq. 2. Eq. 2 describes the usage history of chunk i as the summation of all n successful retrievals of i, where t_j is the time since the jth successful retrieval of i, to the power of the negated decay parameter d. The output is passed through a logarithmic transformation to approximate the log odds that the chunk will be needed at the point of retrieval, given its usage history. After a chunk has been retrieved, the chunk receives an activation boost, followed by decay.

$$B_i = \ln\left(\sum_{j=1}^{n} t_j^{-d}\right) \tag{2}$$

The second term of Eq. 1 reflects the degree of match between chunk i and the retrieval cues. W is the weight associated with each retrieval cue j, which defaults to the total amount of goal activation G available, divided by the number of cues (i.e., G/j). Weights are typically assumed to be equal across all cues. The degree of match between chunk i and the retrieval cues is the sum of the (weighted) associative boost for each retrieval cue S_j that matches a feature value of chunk i. The associative boost that a cue contributes to a matching chunk is reduced as a function of the "fan" of that cue, that is, the number of chunks in memory that match the cue (Anders & Reder, 1999; Anderson, 1974), according to Eq. 3.

$$S_{ji} = S - \ln(\text{fan}_j) \tag{3}$$

The third term of Eq. 1 reflects the penalty for a partial match between the cues of the retrieval probe and the feature values of chunk i. Partial matching makes it possible to retrieve a chunk that matches only some of the cues, creating the opportunity for retrieval interference (Anderson et al., 2004; Anderson & Matessa, 1997). Partial matching is calculated as the matching summation over the k feature values of the retrieval cues. P is a match scale, and M_{ki} reflects the similarity between the retrieval cue value k and the value of the corresponding feature of chunk i, expressed by maximum similarity and maximum difference.

Lastly, stochastic noise is added to the activation level of chunk i, generated from a logistic distribution with a mean of 0, controlled by the noise parameter s, which is related to the variance of the distribution, according to Eqs. 4 and 5.

$$\epsilon \sim \text{logistic}(0, \sigma^2) \tag{4}$$

$$\sigma^2 = \frac{\pi^2}{3} s^2 \tag{5}$$

Activation A_i determines the probability of retrieving a chunk, according to Eq. 6. The probability of retrieving chunk i is a logistic function of its activation with gain $1/s$ and threshold τ. Chunks with higher activation are more likely to be retrieved.

$$P(\text{recall}) = \frac{1}{1 + e^{(-A_i - \tau)/s}} \tag{6}$$

Activation A_i also determines the retrieval latency T_i of a chunk, according to Eq. 7. F is a scaling factor that sets model predictions on an appropriate time scale. Chunks with a higher activation value have a faster retrieval latency.

$$T_i = \text{Fe}_i^{-A_i} \tag{7}$$

Based on Eqs. 6 and 7, retrieval can be viewed as the outcome of a "race": given multiple items in memory, retrieval mechanisms recover the item that would lead to the fastest latency, determined on the basis of activation values, according to Eqs. 6 and 7.

3. Predictions of the ACT-R model

The LV05 ACT-R model predicts that retrieval for linguistic dependency formation should be susceptible to interference from non-target or syntactically irrelevant items that overlap in features with the retrieval cues ("partial match interference"). This prediction is based on the assumptions that retrieval queries all chunks in parallel, and that a partial match between the retrieval cues and a chunk can result in erroneous retrieval of that chunk (see Eq. 1). Many studies have shown that this prediction is borne out for a range of dependencies, including subject-verb agreement (Dillon, Mishler, Sloggett, & Phillips, 2013; Tanner, Nicol, & Brehm, 2014; Wagers, Lau, & Phillips, 2009), anaphora (Parker, Lago, & Phillips, 2015; Parker & Phillips, unpublished data), case licensing (Sloggett, 2013), and ellipsis (Martin, Nieuwland, & Carreiras, 2012, 2014).

Vasishth et al. (2008) used the LV05 model to simulate retrieval interference effects in the processing of NPIs. As noted in our Introduction, NPIs are words like *ever*, *any*, or *yet*, which can be licensed by a negative-like word in a syntactically higher position. The NPI *ever* in Example 2a is licensed because it appears in the scope of the negative phrase *no students*. When negation is absent, as in Example 2b, or is in a syntactically irrelevant position, as in Example 2c, the NPI is not licensed.

Example 2
a. *No students* have *ever* passed the test.
b. The students have *ever* passed the test.
c. The students that *no teachers* liked *ever* passed the test.

