Psychological Review
1992, Vol. 99, No. 1, 122-149

0033-295X/92/$3.00

A Capacity Theory of Comprehension: Individual Differences in Working Memory

Marcel Adam Just and Patricia A. Carpenter
Carnegie Mellon University

A theory of the way working memory capacity constrains comprehension is proposed. The theory proposes that both processing and storage are mediated by activation and that the total amount of activation available in working memory varies among individuals. Individual differences in working memory capacity for language can account for qualitative and quantitative differences among college-age adults in several aspects of language comprehension. One aspect is syntactic modularity: The larger capacity of some individuals permits interaction among syntactic and pragmatic information, so that their syntactic processes are not informationally encapsulated. Another aspect is syntactic ambiguity: The larger capacity of some individuals permits them to maintain multiple interpretations. The theory is instantiated as a production system model in which the amount of activation available to the model affects how it adapts to the transient computational and storage demands that occur in comprehension.

Working memory plays a central role in all forms of complex thinking, such as reasoning, problem solving, and language comprehension. However, its function in language comprehension is especially evident because comprehension entails processing a sequence of symbols that is produced and perceived over time. Working memory plays a critical role in storing the intermediate and final products of a reader's or listener's computations as she or he constructs and integrates ideas from the stream of successive words in a text or spoken discourse. In addition to its role in storage, working memory can also be viewed as the pool of operational resources that perform the symbolic computations and thereby generate the intermediate and final products. In this article, we examine how the human cognitive capacity accommodates or fails to accommodate the transient computational and storage demands that occur in language comprehension. We also explain the differences among individuals in their comprehension performance in terms of their working memory capacity. The major thesis is that cognitive capacity constrains comprehension, and it constrains comprehension more for some people than for others.

This article begins with a general outline of a capacity theory of language comprehension. In the second section we use the capacity theory to account for several phenomena relating individual differences in language processing to working memory capacity. In the third section we describe a computer simulation model that instantiates the capacity theory. In the final section we discuss the implications of capacity theory for other aspects of cognition besides language comprehension.

For the past 100 years, research on working memory (or short-term memory, as it used to be called) has focused on the storage of information for retrieval after a brief interval. A familiar example to illustrate the purpose of short-term memory is the storage of a telephone number between the time that the number is looked up in a directory and the time it is dialed. Short-term memory was typically thought of as a storage device, permitting a person to simply hold items until they were to be recalled. A related function attributed to short-term memory is its role as a stepping stone on the path to long-term memory, while information is being memorized through rehearsal or elaboration. Thus, working memory has long been implicated in both short-term and long-term storage.

A somewhat more modern view of working memory takes into account not just the storage of items for later retrieval, but also the storage of partial results in complex sequential computations, such as language comprehension. The storage requirements at the lexical level during comprehension are intuitively obvious. A listener or comprehender must be able to quickly retrieve some representation of earlier words and phrases in a sentence to relate them to later words and phrases. But storage demands also occur at several other levels of processing. The comprehender must also store the theme of the text, the representation of the situation to which it refers, the major propositions from preceding sentences, and a running, multilevel representation of the sentence that is currently being read (Kintsch & vanDijk, 1978; vanDijk & Kintsch, 1983). Thus, language comprehension is an excellent example of a task that demands extensive storage of partial and final products in the service of complex information processing.

Most recent conceptions of working memory extend its function beyond storage to encompass the actual computations

This work was supported in part by National Institute of Mental Health Grant MH 29617 and Research Scientist Development Awards MH-00661 and MH-00662 as well as a grant from the Andrew W. Mellon Foundation.

Sashank Varma played an important role in developing the capacity constrained interpreter and CC READER. We are also grateful to Jay McClelland, Maryellen MacDonald, Chuck Clifton, Mike Masson, Marilyn Turner, and Jonathan King for their constructive comments on earlier drafts of the manuscript.

Correspondence concerning this article should be addressed to Marcel Adam Just, Department of Psychology, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213.

themselves. The computations are symbolic manipulations that are at the heart of human thinking—such operations as comparison, retrieval, and logical and numerical operations. Of particular relevance are the processes that perform language comprehension. These processes, in combination with the storage resources, constitute working memory for language.

Construing working memory as an arena of computation was first advocated by Baddeley and Hitch (1974; Baddeley, 1986; Hitch & Baddeley, 1976), who constructed tasks that pitted the storage and processing aspects of comprehension against each other. They found that the ability to understand individual sentences rapidly and accurately decreased when listeners also had to encode several digits and later recall them. The trading relation between storage and processing suggested that the two functions were drawing on a common pool of resources. Thus, there are both conceptual and empirical reasons to express the dual roles of working memory within a single system.

