Opinion TRENDS in Cognitive Sciences Vol.10 No.10

Computational principles of working memory in sentence comprehension

Richard L. Lewis[1], Shravan Vasishth[2] and Julie A. Van Dyke[3]

[1] Department of Psychology, University of Michigan, 530 Church St, Ann Arbor, MI 48109, USA
[2] Institute for Linguistics, University of Potsdam, Postfach 601553, D-14415 Potsdam, Germany
[3] Haskins Laboratories, 300 George Street, New Haven, CT 06511, USA

Understanding a sentence requires a working memory of the partial products of comprehension, so that linguistic relations between temporally distal parts of the sentence can be rapidly computed. We describe an emerging theoretical framework for this working memory system that incorporates several independently motivated principles of memory: a sharply limited attentional focus, rapid retrieval of item (but not order) information subject to interference from similar items, and activation decay (forgetting over time). A computational model embodying these principles provides an explanation of the functional capacities and severe limitations of human processing, as well as accounts of reading times. The broad implication is that the detailed nature of crosslinguistic sentence processing emerges from the interaction of general principles of human memory with the specialized task of language comprehension.

Introduction

Understanding spoken or written language in real time requires the rapid, incremental processing of novel compositional structures. The language-processing system must therefore maintain, at least momentarily, some memory of linguistic material. For example, in sentences with long-distance dependencies such as (1), the noun phrase (NP) '*a toy*' must be maintained in memory until it is interpreted as the object of '*like*':

(1) This weekend we bought a toy that Amparo really hoped Melissa would like.

In general, establishing any kind of linguistic relationship requires some memory of the immediate past. This functional requirement gives rise to the following theoretical question: what are the working memory processes that bring prior linguistic material into contact with present material, and what are the constraints on those processes?

Our focus on the memory processes underlying comprehension departs from one of the central research topics in psycholinguistics: understanding the nature of ambiguity resolution [1]. Theories of ambiguity resolution take the form of principles or processes predicated over alternative structures, but they generally do not describe the memories and processes that give rise to those structures. A separate tradition has sought to provide metrics to characterize the memory load associated with linguistic structures [2,3] but this work has not made significant contact with theories of memory in cognitive psychology [4], and has not yielded models of the underlying working memory architecture.

Here, we present an emerging theoretical framework that seeks to bridge this gap and directly answer the question above. The basic framework claims that language processing, although it might operate on specialized representations, is nevertheless subject to general processing principles and constraints that govern other domains of memory [5–7]. The application of these principles to sentence processing results in a surprising breadth of explanation and a range of detailed predictions, which we review here.

This approach builds on several important recent trends:

- Increased empirical attention to crosslinguistic phenomena that reveal constraints on working memory, independently of ambiguity resolution [8–12].
- A shift to more highly constrained and articulated models of verbal working memory that include explicit assumptions about retrieval [13,14].
- The development of detailed computational theories of cognition in the form of cognitive architectures that embody specific assumptions about constraints and functional capacities, and are applicable to a range of domains, including language processing [15–17].

By combining insights from these three lines of work, we believe it is possible to develop a coherent theoretical framework for understanding the working memory system that supports sentence comprehension.

Retrieval from working memory

Where might we look for cognitive principles that will inform a theory of working memory in sentence processing? The influential Baddeley–Hitch framework [18] and its phonological loop is an obvious candidate but substantial evidence suggests that the role of phonological storage and rehearsal processes in sentence processing is limited [6,19]. Instead, investigations into the nature of retrieval from verbal working memory form a more promising foundation [20]. We now briefly describe three core constraints that derive from this work.

Limited focus of attention

Recent models of working memory emphasize a limited focus of attention with capacity much smaller than the

Corresponding author: Lewis, R.L. (rickl@umich.edu).
Available online 1 September 2006.

www.sciencedirect.com 1364-6613/$ – see front matter doi:10.1016/j.tics.2006.08.007

Box 1. Working memory retrieval and cue-based parsing: linguistic processing with an extremely limited attentional focus

Figure I illustrates how cue-based parsing might work to establish the relationship between a verb ('*arrived*') and the distal head of its subject ('*toy*'). Linguistic items in memory are represented as feature bundles (chunks [16,28]). The result of processing the subject noun '*toy*' in the complement clause is an encoding in the working memory of both a representation of the noun phrase itself (shown as NP6 in Figure I) and a representation of the expectation of the predicate that the noun is a subject of (shown as S7). This expectation remains in the memory – but outside the focus of attention – until its retrieval at the verb. This retrieval is driven by a set of retrieval cues set in response to the verb (in the simplified example below, the verb effectively triggers a request for a predicted item of category S whose head is currently empty). The retrieval process is subject to constraints described in the text.