Many studies have shown that NPI licensing is highly susceptible to interference in sentences like (2c), due to the lure of the negative distractor, for example, *no teachers*, that is in a non-target, syntactically irrelevant position for the purpose of NPI licensing (e.g., Drenhaus et al., 2005; Parker & Phillips, 2016; Vasishth et al., 2008; Xiang et al., 2009). This effect manifests as decreased accuracy in judgment tasks and decreased reading time disruptions for sentences with a syntactically irrelevant negative distractor, like (2c), relative to sentences that lack negation, like (2b).

Vasishth et al. (2008) argued that NPI interference effects are a natural consequence of the error-prone retrieval mechanisms embodied in ACT-R. Under this account, NPI licensing is treated as a direct, item-to-item dependency in which a negative licensor is retrieved from memory using syntactic and semantic cues, for example, [+scope], [+negative]. In (2a), retrieval finds an item that matches both cues. In (2b), retrieval fails to find a match to either cue. In (2c), retrieval finds a partially matched item, that is, a semantically appropriate item in a non-target, syntactically irrelevant position. The activation boost from this partial match, combined with stochastic noise, can cause the negative distractor to be erroneously retrieved, creating the illusion that the NPI is licensed. Vasishth et al. (2008) showed using computational simulations that Eqs. 1–7 can achieve good quantitative fits to human reading times and judgments of grammaticality.

4. Challenges for the ACT-R model

The LV05 ACT-R model predicts that interference during NPI licensing should generalize across syntactic and semantic environments, since the effect is attributed to error-prone retrieval mechanisms that are engaged whenever an NPI is encountered. However, this prediction is not borne out. Parker and Phillips (2016) showed using self-paced reading times and speeded acceptability judgments that interference effects for NPIs can be reliably switched on/off, depending on when the NPI is encountered in the sentence.[4] They manipulated the position of the NPI relative to the potential licensors in sentences like Example 3. Self-paced reading times and speeded acceptability judgments revealed converging findings. Interference was consistently observed when the NPI appeared early in the sentence, that is, in the main clause, replicating previous findings, but not when it appeared later in the sentence, that is, in the embedded clause. Parker and Phillips replicated this effect across three sets of experiments (participant sample sizes ranged from 18 to 30 depending on the task).

Example 3
The journalists that no editors recommended (ever) thought that readers would (ever) understand the complicated situation.

These findings suggest that NPI interference effects cannot simply be due to noisy retrieval mechanisms that are engaged whenever an NPI is encountered, as assumed in the LV05 ACT-R model. Furthermore, the effects cannot reflect decay or faulty encoding of the licensing context, since that would predict difficulty in the grammatical conditions, contrary to fact.

Existing accounts of NPI interference effects, such as those proposed by Vasishth et al. (2008) and Xiang et al. (2009), have emphasized that NPI licensing is a function of the licensing conditions on NPIs and the access mechanisms. An additional assumption embodied in ACT-R is that the encoding of items encountered previously in the sentence remains fixed as the parse unfolds. The finding that NPI interference effects can be reliably switched on/off suggests that some component of this licensing

function does not remain constant during parsing. The findings reported by Parker and Phillips (2016) show that NPI interference effects can be switched on/off depending on when the encoding of the licensing context is probed for NPI licensing. These findings point to the status of the encoding as the source of the contrasting profiles, rather than faulty licensing conditions or faulty retrieval mechanisms, as assumed in the LV05 ACT-R model.

Parker and Phillips (2016) argued that the contrasting profiles observed for NPIs reflect changes over time in the memory encoding of the emerging compositional-semantic representations that support NPI licensing. ACT-R assumes that the encoding of previously encountered items remains fixed as the parse unfolds. However, the finding that interference effects can be switched on/off depending on when the encoding of the licensing context is accessed suggests that the encoding is not fixed, but rather changes over time: At one moment, semantic licensing features, such as negation, can be evaluated independently of their position in the sentence structure, creating the opportunity for partial match interference; but then, at a later point in time, those same features are no longer independently evaluable, preventing partial match interference. In short, it appears as though syntactically irrelevant but semantically appropriate licensors become opaque as candidates for causing interference as the parse unfolds (see Parker & Phillips, 2016, for discussion). In the next section, we discuss how such effects are predicted in a dynamically structured holographic memory system.

5. Multiple-stage encoding schemes

The LV05 ACT-R model assumes that the encoding of a sentence remains fixed over time. However, this is not a widespread assumption. Many cognitive models, including the entire class of vector symbolic architectures (VSAs), for example, tensor product models (Smolensky, 1990), holographic memory (Plate, 2003), binary spatter codes (Kanerva, 1994), assume that there is a qualitative shift over time in the format of an encoding in memory.