It is important to point out the differences between our theory and Baddeley's (1986) conception of working memory. In our theory, working memory for language refers to a set of processes and resources that perform language comprehension. In Baddeley's theory, working memory has two components. One component consists of modality-specific storage systems, including the *articulatory loop*, a speech-based rehearsal buffer of fixed duration. The second component is the *central executive*. The central executive is the component that Baddeley addressed least in his empirical research and specified least in his theory; indeed, Baddeley termed the central executive "the area of residual ignorance" in his model (1986, p. 225). The working memory in our theory corresponds approximately to the part of the central executive in Baddeley's theory that deals with language comprehension. The working memory in our theory does not include modality-specific buffers, such as the articulatory loop.

Overview of the Theory

A major purpose of this article is to present a theoretical integration of the storage and processing functions of working memory in language comprehension. We present a computational theory in which both storage and processing are fueled by the same commodity: activation. In this framework, capacity can be expressed as the maximum amount of activation available in working memory to support either of the two functions.

In our theory, each representational element has an associated activation level. An element can represent a word, phrase, proposition, grammatical structure, thematic structure, object in the external world, and so on. The use of the activation level construct here is similar to its widespread use in other cognitive models, both symbolic (e.g., Anderson, 1983) and connectionist (e.g., McClelland & Rumelhart, 1988). During comprehension, information becomes activated by virtue of being encoded from written or spoken text, generated by a computation, or retrieved from long-term memory. As long as an element's activation level is above some minimum threshold value, that element is considered part of working memory, and consequently, it is available to be operated on by various processes. However, if the total amount of activation that is available to the system is less than the amount required to perform a comprehension task, then some of the activation that is maintaining old elements will be deallocated, producing a kind of forgetting by displacement. Thus, representations constructed early in a sentence may be forgotten by the time they are needed later on in the sentence, if enough computational activity has intervened. The activation is used not just for information maintenance—it is also the commodity that underlies computation. The computations are performed within a production system architecture in which productions manipulate symbols by modifying their activation levels. The most common manipulation occurs when a production increases the activation level of one of its action elements. Elements are added or deleted by changing their activation level appropriately.

The computations that are involved in language comprehension also can be expressed as manipulations of activation, as they typically are in connectionist models of comprehension (e.g., Cottrell, 1989; St. John & McClelland, 1990; Waltz & Pollack, 1985). In the current model, a production rule propagates activation from one element to another. The production has the source element as a condition, and the destination element as the action. The same production rule can fire repeatedly over successive cycles, reiteratively incrementing (or otherwise modifying) the target element's activation level, usually until it reaches some threshold. Consider an example in which the encounter with a grammatical subject of a sentence generates an expectation that a verb will occur. A proposition representing a grammatical subject is a source of activation for the proposition that a verb will be encountered. Thus, the rule-based processing typical of a symbolic system works in concert with the graded activation typical of a connectionist system.

Many of the processes underlying comprehension are assumed to occur in parallel. Thus, at the same time that the comprehender develops the expectation of encountering a verb, she or he could also be calculating other syntactic, semantic, and pragmatic features of the sentence. The theory proposes that all enabled processes can execute simultaneously, and generate partial products concurrently. However, if the number of processes (productions) is large or, more precisely, if the amount of activation they try to propagate would exceed the capacity, then their attempts at propagation are scaled back to a level that keeps the total activation within the maximum bound.

The trading relation between storage and processing occurs under an allocation scheme that takes effect when the activation maximum is about to be exceeded. Briefly, if the activation propagation on a given cycle of production firings would exceed the activation maximum, then both the activation propagated and the activation used for maintenance are scaled back proportionally to their current use. This scheme is analogous to imposing an across-the-board percentage budget cut if the spending quota (the amount of activation) is about to be exceeded. The scaling back of the activation propagated will increase the number of cycles required to bring an element to threshold, effectively slowing down the computation. The scheme implies that when the task demands are high (either because of storage or computational needs), then processing will slow down and some partial results may be forgotten. In sum, the time course and content of language processing within

this system depends on the capacity for storage and computation. When the task demands exceed the available resources, both storage and computational functions are degraded. We call this theory *capacity constrained comprehension.*

Processing of Sentences in Context

Because a text can contain an indefinitely large number of sentences whose storage could eventually consume any finite capacity, there must be countervailing mechanisms that reduce the storage demands. Some of these mechanisms selectively retain representations of only the most recent and most central clauses in an activated form, while dampening the activation level of other propositions from earlier sentences (Glanzer, Fischer, & Dorfman, 1984; Kintsch & vanDijk, 1978; vanDijk & Kintsch, 1983). Moreover, analogous mechanisms may selectively retain only the most relevant aspects of world knowledge in an activated form, while dampening the activation level of other knowledge that might be initially activated by the reading of the text (Kintsch, 1988). Storage demands are also minimized through the immediacy of processing, the tendency to semantically interpret each new word or phrase as far as possible when the word is first encountered, in contrast to a wait-and-see strategy that imposes additional storage demands (Carpenter & Just, 1983; Just & Carpenter, 1980). Finally, some lower levels of the hierarchical representations of language may be deactivated after suitable, higher level structures have been formed. For example, the representation of the syntactic structure of a sentence may be dispensable after a referential representation has been constructed. Consistent with this possibility are the empirical findings that indicate that relatively little of the lexical or syntactic information from earlier clauses remains accessible as a reader proceeds in a text (Huey, 1908; Jarvella, 1971; Sachs, 1967).