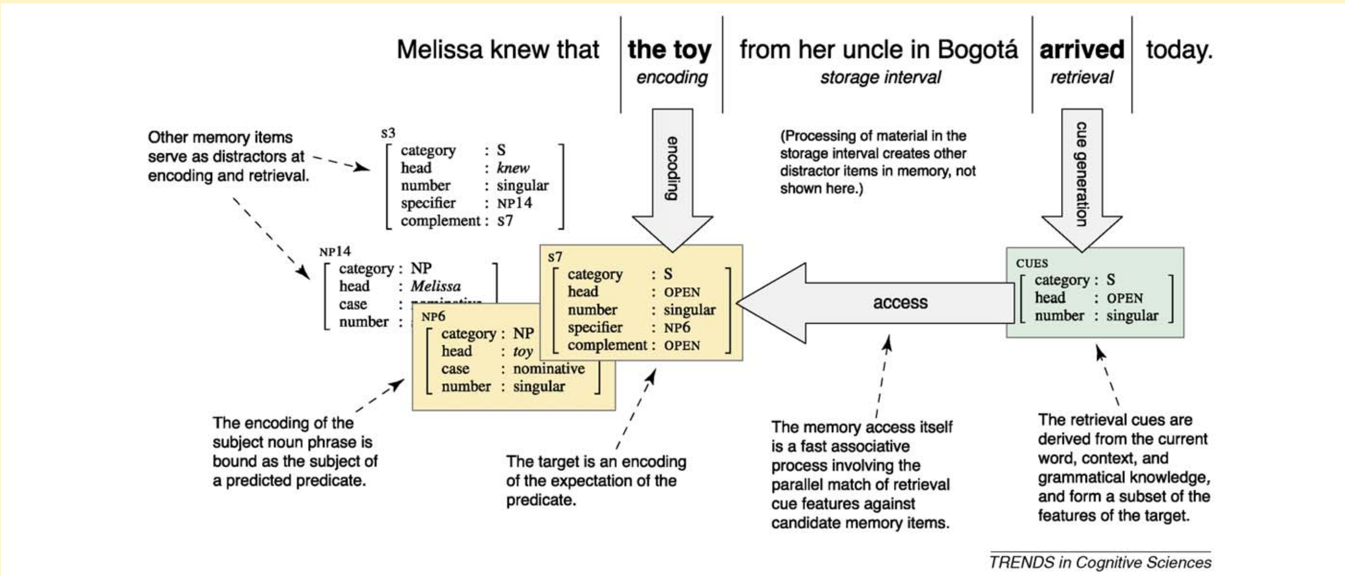

Figure I.

span measured by tasks such as serial recall [13,14]. The minimum capacity required to establish novel relations is two items, and there is growing empirical evidence from speed–accuracy trade-off (SAT) analyses that this minimal capacity corresponds to the actual capacity of the human system [13].

Fast, parallel associative retrieval of item information

Evidence from distributional analyses of reaction times [21] and SAT paradigms [13] suggests a fast retrieval that involves a parallel match of cues against all items in memory. The estimate of memory retrieval times from the SAT studies is about 80–90 ms, which, plausibly, is fast enough to support sentence comprehension.

Slow, serial access to serial order information

SAT paradigms also provide compelling evidence against fast parallel access of serial order information [13]. Instead, the empirical results from tasks requiring order information indicate a sequential access that takes hundreds of milliseconds, depending on serial position – much too slow to support sentence comprehension, which proceeds at a rate of 250–300 ms per word.

Grounding sentence parsing in cue-based retrieval

How can a sentence processor function under such severe constraints – a focus capacity of two with no serial order? Several researchers have recently advocated a theoretical framework – cue-based parsing [11,22,23] – which provides an answer. In this framework, the functional requirements for working memory are met by a combination of a limited focus and mechanisms for rapidly encoding and retrieving information from a secondary store [24].

Box 1 illustrates this view [11,22]. The basic idea is that each incoming word triggers retrievals to integrate that word with the preceding structure. Retrieval is accomplished by a simple type of associative access: content-based retrieval, where the retrieval cues are a subset of the features of the item to be retrieved [16,21,25]. The specific example used in Box 1 shows a retrieval triggered by a verb, although retrievals are not restricted to verbs or any other word class.

The incoming lexical item and the current state of the parse are mapped onto the appropriate set of retrieval cues as specified by the grammar. In our model, this grammatical knowledge is encoded in procedural form [22], consistent with recent cognitive neuroscience models that map grammar onto procedural memory systems in the brain [26,27]. However, the model does not claim that words trigger the retrievals of other words via lexical associations. What is being retrieved are partial representations of linguistic constituents, not words (represented as feature bundles, or 'chunks' [16,28]), and the cues that drive the retrieval are not simply features of the lexical item but are grammatically derived from the current word and context.

Constraints on encoding, storage and retrieval: empirical evidence

Cue-based parsing naturally carves up working memory processes into stages of encoding, storage and retrieval. Figure 1 provides a road map to the empirical evidence