An implicit assumption of VSAs is that compositional structures are encoded in multiple stages. VSAs make a distinction between "atomic," localist representations, in which individual feature values are explicitly represented and independently evaluable versus "complex," distributed representations of feature values for an object that are constructed from atomic representations via some sort of binding operation, for example, convolution, addition, permutation, etc. This binding operation integrates the local atomic features into a complex whole, creating a new representation that is completely dissimilar to any of its bound features. In this format, the atomic features are no longer independently evaluable, and the bound representation must exhibit an "all-or-none" match to the cues of the retrieval probe to be recovered from memory, preventing the possibility for partial matching.[5] This idea of "recoding" is based on Miller's (1956) principle of chunking, which provides a central explanation for how human memory works.

5.1. Proposal

We argue that the two encoding stages described in VSAs, that is, localist versus distributed representations, can be mapped to distinct cognitive processing stages as a principled explanation of the contrasting profiles observed for NPI licensing. Parker and Phillips (2016) suggested that the encoding of a sentence is built in two stages. In the first stage, the parser constructs a localist representation in which the atomic features of the sentence are evaluable independently from their position in the structure, creating the opportunity for partial match interference (as assumed in the LV05 model). At a later stage, those same features are bound together to form a distributed representation that interfaces with the interpretive system. In this stage, the individual features are no longer independently evaluable, preventing partial match interference.

For instance, when processing sentences like those in Example 3, the parser may bind the semantic features, such as the embedded negation, to their position in structure, creating a new composite representation. If the NPI is introduced prior to binding, such as in the main clause position, then the atomic features of the representation may still be independently evaluable, leading to partial match interference. However, if the NPI appears after binding has happened, such as in the embedded clause position, then the atomic features are no longer evaluable independently of their position in the composite representation, preventing partial matching.

Previously, VSAs have not assumed that distinct cognitive processing stages are associated with the two representational states. However, if the format of the encoding changes over time, as assumed in VSAs, then we might expect different behaviors at different points in time, depending on when the encoding is accessed. We discuss the details of this proposal in the next section, and show how it can be implemented to capture the contrasting NPI profiles.

5.2. Encoding linguistic structure in multiple stages

In VSAs, the feature values of a compositional sentence representation can be encoded as high-dimensional vectors that are recursively bound together by compressing their outer product into a single vector. For instance, in a tensor-product scheme (e.g., Smolensky, 1990), features are bound together in memory by taking the outer product of the vector representations of the features, as shown in Example 4.

Example 4

a. Feature vectors: [+scope] = [123]; [+negation] = [abc]

b. Tensor-product feature binding: $\begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix} \otimes \begin{pmatrix} a \\ b \\ c \end{pmatrix} = \begin{pmatrix} 1a & 1b & 1c \\ 2a & 2b & 2c \\ 3a & 3b & 3c \end{pmatrix}$

However, the size of the data structure grows exponentially with the number features encoded, which may be undesirable given the stringent limits on the amount of

information that can concurrently occupy working memory (Cowan, 2001). Plate (2003) proposed a solution using holographic reduced representations (HRRs), which relies on circular convolution to bind features together, according to Eq. 8.[6] With this method, the size of the data structure does not grow as more features are added, since the circular convolution of two n-dimensional vectors, using modulo subscripts, produces a vector with dimensionality n.

$$t_j = \sum_{k=0}^{n=1} c_k x_{j-k} \qquad \begin{array}{c}\text{for } j=0 \text{ to } n-1 \\ \text{(subscripts are modulo-n)}\end{array} \tag{8}$$

Fig. 1 shows circular convolution as the "reduced" outer product t of the feature vectors c and x, corresponding here to the linguistic features [+scope] and [+negation] for $n = 3$. Convolution is calculated as the summation of the outer product values along the paths of the arrows. In the uncompressed form (Encoding stage 1), individual features c and x are independently evaluable, making the representation susceptible to partial matching. In the "reduced" form (Encoding stage 2), the individual features c and x are no longer independently evaluable, preventing partial matching. In this state, the representation must be recovered holistically, that is, with an all-or-none match to the cues of the retrieval probe. In holographic memory, similarity between the retrieval probe p and a memory m is measured by their normalized dot product, that is, cosine similarity, according to Eq. 9.

$$\mathrm{sim}(p,m) = \frac{p \cdot m}{||\ p\ ||||\ m\ ||} = \frac{\sum_{i=0}^{n-1} p_i m_i}{\sqrt{\sum_{i=0}^{n-1} p_i^2}\sqrt{\sum_{i=0}^{n-1} m_i^2}} \tag{9}$$