Processing a sequence of sentences with a finite working memory capacity is possible not only because the storage demands can be limited, but also because the context can provide some processing benefits. The stored context might facilitate the processing of the ensuing sentence by preactivating some concepts, relations, and schemas relevant to its comprehension (Sanford & Garrod, 1981; Sharkey & Mitchell, 1985). Thus, these mechanisms that facilitate processing and minimize the demands on storage may keep the overall demands on working memory manageable, even when an extended text is being processed.

In a later section of this article, we instantiate some parts of the theory in a computer simulation. The simulation is built using an architecture, called CAPS, that is a hybrid of a production system and an activation-based connectionist system. The new simulation is a modification of a computational model of the processing of successive words of a text during reading called READER (Thibadeau, Just, & Carpenter, 1982). The modified model reflects the assumption that comprehension processes are capacity constrained. Hence, the name for the new model is CC READER (Capacity Constrained READER). It is a conventional production system in its use of productions, a working memory, and the recognize–act governance of the flow of control. It is connectionist in that the productions reiteratively propagate activation from source elements to target elements, and all the productions that are satisfied on a given cycle can fire in parallel.

The constraint on capacity is imposed by limiting the total amount of activation that the system has available for maintaining elements in working memory and for propagating activation to other elements in the course of processing. Moreover, as we will describe below, the simulation's account of individual differences in working memory capacity for language is that subjects differ in the maximum amount of activation that they have available. Thus, the total amount of activation in the new CAPS system can express the conjoint constraint as well as any trade-offs that are made between storage and processing.

Individual Differences in Working Memory Capacity

A central thesis of this article is that the nature of a person's language comprehension depends on his or her working memory capacity. We will describe a number of recently uncovered systematic individual differences in reading comprehension that are related to working memory capacity for language. We propose that individuals vary in the amount of activation they have available for meeting the computational and storage demands of language processing. This conceptualization predicts quantitative differences among individuals in the speed and accuracy with which they comprehend language. In addition, it is capable of accounting for some qualitative differences among readers that we have observed.

We have described capacity as though it were an energy source that some people have more of than other people have. According to an analogy proposed by Kahneman (1973) in explaining his capacity theory of attention, a person with a larger memory capacity for language may be able to draw on a larger supply of resources, like a homeowner who can draw on more amperes of current than a neighbor, and can thus generate more units of cooling or heating. However, another account of individual differences is in terms of the efficiency of mental processes. To return to the electrical analogy, it may be that some homeowners have more efficient electrical appliances than others (appliances being the counterparts of mental processes), allowing them to do more with the current, such as produce more units. We can designate these as the *total capacity explanation* and the *processing efficiency explanation.* The two explanations are mutually compatible and the experiments described here do not attempt to discriminate between them. Our theory is expressed in terms of the total capacity account because of the finding reported below that performance differences among college student readers of different working memory capacities are small and often negligible when the comprehension task is easy, but large and systematic when the comprehension task is demanding. This result is easily explained by the total capacity hypothesis, because capacity limitations would affect performance only when the resource demands of the task exceed the available supply. The result is less easily explained in terms of individual differences in the efficiency of certain processes, because efficiency differences should manifest themselves regardless of the total demand. We return to this issue in the final discussion.

Effects of Capacity Constraints

In this section, we present new data and summarize existing research supporting the hypothesis that comprehension is constrained by working memory capacity. We address five capacity-related issues: (a) the influence of pragmatic cues on syntactic processes; (b) the time course of comprehending a complex syntactic embedding; (c) the maintenance of two representations of a syntactic ambiguity; (d) the effect of an external memory load on sentence comprehension processes; and (e) the ability to track long-distance dependencies within and between sentences. The intention is to examine how individual differences in working memory capacity constrain comprehension, producing both qualitative and quantitative performance differences. Before dealing with these five issues, we will describe the measurement of working memory capacity.