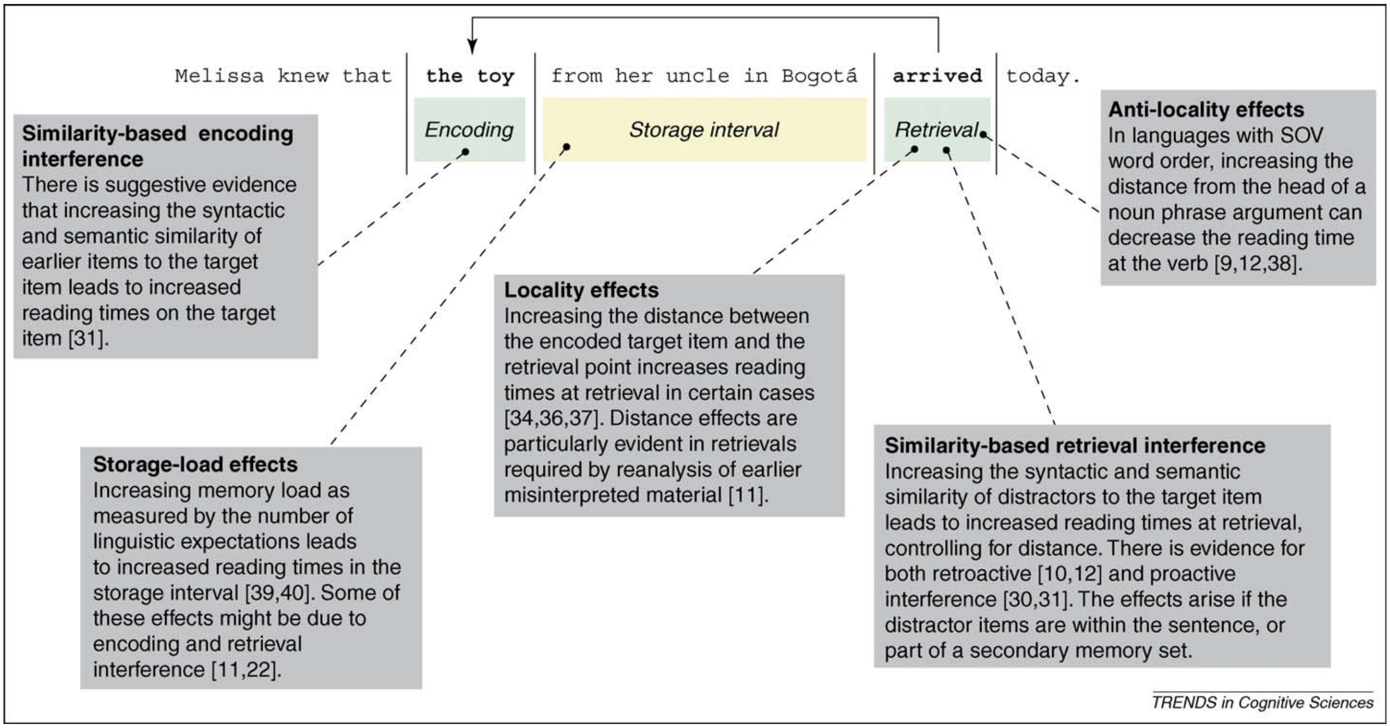

Figure 1. Recent empirical work investigating incremental structure building in sentence processing can be usefully summarized in terms of distinct effects on the working memory processes of encoding, storage and retrieval. The figure illustrates the three processes with a simple example, showing the encoding of a subject noun-phrase, '*the toy*', and its later retrieval at the associated verb, '*arrived*'. See the text and Box 1 for more detail about the nature of the encoded and retrieved structures. Abbreviation: SOV, subject-object-verb word order.

associated with these stages and indicates their locus in online reading. This evidence is consistent with the view that working memory processes are constrained by similarity-based interference and activation decay (more generally, activation as a function of time and history of retrievals). We first review the evidence for interference, and then consider locality and storage-load phenomena and their relationship to decay and interference.

Similarity-based retrieval interference

Similarity effects are the hallmark of interference, and there is considerable evidence for similarity-based retrieval interference in the verbal memory literature [6,29]. Several theorists have posited that such interference also constrains the rapid retrievals from working memory in sentence processing [10,11,22,30].

An effect of similarity-based retrieval interference is observed if increasing the similarity of distractors (either preceding or following the target) to the retrieval cues used to access the target increases difficulty (and thus reading time or errors) at the point of establishing the relationship that requires the retrieval. Van Dyke and Lewis [11] found direct evidence for retroactive interference by manipulating the structural similarity of the intervening region while keeping the distance of the linguistic relationship constant (Box 2). Other evidence of retroactive interference comes from experiments in Hindi that manipulated the semantic similarity of intervening nouns to the target nouns, resulting in increased reading times at the crucial verbs [12].

There is also evidence for proactive interference from distractors that come before the desired target. In one paradigm, introduced by Gordon *et al.* [31], proactive interference is created by an extra-sentential memory load of a small set of nouns that must be maintained while reading a sentence. Reading times in the verb regions that trigger the retrievals increased when the semantic type of the memory-load nouns matched the target noun.

Van Dyke and McElree [30] reported a related effect but showed that proactive interference is due to retrieval cue overlap. Using the Gordon *et al.* [31] paradigm, they manipulated the verb and its semantic constraints, while keeping the memory load items constant. Subjects remembered a short list of nouns as shown in (2a), and then read sentences containing a verb that either semantically selected for the class of nouns in the memory list (2b) or did not (2c):

(2a) MEMORY LIST: (table, sink, truck)
(2b) INTERFERING: It was the boat that the guy who lived by the sea *fixed* in two sunny days.
(2c) NON-INTERFERING: It was the boat that the guy who lived by the sea *sailed* in two sunny days.

The key prediction of retrieval interference was an interaction of load (present or not) and verb type (interfering or not) in reading times at the crucial verb region. This was the pattern observed in the data: in the interference conditions, subjects took longer to read the crucial verb than in the noninterference condition, relative to a nonload baseline.

There is also limited but suggestive evidence for similarity-based encoding interference: an increase in processing time associated with encoding an item as a function of its similarity to preceding items (Figure 1).

In summary, both retroactive and proactive similarity-based retrieval interference – and, possibly, encoding interference – arises in sentence processing, providing

Box 2. Empirically distinguishing the effects of decay and interference in sentence processing

Van Dyke and Lewis [11] independently varied distance and interference and found distinct effects of both. There were three manipulations: distance (short versus long), structural interference (low versus high) and ambiguity (ambiguous versus unambiguous). Examples are shown below:

(a) (SHORT, UNAMBIGUOUS) The assistant *forgot that* the student *was standing* in the hallway.
(b) (LONG, LOW-INTERFERENCE, UNAMBIGUOUS) The assistant *forgot that* the student who was waiting for the exam *was standing* in the hallway.
(c) (LONG, HIGH-INTERFERENCE, UNAMBIGUOUS) The assistant *forgot that* the student who knew that the exam was important *was standing* in the hallway.
(d) (SHORT, AMBIGUOUS) The assistant *forgot* the student *was standing* in the hallway.
(e) (LONG, LOW-INTERFERENCE, AMBIGUOUS) The assistant *forgot* the student who was waiting for the exam *was standing* in the hallway.
(f) (LONG, HIGH-INTEREFERENCE, AMBIGUOUS) The assistant *forgot* the student who knew that the exam was important *was standing* in the hallway.