One concern is that encoding n-dimensional bindings using circular convolution can be slow, since convolution calculates the sum of products, e.g., convolution with modulo subscripts takes $O(n^2)$ time, where n is the size of the data structure. Processing can be sped up by performing convolution in the frequency domain with the Fast Fourier Transform (FFT), which involves element-wise multiplication, as shown in Eq. 10, where $f()$ represents FFT. This process implements circular convolution in $O(n \log n)$ time, again where n is the size of the data structure.

$$[+\text{scope}]_C \circledast [+\text{negation}]_x = f'(f(c) \odot f(x)) \tag{10}$$

The most important property of HRRs, for present purposes, is that the encoding changes with the passage of time, such that the internal items become opaque for partial matching. This property can provide a principled explanation for the contrasting profiles observed for NPIs. If the format of the encoding changes over time, as assumed in a holographic memory system, then we should see different behaviors at different points in time, depending on when the encoding is accessed for NPI licensing.

In the next section, we show how a holographic memory system can be integrated into the LV05 ACT-R model to simulate human reading times and judgments of grammaticality.

$$\mathbf{Binding}_t = [+\mathbf{scope}]_c \circledast [+\mathbf{negation}]_x$$

$$\begin{aligned} t_0 &= c_0x_0 + c_2x_1 + c_1x_2 \\ t_1 &= c_1x_0 + c_0x_1 + c_2x_2 \\ t_2 &= c_2x_0 + c_1x_1 + c_0x_2 \end{aligned}$$

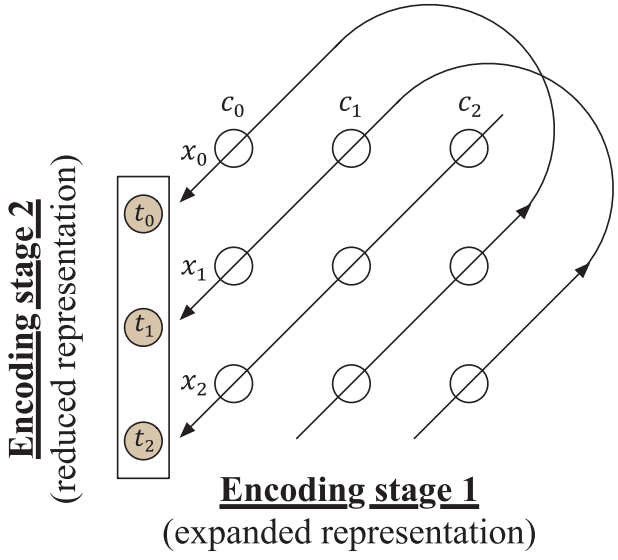

Fig. 1. Circular convolution represented as the compressed outer product t of the feature vectors c and x. Adapted from Plate (2003).

6. Integrating HRRs into ACT-R

To implement our proposal, a new memory module for the LV05 ACT-R model was developed using HRRs, replacing traditional ACT-R chunks with holographic vectors. Holographic vectors retain the same expressive power of the chunks used in the LV05 model, but allow for dynamic changes in the format of the encoding. To integrate HRRs into ACT-R, we implemented a modified version of the LV05 ACT-R model of sentence processing, using code originally developed by Badecker and Lewis (2007).[7] We made the following changes to the model. First, linguistic feature-value specifications and retrieval cues were encoded as vectors (one dimensional arrays) of n numbers, randomly sampled from a normal distribution. For our simulations, n = 10,000. In this format, different feature-value specifications and the corresponding retrieval cues are represented by different array patterns.

In Encoding stage 1 (expanded representation), feature-value pairs and retrieval cues are defined as bundles of independent vectors, corresponding to the linguistic chunks assumed in the LV05 ACT-R model. In this state, the individual features of a chunk are independently evaluable at retrieval and hence susceptible to partial matching. In Encoding stage 2 (reduced representation), convolution is used to bind the feature-value vectors

within a chunk, according to Eq. 10. In this stage, a chunk represents a single, integrated composite encoding that must exhibit an all-or-none match to the cues of the retrieval probe to be recovered, that is, partial matching is not possible. Retrieval probe vectors are constructed in the same fashion. Thus, successful retrieval of a chunk necessitates an enriched control structure to ensure that the parser matches the format of the retrieval probe to the format of the current encoding state, such that the match to the retrieval cues is evaluated in an all-or-none fashion, that is, without partial matching. For present purposes, we assumed that the transition to encoding stage 2 was triggered upon encountering the main clause verb of a sentence during comprehension. According to Parker and Phillips (2016), encountering a main clause verb may force the parser to "wrapup" and consolidate the encoding of the previous context to conserve memory resources.