Assessing Working Memory Capacity

To assess individual differences in working memory capacity for language, we have used the *Reading Span* task (Daneman & Carpenter, 1980), which was devised to simultaneously draw on the processing and storage resources of working memory. The task requires subjects to read a set of unrelated sentences, such as: "When at last his eyes opened, there was no gleam of triumph, no shade of anger"; "The taxi turned up Michigan Avenue where they had a clear view of the lake." After reading these two sentences, the subject tries to recall the final word of each sentence, in this case, "anger" and "lake." The test determines the maximum number of sentences per set for which the subject can recall all of the final words. The largest set size for which the subject successfully recalls all of the final words for at least three of five sets is defined as his or her *reading span.* If the subject is correct on only two of the five sets, she or he is assigned a span that is intermediate between that size and the next lower one. Among college students, reading spans typically vary from 2 to 5.5 for sentences of this type. In most of the studies described below, *high span* individuals have spans of four words or more, *medium span* individuals have spans of three or three and a half words, and *low span* individuals are those with spans of less than three words. The label of *low span* is relative to our sample; our low span subjects are in the top half of the distribution of verbal ability in standardized tests such as the Scholastic Aptitude Test (SAT).

The rationale behind the test is that the comprehension processes used in reading the sentences should consume less of the working memory resources of high span readers. These readers would thus have more capacity left to hold the final words of the sentences. This measure of individual differences builds on the research of Baddeley and Hitch (1974) and Hitch and Baddeley (1976), which showed that language comprehension and simultaneous digit recall can draw on a shared resource. Thus, there can be a trading relation between the performance of the two tasks when they are done simultaneously, reflecting the constrained capacity.

The Reading Span task measure correlates highly with certain aspects of reading comprehension, such as the verbal SAT, with these correlations lying between .5 and .6 in various experiments (Daneman & Carpenter, 1980; Masson & Miller, 1983). The correlation between reading span and particular comprehension skills is even higher. For example, the ability to answer a factual question about a passage correlates between .7 and .9 with reading span in various studies. For easy texts, low span subjects read only marginally slower than high span subjects, but on particularly difficult portions of a text, low span subjects tend to be substantially slower than high span subjects. In contrast to the strong relation between reading span and various comprehension indices, passive short-term memory span performance (e.g., recalling a list of digits or unrelated words) is not significantly correlated with reading comprehension (Perfetti & Goldman, 1976).[1]

A listening version of the Reading Span task produces similar results to the reading version (Daneman & Carpenter, 1980). In general, individual differences in reading and listening are highly correlated for educated adults, such as college students, who are the primary target population in these studies. Thus, the individual differences of interest here are due to language processes, and are not restricted to reading processes. Although others have suggested that the Reading Span task may measure a more general factor than just working memory for language (Turner & Engle, 1989), the evidence on this point is not yet conclusive (Baddeley, Logie, Nimmo-Smith, & Brereton, 1985). In the present article, the term *working memory* refers to working memory for language.

Modularity of Syntactic Processing

Capacity constraints have the potential of creating boundaries between different types of processes when the total processing resources are insufficient to permit direct interaction between different processes. Interaction between processes, like other forms of computation, requires operational resources, such as storage of partial products and communication from one process to another. In the absence of resources sufficient to support interaction, two processes that have the requisite interconnectivity may fail to interact; that is, they may fail to influence each other's ongoing computations. But the boundaries created by capacity constraints are outcomes of resource limitations, and not of architectural barriers.

---

[1] The lack of correlation between the standard digit span task and reading comprehension indicates that the standard digit span task does not draw on the same resources as those used in most language comprehension tasks. The source of individual differences in standard span tasks is not clear (Lyon, 1977). One possibility is that such differences primarily reflect individual differences in the phonological store and the articulatory loop, an internal auditory–articulatory rehearsal process of fixed duration (Baddeley, 1986). The involvement of the articulatory loop in digit span performance has been implicated by cross-linguistic studies; in particular, the smaller digit span associated with the Welsh language has been attributed to the fact that Welsh vowels are longer and so Welsh digits take longer to subvocally rehearse than English digits (Ellis & Hennelley, 1980). Moreover, neuropsychological data suggest that impairments in digit span are not necessarily correlated with impaired language comprehension; some patients with very severely impaired digit span have relatively preserved sentence comprehension (Shallice, 1988). These neuropsychological data also support a dissociation between the standard digit span task and the mechanisms that are used in normal sentence comprehension.

Capacity constraints thus provide an interesting new perspective on the controversy concerning the modularity of language processing (Fodor, 1983; Garfield, 1989). A cognitive module is defined as a fast, domain-specific set of processes that is mandatory and informationally encapsulated. The interesting conjecture offered by Fodor is that the syntactic level of language processing is performed by a cognitive module. We propose that the postulated modularity of syntactic processing may be better explained as a capacity constraint that sometimes imposes informational encapsulation.