The crucial region is the phrase '*was standing*', which must be attached to the predicted sentential complement of '*forgot*' in all six conditions. Consider first the unambiguous conditions (a–c), which provide an estimate of this attachment cost independently of ambiguity (the cost of retrieving the predicted complement). In the short condition (a), nothing intervenes except the subject '*the student*'. In the low-interference condition (b), there is an intervening relative clause and prepositional phrase. The high-interference condition (c) is approximately as long but includes another sentential complement, which provides a similar distractor to the target. Van Dyke and Lewis argued that the contrast between (a) and (b) provides an estimate of a distance effect, and the contrast between (b) and (c) provides an estimate of an additional interference effect.

Interference had a significant effect on reading times at the crucial region (Figure Ia) as well as on grammaticality judgement accuracies (Figure Ib), but distance did not (indeed there was an antilocality trend). This result is consistent with the view that pure distance effects might be difficult to find because memory elements might be reactivated during the parse; in this case, the predicted sentential complement is reactivated at least once to attach '*the student*' as subject.

Van Dyke and Lewis reasoned that decay effects would be most evident in cases where reactivations were eliminated. Locally ambiguous garden path structures provide such cases because the less-preferred interpretation is not actively pursued. Conditions (d–f) drop the complementizer '*that*', introducing a local ambiguity between a direct object and sentential complement structure. Increased reading time in the ambiguous conditions indicates a garden path effect – the cost of reactivating the discarded sentential complement prediction. Distance should have an effect on this garden path cost because the target structure suffers decay without reactivation, whereas interference should have no further effect because the interfering material is identical in the ambiguous and unambiguous conditions. The results confirmed this prediction: the effects of distance and interference were reversed.

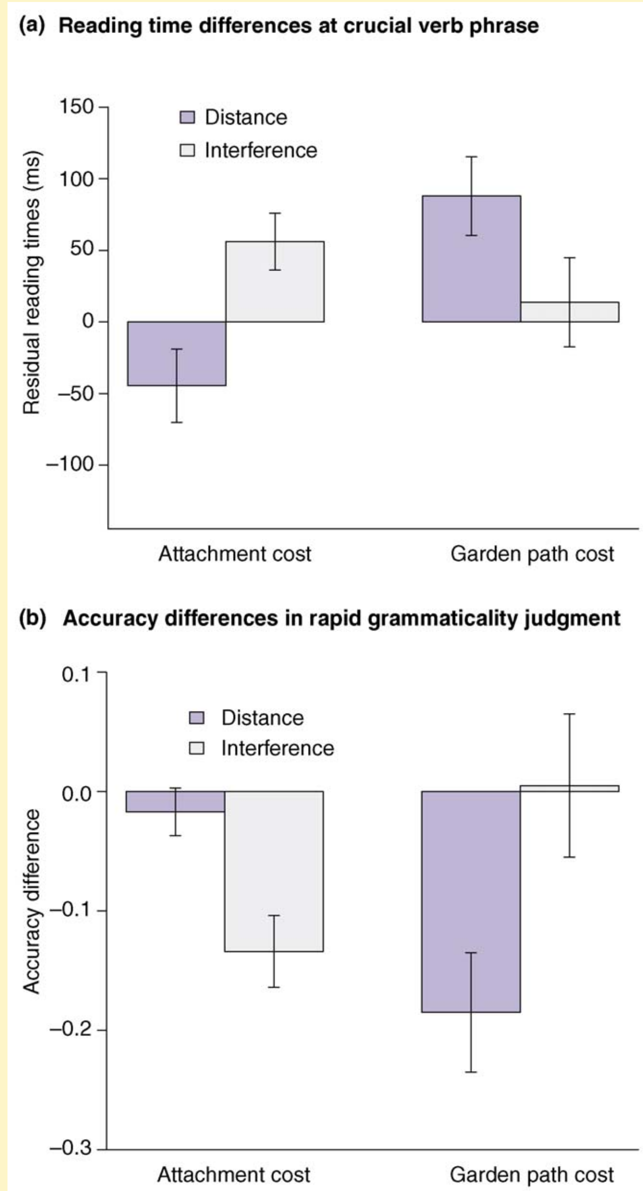

Figure I. Effects of decay and interference in sentence processing. See Box 2 text for conditions (a–f). For the attachment cost, the distance effect is the contrast (b)–(a) and the interference effect is (c)–(b). For garden path cost, the distance effect is [(e)–(b)] – [(d)–(a)] and the interference effect is [(f)–(c)] – [(e)–(b)]. Reading time differences **(a)** are computed from residuals after regressing out effects of word length. Error bars represent standard errors. Reproduced, with permission, from Ref. [11].

an important link between verbal memory theory and parsing. Furthermore, both structural and semantic similarity effects have been found, suggesting that retrieval cues include grammatical and semantic features [30], consistent with some constraint-based approaches to comprehension [32].

A second phenomenon broadly implicated in verbal memory is decay [18]. One possibility is that increasing the distance between two linguistically dependent items creates difficulty for language processing because the distant item decays over time [33]. We see below that the role of decay in these sorts of nonlocal attachments might actually be considerably more limited.

Locality effects

Locality is a central concept in several prominent accounts of sentence complexity [3,34] and also several influential principles of ambiguity resolution [35]. In Dependency Locality Theory [2], distance is quantified by the number of new discourse referents that intervene between a head (e.g. a verb) and its dependents (its arguments); this calculation includes verbal elements and the dependent itself. For example, consider the object relative clause sentences below [36]. The distance between the embedded verb '*supervised*' and the head noun of its subject '*nurse*' is indicated for each sentence type:

(3a) (DISTANCE = 1) The administrator who the nurse$_1$ *supervised* scolded the medic while...

(3b) (DISTANCE = 2) The administrator who the nurse$_1$ from the clinic$_2$ *supervised* scolded the medic while...

(3c) (DISTANCE = 3) The administrator who the nurse$_1$ who was$_2$ from the clinic$_3$ *supervised* scolded the medic while...