Second, we modified the standard ACT-R equation for activation values (Eq. 1) to accommodate HRR vectors. This required us to substitute the calculation of cosine similarity (Eq. 9) for the third term of the standard ACT-R activation equation (Eq. 1). This is the term that computes the penalty for a partial match between the cues of the retrieval probe and the feature values of chunk *i*. In stage 1, cosine similarity is computed over individual feature vectors, whereas in stage 2, it is computed over reduced representation vectors.

7. Simulations

Our goal was to determine whether the contrasting NPIs profiles reported in Parker and Phillips (2016) would be best captured by the original LV05 ACT-R model or the integrated HRR/ACT-R model. To this end, we conducted side-by-side comparisons of the LV05 model with the integrated model, without adjusting model parameters.

7.1. Procedure

Previous implementations of ACT-R have included a wide range of modules for visual information processing, lexical access, memory retrieval, and syntactic parsing (e.g., Lewis & Vasishth, 2005; Vasishth et al., 2008). However, the simulations reported here focus solely on the retrieval module and abstract away from the contribution of the peripheral modules by stipulating the chunks in memory and retrievals required to parse a sentence. There are additional processes associated with sentence comprehension that contribute to behavioral measures, but for current purposes, we adopt the standard assumption that the dynamics and output of memory retrieval map monotonically to the behavioral measures of interest (Anderson & Milson, 1989; Vasishth et al., 2008).

To maximize transparency and simplicity, we implemented the memory retrieval module of ACT-R (i.e., Eqs. 1–7) in the R software environment (R Core Team, 2014), using code originally developed by Badecker and Lewis (2007). Three conditions were simulated, manipulating the presence and location of an NPI licensor (appropriate licensor, irrelevant licensor, no licensor) and the position of the NPI (main clause, embedded

clause), based on the sentence structures in Example 3 from Parker and Phillips (2016). For each condition, a schedule of constituent creation times and retrievals was estimated from the reading times reported in Parker and Phillips (2016). Differences between conditions were modeled only as differences in NPI position and the feature composition of the licensors (±scope, ±negation).

To ensure that the modeling results for the LV05 and integrated HRR/ACT-R model would be directly comparable, all models used the same default parameter settings reported in Lewis and Vasishth (2005) and Vasishth et al. (2008). The only exception was the scaling parameter F, which was optimized to fit the behavioral time scale (in all models, $F = 0.6$). A total of 5,000 Monte Carlo simulations were run for each condition, yielding a solid representation of the model's behavior (Ritter, Schoelles, Quigley, & Klein, 2011).

We report two measures of interest, following Vasishth et al. (2008). Retrieval error rate reflects the percentage of runs for which the distractor (the item in the irrelevant licensor position) was retrieved, rather than the target (the item in the relevant licensor position). This measure maps monotonically to human speeded acceptability judgments, with higher retrieval error rates corresponding to increased rates of judgment errors. Retrieval latencies reflect the average amount of time it took to retrieve the most probable item, and map monotonically to human reading times, with higher latencies corresponding to increased reading times. These measures were used to determine the predicted interference effect, which was calculated as the difference in predicted mean error rates and mean retrieval latencies between the ungrammatical conditions with and without a negative distractor. We focused on these conditions because NPI interference is observed only in ungrammatical conditions. Thus, for predicted error rates, a positive value corresponds to an interference effect, reflecting increased rates of acceptance for sentences with a distractor, relative to sentences with no distractor, and a larger positive value corresponds to more interference. For predicted retrieval latencies, a negative value corresponds to an interference effect, reflecting facilitated processing for sentences with a distractor, relative to sentences with no distractor, and a smaller negative value corresponds to more interference.

7.2. Simulation results

We compared the interference effects observed in Parker and Phillips (2016) with those predicted by the LV05 model and the integrated HRR/ACT-R model for the reading time measures (Fig. 2) and judgment data (Fig. 3). Error bars show levels of variance in the model and observed data using standard error of the mean across model runs (model) and participants (observed) for each condition.

Across both behavioral measures, the integrated HRR/ACT-R model provided a better fit to the observed data, without the need to adjust key model parameters (fit with the HRR/ACT-R model was adjusted $R^2 = 0.79$; fit with the LV05 model was adjusted $R^2 = 0.28$; values were based on the four conditions for reading time and judgment data, i.e., Figs. 1 and 2 combined). The LV05 model failed to capture the observed on/off

behavior, predicting similar rates of interference across NPI positions. The integrated model, on the other hand, captured the basic contrast between NPI positions, with a substantially attenuated interference effect for NPIs in an embedded clause position, relative to NPIs in the main clause position.