Informational encapsulation is the single most distinguishing property of a cognitive module. It refers to a module's activities and outputs being uninfluenced by certain classes of information that may exist elsewhere in the system. One of Fodor's (1983) examples of encapsulation is that when you push your eyeball with your finger you see motion, even though another part of your cognitive system has the information that the motion is not real. The information about the finger movement is apparently not available to or not used by the processes that interpret input from the retina. These perceptual interpretation processes are in some sense encapsulated from the finger motion information. Fodor also considered the syntactic processing of language to constitute a module that is encapsulated from nonsyntactic information. Fodor said "as things stand I know of *no* convincing evidence that syntactic parsing is ever guided by the subject's appreciation of pragmatic context or of 'real world' background" (p. 78). The rationale that Fodor offered for encapsulation is that input systems have to operate rapidly (without allocating time to consider all possible relevant information) and veridically, somewhat like a reflex. Despite the centrality of encapsulation in the debate about syntactic modularity, Garfield (1989) wrote, "Nevertheless, it [encapsulation] is one of the most difficult of the central properties to detect experimentally. . . . As Fodor concedes, and as the debate in this volume documents, encapsulation is a vexing issue in psycholinguistics" (p. 4). The information from which syntactic processing is encapsulated is ultimately brought to bear at some time on the final interpretation of a word, phrase, or sentence. Fodor's conjecture is that the information does not affect syntactic processing itself, but instead affects some later integrative process.

The theory we propose reinterprets syntactic encapsulation as an issue of capacity rather than of architecture. According to our view, people with small working memories for language may not have the capacity to entertain (keep activated and propagate additional activation from) nonsyntactic information during the syntactic computations, or at least not to the degree that the nonsyntactic information can influence the syntactic processing. In this view, the syntactic processing of a person with a small working memory is encapsulated only by virtue of a capacity constraint, not an architectural constraint. Individuals with a large working memory capacity may be more able to keep both syntactic and nonsyntactic information activated, and hence their syntactic processing would be more likely to be influenced by the nonsyntactic information. Thus, some people's syntactic processing might seem more modular than others. The degree of modularity would depend on working memory capacity for language, not on some structural separation between modules. But any variation in modularity across subjects destroys the value of the concept of modularity. To replace it, we offer the concept of capacity for interaction, which can differ among individuals.

First, we briefly summarize the previous and new empirical findings, and then we report them in more detail. The empirical support for our position comes from a study in which we examined individual differences in encapsulation of syntactic processing. The study used a task that had previously provided the strongest support for the modularity of syntax (Ferreira & Clifton, 1986). Ferreira and Clifton constructed a task in which the reader could avoid being led down a garden path only by making immediate use of nonsyntactic information. Their surprising result was that readers were led down the garden path, despite the presence of prior disambiguating information about the proper resolution of a syntactic ambiguity. That disambiguating information was nonsyntactic, so its lack of influence on the syntactic processing was attributed to the encapsulation of the syntactic module. We repeated that experiment, separating subjects of different reading spans, and replicated Ferreira and Clifton's result in the case of the low span subjects. However, just as our theory predicted, the high span subjects did take the nonsyntactic information into account in initially interpreting the syntactic ambiguity. In effect, the syntactic processing of the high span subjects was not modular, but interactive.

Ferreira and Clifton (1986) examined the reading time on sentences like Sentence 1, "The evidence examined by the lawyer shocked the jury"; and Sentence 2, "The defendant examined by the lawyer shocked the jury." Because these sentences omit the complementizers and the verb ("who was" or "that was") of the relative clauses, the sentences are called *reduced relative clauses.* The initial part of Sentence 2 "The defendant examined" is temporarily ambiguous between the main verb interpretation (which could have continued as "The defendant examined the courtroom") and the eventually correct interpretation as a relative clause. The study varied whether or not a pragmatic cue, the animacy of the head noun, signalled the correct interpretation of the initial ambiguous portion of the sentences. In Sentence 1, the initial noun is inanimate and, consequently, an implausible agent of the following verb. If the pragmatic information concerning inanimacy influences the parsing decision, readers should be more likely to interpret the verb as a reduced relative verb than as a main verb. Moreover, they might expect that an agent will be specified later in the sentence. When an agentive phrase, such as "by the lawyer," occurs, it should be no surprise and it should present no particular processing difficulty. In contrast, in Sentence 2, "the defendant" is a plausible agent of the verb "examined." Consequently, readers are likely to interpret the verb as a main verb, and begin their trip down the garden path. The subsequent agentive phrase is inconsistent with the main verb interpretation, so the encounter with the *by* phrase should present a processing difficulty. The garden path effect can be measured by comparing the reading time on "by the lawyer" in sentences like 1 and 2, to determine whether or not the inanimacy of the noun "evidence" alleviates any of the surprise on the encounter with the *by* phrase.