According to the dependency locality theory, reading times at the embedded verb 'supervised' should be a function of distance; that is, the fastest times should be observed in (3a), with progressively slower times in (3b) and (3c). This is the pattern reported by Grodner and Gibson [36] in a self-paced reading study involving these constructions.

This relatively simple complexity metric can account for a variety of offline and online behavioral data, at least in English and Japanese [2,8,34]. The basis of the metric in discourse processing complexity receives support from a study [37] that manipulated the referential type of intervening noun phrases while keeping length constant; reading times were faster at crucial verbs when the intervening nouns were referentially more accessible. We suggest that locality effects have their source in both interference and decay but first we will discuss the surprising presence of antilocality effects, which place considerable constraint on any processing explanation of locality.

Antilocality effects

Recent crosslinguistic studies have indicated that there are both limits to locality and direct counterexamples to it. For example, locality effects were not observed at the main verb as a function of the distance to the head noun of the subject noun phrase, using the same distance manipulations as used in example (3) above [36], suggesting that locality might be most evident at points that are higher in processing load for independent reasons [22].

Direct evidence against locality comes from cases where increasing distance speeds processing. Such antilocality effects have been observed in studies of head-final languages. For example, in a German self-paced reading study, Konieczny [9] showed that in (4a) the verb '*hingelegt*' was read faster than in (4b), despite the longer distance between the verb and its argument.

(4a) (DISTANCE = 2) Er hat das Buch, das Lisa gestern gekauft hatte, hingelegt
He has the book, that Lisa yesterday bought had, laid down
'He has laid down the book that Lisa had bought yesterday'.

(4b) (DISTANCE = 0) Er hat das Buch hingelegt, das Lisa gestern gekauft hatte
He has the book laid down, that Lisa yesterday bought had
'He has laid down the book that Lisa had bought yesterday'.

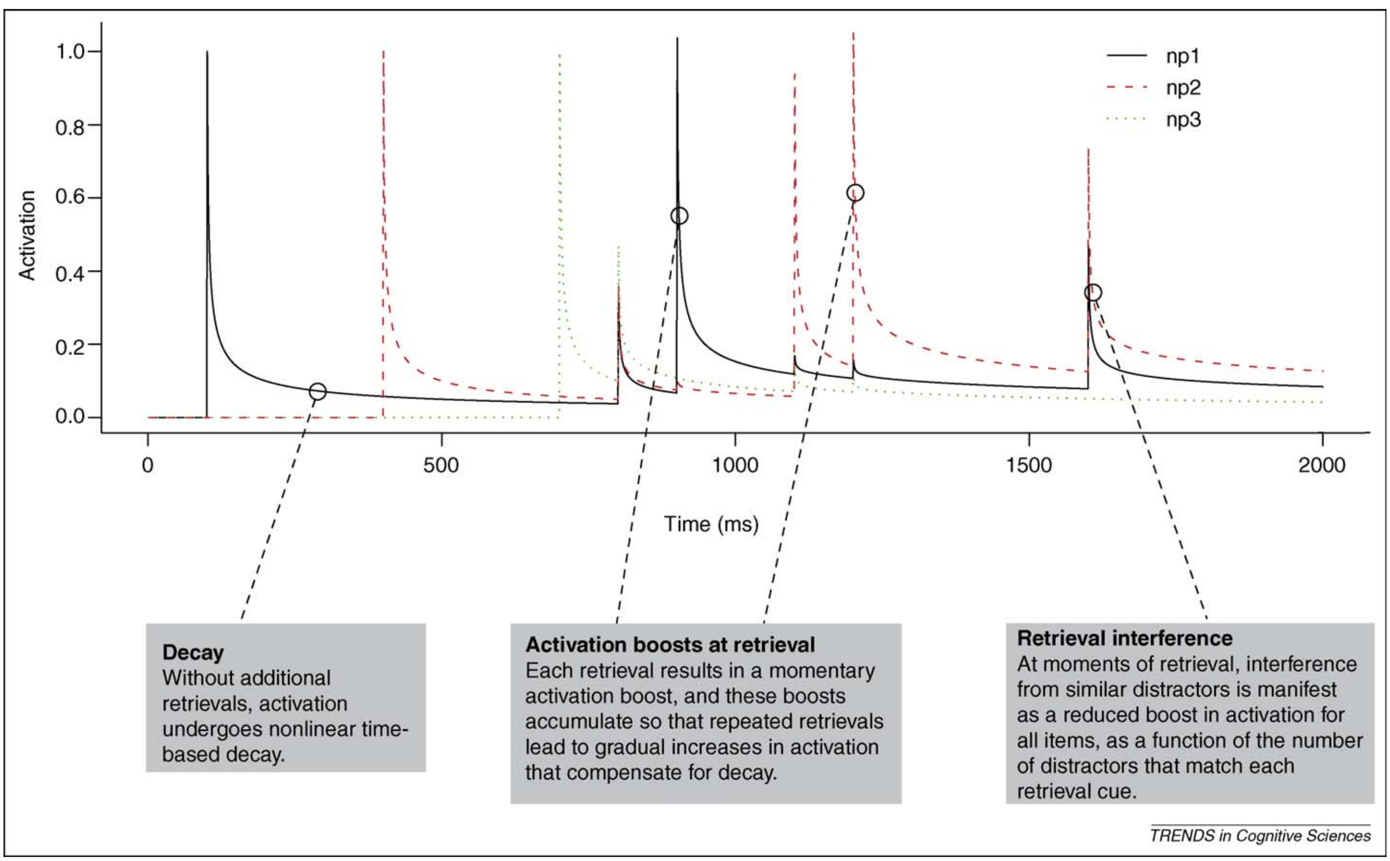

Figure 2. The activation profiles over time of three items in memory undergoing a series of hypothetical retrievals. These activation profiles are generated by a set of simple mathematical equations that form the basis of the ACT-R theory of declarative memory, and provide a unified account of interference, locality, antilocality and storage effects in sentence processing. Locality effects are a consequence of activation decay; antilocality results from repeated activation boosts at retrieval. Similarity-based retrieval interference arises because the strength of association from a cue is reduced as a function of the number of items associated with the cue, reducing the activation boost at retrieval.