Although the values predicted by the integrated HRR/ACT-R model did not match the observed data perfectly, the predicted profiles were qualitatively similar to the observed data. We could explore different parameter values to achieve an even better fit with the observed data, but this was not our goal. Rather, our goal was to determine whether the ACT-R model enhanced with a holographic declarative memory system would predict the basic contrasts across NPI positions, without adjusting previously fixed parameter

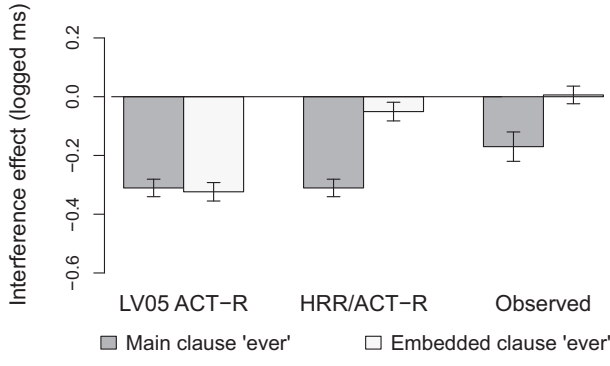

Fig. 2. Comparison of predicted and observed interference effects for reading time measures of main clause *ever* versus embedded clause *ever*. Predicted model values are based on 5,000 runs for each condition. Observed data from Parker and Phillips (2016). Error bars show levels of variance in the model and observed data using standard error of the mean across model runs (model) and participants (observed) for each condition.

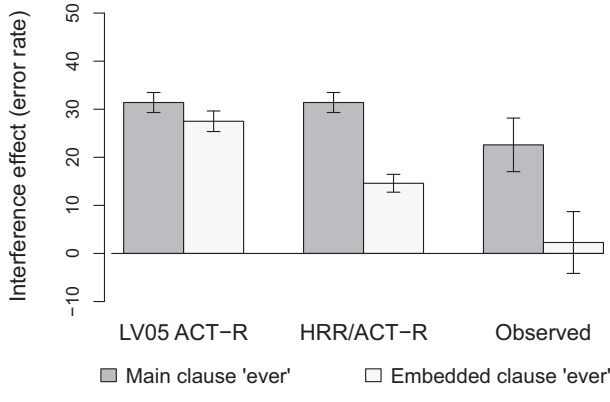

Fig. 3. Comparison of predicted and observed interference effects in judgment accuracy for main clause *ever* and embedded clause *ever*. Predicted model values are based on 5,000 runs for each condition. Observed data from Parker and Phillips (2016).

values. Our simulation task can be viewed as a success, as it confirmed that the integrated HRR/ACT-R model can better capture the basic contrasts.

7.3. Discussion

The contrasting profiles predicted by the integrated HRR/ACT-R model are consistent with the hypothesis proposed by Parker and Phillips (2016) that the accessibility of compositional-semantic features in the encoding is not fixed, as assumed in previous work, but rather, changes over time. In the initial stage, the individual features of a compositional representation are independently evaluated, creating the opportunity for partial match interference. Then, at a later stage, those same features are bound together, such that the representation must exhibit an all-or-none match to the cues of the retrieval probe in order to be recovered, reducing the possibility for partial match interference. Our simulations showed that the integrated HRR/ACT-R model provides a good quantitative fit to the observed human data, without adjusting model parameters.

These findings suggest several avenues for future research. The results raise the question of where else we might observe similar effects. Recent work suggests that an integrated HRR/ACT-R system can explain a wide range of general cognitive effects. For instance, as noted in our Introduction, Rutledge-Taylor et al. (2014) and Kelly et al. (2015) have shown that a holographic declarative memory system, similar to the one proposed here, can be integrated into ACT-R to explain decision-making tasks, the fan effect, and delayed learning. These results suggest that our model is not simply a "one off" model built to explain a narrow range of effects. Instead, our study demonstrates that this unified framework can also capture specialized cognitive abilities involving sentence processing. Specific to language processing, it is important to determine what other types of linguistic dependencies might be impacted by changes in the format of the encoding. Parker and Phillips (2016) found that such effects are likely limited to semantically or pragmatically licensed dependencies, and our model predicts that other types of semantically licensed dependencies, such as those involving certain types of ellipsis, might show similar effects. We leave further investigation of this issue to future research.