The striking result that Ferreira and Clifton (1986) obtained was that readers still spent a long time on the *by* phrase when

the head noun of the sentence was inanimate. This result suggested that readers were led down the garden path. In fact, the first-pass reading times on the *by* phrase (as indicated in eye fixation data) were no shorter in the sentences containing an inanimate head noun than in sentences with an animate head noun. In other words, the inanimacy of a head noun like "evidence" appeared not to penetrate the syntactic analysis of the verb "examined." Ferreira and Clifton concluded that the lack of effect of inanimacy occurred because syntactic analysis is a modular cognitive process that is informationally encapsulated. They reasoned that even if there were information available to influence the syntactic analysis, that kind of information would not be used on the first pass of the syntactic processing because syntactic processing, according to theories like Fodor's (1983), is impermeable to other sources of information.

Our experiment was similar to Ferreira and Clifton's (1986) in most respects, except that the data were separated for subjects of different reading spans—40 high span readers (spans of 4.0 or higher) and 40 low span readers (spans of 2.5 or lower). In addition, we improved some of the stimulus sentences so that the grammatical subject could not be interpreted as an instrument; this eliminated some of the sentences used by Ferreira and Clifton, such as "The car towed. . . ." Each of 20 sentences, interspersed with a variety of filler sentences, was presented in random order. In addition to sentences with reduced relative clauses, our experiment (like Ferreira and Clifton's) presented sentences with unreduced relative clauses, such as the following, syntactically unambiguous sentences: Sentence 3, "The evidence that was examined by the lawyer shocked the jury"; and Sentence 4, "The defendant who was examined by the lawyer shocked the jury." Across four groups of subjects, each sentence occurred in each of the four forms.

While the subject read sentences on a graphics monitor, his or her eye fixations were recorded using an ISCAN Model RK-426 Pupil/Corneal Tracking System, and a VAXstation 3200 computed the point of regard every 16.7 ms. Each screen display consisted of a neutral, introductory sentence, followed by the target sentence, followed by a true–false comprehension question that the reader answered by pressing one of two buttons.

The primary analysis focused on the first-pass reading times on the *by* phrase and on the initial verb and the main verb of the target sentence. Subsequent analyses examined any reading beyond a first pass (typically regressions back to the *by* phrase). The analysis used only trials in which the subject fixated on the first verb for at least 150 ms, and then on the *by* phrase for at least 150 ms.

The main result was a difference between the high and low span subjects in their reaction to an inanimate noun. Inanimacy decreased the first-pass reading time on the *by* phrase for high span subjects, but not for the low span subjects. This result suggests that high span readers were sensitive to the pragmatic cue during the first-pass syntactic analysis. In the case of the reduced relative clauses, the presence of an inanimate noun reduced the first-pass reading time on the *by* phrase by 75 ms for the high span subjects, but did not reduce it for the low span subjects, as shown in the bottom right side of Figure 1. Similarly, in the case of the unreduced relative clauses, the presence of an inanimate noun reduced the first-pass reading time on the *by*