Subsequent work on Hindi [12,38] has uncovered more evidence for antilocality. One plausible explanation of these effects, discussed further below, is that the interposed material helps to strengthen the representation of the noun-phrase argument by reactivating it through modification (Figure 2). The bottom line is that expected locality effects do not always arise: a comprehensive theory must account for the highly circumscribed contexts in which they do arise and also their complete reversal in some cases.

The relationship of locality, decay and interference

As described earlier, existing locality metrics are predicated over linguistic objects. Such metrics are therefore closer in form to retroactive interference theories than to decay theories, although they abstract away from commitments to either.

Distance effects are therefore consistent with both decay and interference, and do not provide unambiguous evidence for either. Teasing the two apart is one of the classic empirical problems in memory theory, although one recent study provided evidence that both are manifest in sentence processing [11]. Box 2 describes the result in more detail. The broader implication is that it might not be necessary to make assumptions about the appropriate linguistic metric of distance – the fundamental metric instead might be *time*. Discourse effects [37] remain a challenge to this approach.

In the same way that interference and decay might enable us to dispense with locality metrics, we will now show that interference can also enable us to dispense with storage metrics.

Storage-load effects: distinct from interference?

The working memory theory of Just and Carpenter [17] posits a trade-off between storage and processing: increasing cognitive resources devoted to storage decreases the cognitive resources available for processing. The locality model of Gibson [2] formalizes this in sentence processing by introducing a storage-cost component that increases processing difficulty during the storage interval.

Gibson and co-workers [39,40] present a set of experiments providing evidence for storage effects. The paradigm involves comparing reading times over a region that is held constant while manipulating syntactic load by changing the preceding syntactic context. For example, one experiment manipulated the number of verb predictions (the number of verbs that must appear to complete a grammatical sentence) maintained over the italicized clause below:

(5a) (ZERO PREDICTED VERBS) The detective suspected that the thief knew that *the guard protected the jewels*...

(5b) (TWO PREDICTED VERBS) The suspicion that the knowledge that *the guard protected the jewels*...

In (5b), after reading '*The suspicion that the knowledge that*', at least two verbal predictions must be kept in mind (one for which '*the suspicion*' will be subject, and one for which '*the knowledge*' will be subject). Consistent with the storage-load account, reading times increased over the embedded clause as a function of the number of predicted verbs.

How could interference account for storage effects? In interference theories, there is not a distinct memory cost associated with storage: there is only passive storage and decay. However, stored memories have their observable effects through encoding and retrieval interference, and in this way both kinds of interference predict storage effects. For example, retrieval interference predicts increased reading times on the crucial embedded clause in example (5) because the retrievals involved in integrating the clause suffer proactive interference from the increased number of similar clause predictions.

Interference and storage load theories do make distinct predictions about the localization of processing difficulty. By focusing on regions in the storage interval before crucial retrievals, it should be possible to find evidence for storage effects that cannot be accounted for by interference. In one study investigating reading times in such regions, no storage effects were observed [11]. Thus, interference might account for both the similarity effects described above and also the details of storage effects.

The best known cases of apparent storage overload in sentence processing are the self-embeddings identified by Miller and Chomsky [41], such as example (6):

(6) The diplomat that the editor that the newspaper hired interviewed was...

Can a theory without explicit storage-load constraints account for these classic effects? The following section suggests that such effects arise naturally from similarity-based interference and a severe limit on the use of serial order information.

The role of serial order and difficult self-embeddings

One of the core memory principles identified earlier is that access to serial order information is a slow, serial process [13]. This led to the surprising conclusion that serial order information cannot be used in real-time sentence processing [23], given the time course of its access. This conclusion is joined by computational evidence leading to an equally surprising result: explicit serial order information is almost never needed in sentence processing. The computational solution is embodied in the parser described by Lewis and Vasishth [22]. This parser functions by relying on discriminating retrieval cues and the simple architectural distinction between the past (things outside the focus) and the present (the item just encoded).

A simple example in Japanese will help to make the problem and solution clear. Consider the embedded structure in example (7) below:

(7) Mary-ga John-ga butler-o korosita-to omotta.
Mary-nom John-nom butler-acc killed-comp thought
'Mary thought that John killed the butler'.

The requirement for serial order is clear. Who killed the butler? Upon encountering the verbs, the only way to distinguish the two candidate subject nouns (marked by '-ga') is by their serial order – they are otherwise syntactically and semantically indistinguishable. The system must, it appears, be able to distinguish two distal items exclusively on the basis of serial order.

The solution is straightforward and falls out naturally from cue-based, predictive parsing. The key observation is that the preverbal nouns will already have been structured as arguments of distinctive predicted verb types [42,43] ('*John*' is the subject of a predicted embedded verb taking an NP object, and '*Mary*' is the subject of a predicted verb taking a sentential complement). When the verbs are processed, their retrieval cues include type features that unambiguously identify the correct preceding structure to retrieve. Some interference arises but serial order is not the unique disambiguating cue.

This solution has limits. There are constructions that do require serial order to parse correctly; these are the difficult self-embeddings originally identified by Miller and Chomsky [41]. Lewis and Vasishth [22] present simulations on a range of embedded structures tested in a difficulty rating study [44], ranging from the nearly impossible to the reasonably comprehensible. The difficulty ranking that the simulations imposed was consistent with the difficulty ranking that the human subjects gave. In the worst cases, the parser relies on the residual order information in the noisy activation levels (Figure 2).