Another issue concerns the algorithm for generating reduced or compressed representations. There are numerous methods for generating reduced representations, including convolution, element-wise multiplication (Gayler, 2003; Kanerva, 1994, 1996, 1997), and permutation-based thinning (Rachkovskij & Kussel, 2001). An important task for future research is to verify the predictions of these different binding methods and to explore their empirical consequences for a wide range of cognitive tasks.

8. Conclusion

We presented a computational model that integrates a holographic memory system into the ACT-R model of sentence processing to explain how certain types of compositional

linguistic structures are encoded and accessed in memory. Modeling results showed that the integrated system is better suited to capture the observed behavioral profiles, compared to existing models, yielding a good quantitative fit to data from several behavioral tasks. These results provide proof-of-concept for the unification of two independently developed computational cognitive frameworks and offer new insights into how humans encode and access compositional representations in memory.

Acknowledgments

The code for the Lewis and Vasishth (2005) ACT-R model was generously made available by Rick Lewis. We thank Alan Du for his revisions and additions to this code. We would like to thank John Hale, Luiza Newlin-Łukowicz, Colin Phillips, David Reitter, Frank Ritter, Matt Wagers, and three anonymous reviewers for their feedback on this work.

Notes

1. In the sentence *The bills that no senators supported will ever become law*, the negative distractor *no senators* is embedded inside a subject-modifying relative clause and hence is not syntactically higher than the NPI *ever*, which appears in the main clause.
2. An alternative account proposed by Xiang et al. (2009) argues that NPI interference reflects over-application of pragmatic inferencing mechanisms, rather than misretrieval. This account also predicts that NPI interference effects should generalize across environments.
3. We have based our description of Eqs. 1–7 on ACT-R 6.0. Readers familiar with the LV05 ACT-R model may notice the non-standard presentation of Eq. 1: the sign on the partial match component has been moved outside of the summation to indicate its penalizing nature.
4. In the self-paced reading task reported in Parker and Phillips (2016), sentences were initially masked by dashes, with white spaces and punctuation intact. Participants pressed the space bar to reveal each word. Presentation was non-cumulative, such that the previous word was replaced with a dash when the next word appeared. Each sentence was followed by a comprehension question to ensure that participants were reading the sentences. In the speeded acceptability task, sentences were presented one word at a time at a fixed rate. At the end of the sentence, participants had 3s to make a "yes/no" response about the perceived acceptability of the sentence. Both tasks are widely used in psycholinguistics.
5. Importantly, the component features of the representation are not forever lost but require time-consuming decomposition operations to be recovered.
6. Convolution is the core mathematical operation behind holography, hence the term "holographic."

7. The integrated HRR/ACT-R model can be downloaded from https://github.com/WM-CELL/HRR-ACT-R.

References

Anders, J. R., & Reder, L. M. (1999). The fan effect: New results and new Theories. *Journal of Experimental Psychology: General*, *128*, 186–197.

Anderson, J. R. (1974). Retrieval of propositional information from long-term memory. *Cognitive Psychology*, *6*, 451–474.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R., Bothell, D., Byrne, M. D., Douglass, S., Lebiere, C., & Qin, Y. (2004). An integrated theory of the mind. *Psychological Review*, *111*, 1036–1060.

Anderson, J. R., & Matessa, M. (1997). A production system theory of serial memory. *Psychological Review*, *104*, 728–748.

Anderson, J. R., & Milson, R. (1989). Human memory: An adaptive perspective. *Psychological Review*, *96*, 703–719.

Badecker, B., & Lewis, R. L. (2007). A new theory and computational model of working memory in sentence production: Agreement errors as failures of cue-based retrieval. Talk at the 20th Annual CUNY Conference on Human Sentence Processing. University of California, San Diego.

Cowan, N. (2001). The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, *24*, 87–185.

Dillon, B., Mishler, A., Sloggett, S., & Phillips, C. (2013). Contrasting intrusion profiles for agreement and anaphora: Experimental and modeling evidence. *Journal of Memory and Language*, *69*, 85–103.

Drenhaus, H., Saddy, D., & Frisch, S. (2005). Processing negative polarity items: When negation comes through the backdoor. In S. Kepser & M. Reis (Eds.), *Gradience in grammar: Generative perspectives* (pp. 145–164). New York: Oxford University Press.

Fodor, J. A. (2001). Language, thought, and compositionality. *Mind & Language*, *16*, 1–15.

Gayler, R. W. (2003). Vector symbolic architectures answer Jackendoff's challenges for cognitive neuroscience. In P. P. Slezak (Ed.), *Proceedings of the Joint International Conference on Cognitive Science* (pp. 133–138). Sydney: University of New South Wales.