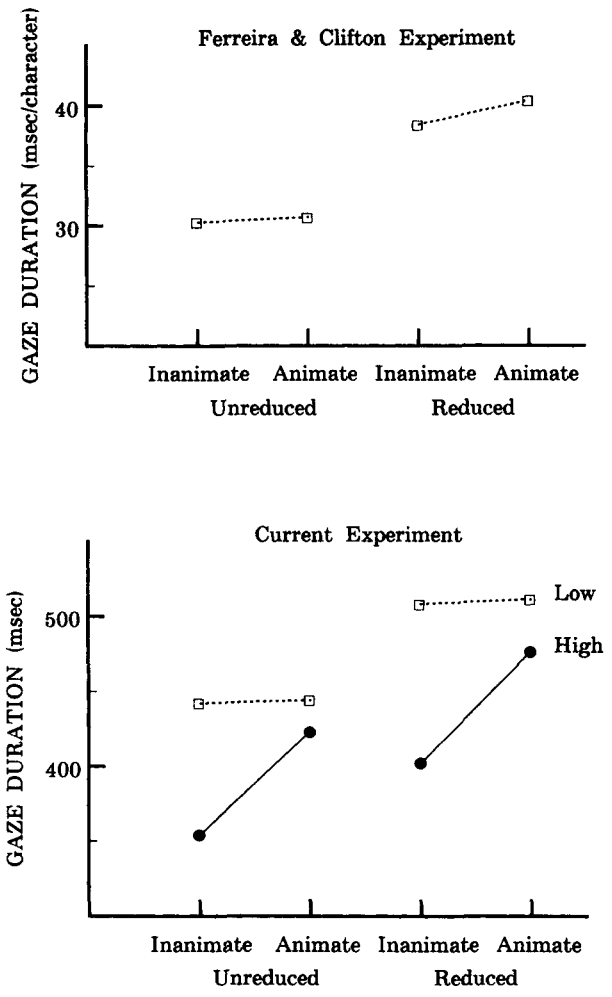

Figure 1. First-pass reading times on *by* phrase for the untested subjects of Ferreira and Clifton (1986) on the top and for high and low span subjects in the present study on the bottom. (The slope for the high span subjects between the inanimate and animate conditions indicates that their processing is faster if the grammatical subject of the sentence is inanimate; in contrast, the lack of difference between the inanimate and animate conditions indicates a negligible influence for Ferreira and Clifton's subjects and the low span subjects in the present experiment.)

phrase by 69 ms for the high span subjects, but did not reduce it for the low span subjects. This pattern produced a reliable interaction of animacy and span group, $F(1, 66) = 5.36$, $p < .025$, $MS_e = 15{,}487$. The results for the low span group parallel those of Ferreira and Clifton (1986); their data were reported as reading time per character and they are plotted in the top part of Figure 1.

Not surprisingly, the first-pass reading time on the *by* phrase was shorter (by 58 ms on average) for both span groups if the relative clause was not reduced, $F(1, 66) = 14.13$, $p < .01$, $MS_e = 16{,}334$. There was no interaction between span group and reduction. Taken together, these results suggest that the two groups of subjects make similar use of the explicit syntactic cue

(i.e., the expanded relative clause), whereas only the high span subjects make use of the pragmatic cue of inanimacy.

Averaged over all four conditions, the high span subjects had marginally shorter first-pass reading times on the *by* phrase (by 62 ms), $F(1, 66) = 3.76$, $p < .06$, $MS_e = 70{,}308$.

This pattern of results is most easily explained in terms of a capacity difference between the two groups of subjects, such that only the high span subjects have the capacity to take the pragmatic information into account. The modularity explanation does not fit this pattern of results, unless one postulated that the syntactic processing of low span subjects is modular, and the syntactic processing of high span subjects is interactive. But modularity was construed as a hypothesis about a universal functional architecture, a construal that is violated by a finding of individual differences. In contrast, the capacity constrained comprehension model simply postulates that interaction requires capacity, so only those subjects with greater capacity have the resources to support interaction. The specific interaction in this case is the ability to represent, maintain, and use the inanimacy cue during the processing of the syntactic information.

The subsequent reading times on the *by* phrase after the first pass show the same general pattern of results as the first-pass reading, except that the critical interaction of animacy by span is no longer reliable, just as expected (as shown in Table 1). This result indicates that all subjects can ultimately take inanimacy into account, but only the high span subjects can take it into account during the first pass. The reading time on the *by* phrase after the first pass was shorter for both span groups if the relative clause was not reduced, $F(1, 66) = 7.95$, $p < .01$, $MS_e = 7{,}798$. There was no interaction between span group and reduction. These times were marginally shorter for the high span subjects than for the low span subjects, $F(1, 66) = 3.63$, $p < .07$, $MS_e = 15{,}140$.

For completeness, we also report the reading times on the two verbs that bounded the *by* phrase ("examined" and "shocked" in the sample sentence) in Table 1. The result that is important to our thesis is that there are no interactions between the two span groups and any other variable, for either the first or the second verb, nor is there a reliable main effect of span group in either case. In the case of the first verb, there is a main effect of both animacy and reduction, in that readers take less time on the first verb if the preceding context imposes more constraint (inanimacy or an unreduced structure) than if it imposes less constraint (animacy or a reduced structure).

Table 1
Reading Times (in Milliseconds) on Bypass, First Verb, and Second Verb

Group	Unreduced inanimate	Unreduced animate	Reduced inanimate	Reduced animate
	Subsequent reading time on *by* phrase			
Low span	55	56	87	100
High span	22	48	49	83
	First-pass reading time on first verb			
Low span	234	253	249	301
High span	212	232	265	284
	First-pass reading time on second verb			
Low span	287	268	308	295
High span	252	269	264	278

The general pattern of results can be viewed from another useful perspective—the effects of various cues to sentence structure. One cue to the complex relative clause structure is the complementizer and auxiliary verb ("that was") that occur in the unreduced relative clause. This is a reliable and valid cue used by both high and low span subjects; when this cue is present, the reading times on the first verb and on the *by* phrase are shorter than if it is absent. The inanimacy of the head noun is another cue, which indicates that the head noun will not be the agent of the verb, but will instead occupy some other thematic role in the main clause and in any subordinate clauses. Only the high span subjects use this cue in their first-pass reading; their reading time on the first verb and on the *by* phrase is shorter if the inanimacy cue is present and longer if it is absent. In the case of the high span subjects, the two cues make approximately additive contributions to performance. The low span subjects show no effect of the inanimacy cue on the first-pass reading of the *by* phrase, indicating the insensitivity of their syntactic parsing process to this cue, as predicted by Ferreira and Clifton's (1986) instantiation of the syntactic modularity hypothesis. Completely unpredicted by the modularity hypothesis, but predicted by the capacity constrained comprehension model, is the high span subjects' use of this cue. Thus the syntactic encapsulation that some subjects exhibit is better explained in terms of a capacity constraint. The results contradict the view that syntactic processing is informationally encapsulated by an architectural barrier.