Crucially, although the model fails to parse difficult self-embeddings, it succeeds on a wide range of complex syntactic constructions, including some center-embedded structures [22]. It is the combination of successes and failures that forms the basis of the claim that explicit serial order information is not used in online sentence processing. From this perspective, the problem with nested center-embedding is not storage overload but rather impoverished discrimination combined with poor support for serial order. This view is consistent with the early speculation of Chomsky and Miller that susceptibility to similarity is the key characteristic of human memory that gives rise to difficulty with self-embeddings [33].

Unifying principles

Together with the research on working memory retrieval dynamics summarized earlier, the psycholinguistic phenomena and modeling results reviewed here suggest that the working memory system underlying sentence processing operates according to the following computational principles:

(i) Extremely limited focus of attention;
(ii) Fast content-addressed access to item information but not serial order information;
(iii) Similarity-based retrieval interference;
(iv) Fluctuating activation as a function of decay and retrieval history;
(v) Similarity-based encoding interference.

The first four of these principles have already been given a unified mathematical treatment in adaptive control of thought-rational (ACT-R), a domain-independent theory of cognitive architecture that takes computational form [16]. Unification is achieved through a simple set of equations that govern the activation level of memory items at moments of retrieval and storage, and transform the activation levels into retrieval latencies (Figure 2). These computational principles give rise to locality effects, antilocality effects, interference effects and storage-load effects, and can be used to derive quantitative predictions of reading times [22].

Several other recent computational models of parsing are also converging on a type of processing that is similar in important respects to that described here. The models of Tabor *et al.* [45] and Vosse and Kempen [46] share the general property that all partial structures are freely accessible as a function of their content, not as a function of some special accessing mechanisms such as stacks, queues or buffers. As a consequence, many of the significant processing limitations arise as a result of limitations in discriminating similar items, rather than limits on storage.

Conclusions and outstanding research questions

We have proposed a new theoretical unification between sentence processing and independent work on verbal working memory that is beginning to yield significant explanatory dividends. It also opens the door to explicit computational models that account for both the severe limits and extraordinary functional capacity of human sentence parsing.

An important benefit of developing more precise models of working memory in language processing is that they naturally lead to new theoretical questions and new ways of addressing existing questions (Box 3). Although the principles advanced here are likely to undergo major

Box 3. Some outstanding questions for future research

- What is the nature of individual differences in working memory and how does this affect sentence processing?
  The theoretical framework provides new avenues for pursuing the research program initiated by Just and Carpenter [17]; for example, rather than activation differences, individual differences might be manifest in different capacities to resolve interference.
- What are the shared processing resources underlying sentence comprehension and other cognitive processes?
  Shared processing principles need not imply shared processing resources. For example, the principles outlined here might be associated with specialized memories dedicated to syntactic parsing [19], or might characterize a more general system giving rise to graded similarity effects.
- What are the principles of neural computation that give rise to the processing constraints described here, and what is their basis in cortical circuits?
  For example, similarity-based interference and decay are emergent properties of certain attractor neural networks [47], and interference resolution figures prominently in some recent theories of the function of Broca's area [48].
- What is the relationship of the principles (and resources) of working memory in comprehension and production?
  To the extent that processes of language production might be cast as a series of memory retrievals, the effects of interference and decay should arise in production as well.
- How can other levels of linguistic processing (such as prosodic structure and information structure) be integrated, and how might these additional levels be exploited to provide more discriminating retrieval cues?
  For example, one understudied question concerning complex syntactic embeddings is the role that prosodic structure might have in facilitating processing.
- What are the relative roles of working memory principles and principles of information theory accounts of sentence processing such as surprisal?
  For example, in antilocality effects, surprisal predicts facilitation, irrespective of how much distance increases, whereas activation-based accounts predict that increasing distance should result in a speed-up only if the decay does not offset the benefit of reactivation [49].

revision as these issues are pursued, the application of general cognitive principles to the detailed problems of language processing should prove a fruitful research path for some time to come.

Acknowledgements
This paper benefited from the patient criticism of numerous colleagues, including William Badecker, Brian Bartek, Marc Berman, Andrew Howes, John Jonides, Pawel Logatschew and Cindy Lustig, as well as four anonymous reviewers. Any errors and unclarity remain in spite of their generous help.