Kanerva, P. (1994). The binary spatter code for encoding concepts at many levels. In M. Marinaro & P. Morasso (Eds.), *ICANN '94: Proceedings of the International Conference on Artificial Neural Networks* (pp. 226–229). London: Springer-Verlag.

Kanerva, P. (1996). Binary spatter-coding of ordered K-tuples. In von der Malsburg C., von Seelen W., J. C. Vorbruggen, & B. Sendhoff (Eds.), *Artificial Neural Networks — ICANN 96 proceedings (Lecture Notes in Computer Science*, Vol. 1112), pp. 869–873. Berlin: Springer.

Kanerva, P. (1997). Fully distributed representation. In *Proceedings of the 1997 Real World Computing Symposium: Real World Computing Partnership* (pp. 358–365). Tsukuba-city, Japan: Real World Computing Partnership.

Kelly, M. A., Kwock, K., & West, R. L. (2015). Holographic declarative memory and the fan effect: A test case for a new memory module for ACT-R. In N. A. Taatgen et al. (Eds.), *Proceedings for the 2015 International Conference on Cognitive Modeling (ICCM)* (pp. 148–153). Groningen, The Netherlands.

Lewis, R. L., & Vasishth, S. (2005). An activation-based model of sentence processing as skilled memory retrieval. *Cognitive Science*, *29*, 375–419.

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

Martin, A. E., Nieuwland, M. S., & Carreiras, M. (2012). Event-related brain potentials index cue-based retrieval interference during sentence comprehension. *NeuroImage*, *59*, 1859–1869.

Martin, A. E., Nieuwland, M. S., & Carreiras, M. (2014). Agreement attraction during comprehension of grammatical sentences: ERP evidence from ellipsis. *Brain & Language*, *135*, 42–51.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*, 81–97.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Parker, D., Lago, M., & Phillips, C. (2015). Interference in the processing of adjunct control. *Frontiers in Psychology*, *6*, 1–13.

Parker, D., & Phillips, C. (2016). Negative polarity illusions and the encoding of compositional representations. *Cognition*, *157*, 321–339.

Parker, D., & Phillips, C. (submitted). *Reflexive attraction in comprehension is selective*.

Phillips, C., Wagers, M. W., & Lau, E. F. (2011). Grammatical illusions and selective fallibility in real-time language comprehension. In J. Runner (Ed.) *Experiments at the interfaces: Syntax & Semantics* (Vol. 37, pp. 147–180). United Kingdom: Emerald Group.

Plate, T. (2003). *Holographic reduced representation: Distributed representation for cognitive structures*. California: CSLI Publications.

Pollard, C., & Sag, I. A. (1994). *Head-driven phrase structure grammar*. Chicago: University of Chicago Press.

R Core Team. (2014). The R Project for Statistical Computing. Available at https://www.r-project.org.

Rachkovskij, D. A., & Kussel, E. M. (2001). Binding and normalization of binary sparse distributed representations by context-dependent thinning. *Neural Computation*, *2*, 411–452.

Ritter, F. E., Schoelles, M. J., Quigley, K. S., & Klein, L. C. (2011). Determining the number of model runs: Treating cognitive models as theories by not sampling their behavior. In L. Rothrock & S. Narayanan (Eds.), *Human-in-the-loop simulations: Methods and practice* (pp. 97–116). London: Springer-Verlag.

Rutledge-Taylor, M. F., Kelly, M. A., West, R. L., & Pyke, A. A. (2014). Dynamically structured holographic memory. *Biologically Inspired Cognitive Architectures*, *9*, 9–32.

Sloggett, S. (2013) Case licensing in processing: Evidence from German. Poster at the 26th CUNY Conference on Human Sentence Processing. Columbia, South Carolina.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159–216.

Tanner, D., Nicol, J., & Brehm, L. (2014). The time-course of feature interference in agreement comprehension: Multiple mechanisms and asymmetrical attraction. *Journal of Memory and Language*, *76*, 195–215.

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

Vasishth, S., Brüssow, S., Lewis, R. L., & Drenhaus, H. (2008). Processing polarity: How the ungrammatical intrudes on the grammatical. *Cognitive Science*, *32*, 685–712.

Wagers, M., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Representations and processes. *Journal of Memory and Language*, *61*, 206–237.

Xiang, M., Dillon, D., & Phillips, C. (2009). Illusory licensing effects across dependency types: ERP evidence. *Brain & Language*, *108*, 40–55.

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Data S1. Code for the Integrated HRR/ACT-R Model