References
1 Mitchell, D.C. (1994) Sentence parsing. In *Handbook of Psycholinguistics* (Gernsbacher, A.M., ed.), Academic Press
2 Gibson, E.A. (2000) The dependency locality theory: a distance-based theory of linguistic complexity. In *Image, Language, Brain* (Miyashita, Y. *et al.*, eds), MIT Press
3 Hawkins, J.A. (1994) *A Performance Theory of Order and Constituency*. Cambridge University Press
4 Lewis, R.L. (2000) Specifying architectures for language processing: process, control, and memory in parsing and interpretation. In *Architectures and Mechanisms for Language Processing* (Crocker, M.W. *et al.*, eds), Cambridge University Press
5 Bever, T.G. (1970) The cognitive basis for linguistic structures. In *Cognition and the Development of Language* (Hayes, J.R., ed.), pp. 279–362, Wiley
6 Lewis, R.L. (1996) Interference in short-term memory: the magical number two (or three) in sentence processing. *J. Psycholinguist. Res.* 25, 93–115
7 Marcus, G. (2006) Cognitive architecture and descent with modification. *Cognition* DOI: 10.1016/j.cognition.2006.04.009 (www.sciencedirect.com/science/journal/00100277)
8 Babyonyshev, M. and Gibson, E. (1999) The complexity of nested structures in Japanese. *Language* 75, 423–450
9 Konieczny, L. (2000) Locality and parsing complexity. *J. Psycholinguist. Res.* 29, 627–645
10 Gordon, P.C. *et al.* (2001) Memory interference during language processing. *J. Exp. Psychol. Learn. Mem. Cogn.* 27, 1411–1423
11 Van Dyke, J.A. and Lewis, R.L. (2003) Distinguishing effects of structure and decay on attachment and repair: a cue-based parsing account of recovery from misanalyzed ambiguities. *J. Mem. Lang.* 49, 285–316
12 Vasishth, S. and Lewis, R.L. Argument-head distance and processing complexity: explaining both locality and anti-locality effects. *Language* (in press)
13 McElree, B. (2006) Accessing recent events. In *The Psychology of Learning and Motivation* (Vol. 46) (Ross, B.H., ed.), Academic Press
14 Cowan, N. (2001) The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *Behav. Brain Sci.* 24, 87–114
15 Newell, A. (1990) *Unified Theories of Cognition*. Harvard University Press
16 Anderson, J.R. *et al.* (2004) An integrated theory of mind. *Psychol. Rev.* 111, 1036–1060
17 Just, M.A. and Carpenter, P.A. (1992) A capacity theory of comprehension: individual differences in working memory. *Psychol. Rev.* 99, 122–149
18 Baddeley, A. (2000) The episodic buffer: a new component of working memory? *Trends Cogn. Sci.* 4, 417–423
19 Caplan, D. and Waters, G. (2002) Working memory and connectionist models of parsing: a reply to MacDonald and Christiansen (2002). *Psychol. Rev.* 109, 66–74
20 Sternberg, S. (1966) High-speed scanning in human memory. *Science* 153, 652–654
21 Ratcliff, R. (1978) A theory of memory retrieval. *Psychol. Rev.* 85, 59–108
22 Lewis, R.L. and Vasishth, S. (2005) An activation-based model of sentence processing as skilled memory retrieval. *Cogn. Sci.* 29, 375–419
23 McElree, B. *et al.* (2003) Memory structures that subserve sentence comprehension. *J. Mem. Lang.* 48, 67–91
24 Ericsson, K.A. and Kintsch, W. (1995) Long-term working memory. *Psychol. Rev.* 102, 211–245
25 Gillund, G. and Shiffrin, R.M. (1984) A retrieval model for both recognition and recall. *Psychol. Rev.* 91, 1–67
26 Ullman, M.T. (2004) Contributions of memory circuits to language: the declarative/procedural model. *Cognition* 92, 231–270
27 Vannest, J. *et al.* (2005) Dual-route processing of complex words: new fMRI evidence from derivational suffixation. *Cogn. Affect. Behav. Neurosci.* 5, 67–76
28 Miller, G.A. (1956) The magical number seven plus or minus two: some limits on our capacity for processing information. *Psychol. Rev.* 63, 81–97
29 Anderson, M.C. (2003) Rethinking interference theory: executive control and the mechanisms of forgetting. *J. Mem. Lang.* 49, 415–445
30 Van Dyke, J. and McElree, B. (2006) Retrieval interference in sentence comprehension. *J. Mem. Lang.* 55, 157–166
31 Gordon, P.C. *et al.* (2002) Memory load interference in syntactic processing. *Psychol. Sci.* 13, 425–430
32 MacDonald, M.C. *et al.* (1994) The lexical nature of syntactic ambiguity resolution. *Psychol. Rev.* 101, 676–703
33 Chomsky, N. (1965) *Aspects of the Theory of Syntax*. MIT Press
34 Gibson, E.A. (1998) Linguistic complexity: locality of syntactic dependencies. *Cognition* 68, 1–76
35 Frazier, L. and Fodor, J.D. (1978) The sausage machine: a new two-stage parsing model. *Cognition* 6, 291–325
36 Grodner, D.J. and Gibson, E.A. (2005) Consequences of the serial nature of linguistic input for sentential complexity. *Cogn. Sci.* 29, 261–290
37 Warren, T. and Gibson, E. (2002) The influence of referential processing on sentence complexity. *Cognition* 85, 79–112
38 Vasishth, S. (2003) *Working Memory in Sentence Comprehension: Processing Hindi Center Embeddings*. Garland Press
39 Grodner, D.J. *et al.* (2002) Syntactic complexity in ambiguity resolution. *J. Mem. Lang.* 46, 267–295
40 Chen, E. *et al.* (2005) Online syntactic storage costs in sentence comprehension. *J. Mem. Lang.* 52, 144–169
41 Miller, G.A. and Chomsky, N. (1963) Finitary models of language users. In *Handbook of Mathematical Psychology* (Vol.II) (Luce, D.R. *et al.*, eds), John Wiley
42 Frazier, L. (1987) Syntactic processing: evidence from Dutch. *Nat. Lang. Linguist. Th.* 5, 519–560
43 Bader, M. and Lasser, I. (1994) German verb-final clauses and sentence processing: evidence for immediate attachment. In *Perspectives in Sentence Processing* (Clifton, C.J., ed.), Lawrence Erlbaum
44 Gibson, E.A. and Thomas, J. (1996) The processing complexity of English center-embedded and self-embedded structures. In *Proceedings of the NELS 26 Workshop on Sentence Processing (MIT Occasional Papers in Linguistics 9)* (Schutze, C., ed.), pp. 45–71, MIT Press
45 Tabor, W. *et al.* (2004) Effects of merely local syntactic coherence on sentence processing. *J. Mem. Lang.* 50, 355–370
46 Vosse, T. and Kempen, G. (2000) Syntactic structure assembly in human parsing: a computational model based on competitive inhibition and a lexicalist grammar. *Cognition* 75, 105–143
47 Jones, M. and Polk, T.A. (2002) An attractor network model of serial recall. *Cogn. Syst. Res.* 3, 45–55
48 Novick, J.M. *et al.* (2005) Cognitive control and parsing: reexamining the role of Broca's area in sentence comprehension. *Cogn. Affect. Behav. Neurosci.* 5, 263–281
49 Hale, J. (2003) The information conveyed by words in sentences. *J. Psycholinguist. Res.* 32, 101–123