Processing Complex Embeddings

Capacity constraints, as measured by working memory capacity, should produce quantitative differences among individuals in the time course and accuracy of their processing. Moreover, these quantitative differences should be most apparent when the sentence or task is particularly capacity demanding. Several recent studies have obtained the predicted quantitative differences, that is, high span readers were both faster and more accurate in their comprehension of difficult sentences (King & Just, 1991). These studies demonstrated that much of the quantitative difference that results from working memory capacity can be localized to those portions of the sentence that are particularly capacity demanding.

The classic example of a syntactic structure that makes large demands on working memory capacity is a sentence containing a center-embedded object-relative clause, such as Sentence 5, "The reporter that the senator attacked admitted the error." It is called an object-relative clause because the head noun is the object of the relative clause. Subjects who hear such a sentence and then try to paraphrase it make errors approximately 15% of the time (Larkin & Burns, 1977). There are two processing demands that in combination make such sentences difficult to understand, but the nature of the demands is not critical to our argument. Briefly, one source of difficulty is that the embedded clause interrupts the main clause, requiring that the interrupted clause representation either be retained in working

memory or be reactivated at the conclusion of the embedded clause. The second source of difficulty is that one of the syntactic constituents ("reporter," in the example above) is the subject of the main clause, but it is also the grammatical object of the embedded clause. Associating a single concept with two different roles simultaneously poses a difficulty in language comprehension (Bever, 1970; Sheldon, 1974). In contrast, it is easier to comprehend a sentence containing a subject-relative clause, such as Sentence 6, "The reporter that attacked the senator admitted the error." Like Sentence 5, this example has an interruption of the main clause, but in this case, the head noun plays the same role in both clauses (Holmes & O'Regan, 1981). Thus, clause interruption together with the assignment of nonparallel roles to the same entity combine to make object-relative sentences particularly difficult to understand.

The processing demands imposed by object-relative sentences provide a way to examine how working memory constrains the comprehension of normal (albeit difficult) sentences. The model predicts that each person's working memory capacity will determine how difficult each sentence type will be for him or her. All readers should find the object-relative sentences more difficult to comprehend than subject-relative sentences; more important for the thesis of this article, readers with lower capacity should have relatively more difficulty with object-relative sentences. King and Just (1991) measured the word-by-word reading times as subjects read sentences containing either an object-relative or subject-relative clause and then answered a question to assess the accuracy of their interpretation.

Reading time was assessed with the self-paced moving window paradigm, in which the sentences are initially displayed with dashes replacing the alphabetic characters (Just, Carpenter, & Woolley, 1982). Each time the subject presses a small hand-held microswitch lever, the characters of the next word ahead of the current point of advance replace the corresponding dashes, and the previously displayed word is replaced by dashes. This method yields a reading time for each word of the text.

The results confirmed several important predictions of the theory, as shown in Figure 2. First, there were large individual differences in reading times, and these differences were primarily localized to the object-relative sentences which are shown on the right-hand panel of the figure. The fact that the intergroup differences were larger on the more demanding object-relative sentences suggests that working memory constraints are manifested primarily when processing demands exceed capacity. Second, the word-by-word reading times localized the processing difficulty of object-relative clauses to the point at which the critical syntactic information becomes available. All three groups of subjects showed a selective increase in reading time at the verb of the embedded relative clause ("attacked") and at the verb of the main clause ("admitted"). The increase was larger for subjects with smaller spans, so the three curves diverge in the right-hand panel of Figure 2 precisely at the location where the processing load is at its peak.

The subjects with lower spans not only took longer to read the more complex sentences, but their comprehension accuracy was also poorer than that of higher span subjects. The accuracy of the low span subjects in answering true–false comprehension questions, such as "The senator admitted the error," was 64%, compared with 85% for the high span subjects (mid-span subjects' comprehension was 83%). The combination of reading time and comprehension-accuracy results shows that readers with lower reading spans have poorer comprehension, even though they may spend considerably more time processing in the syntactically critical area of the sentence. These results demonstrate systematic and localized individual differences in the comprehension of difficult syntactic structures, differences that are modulated by working memory capacity.

The near absence of differences among the groups for undemanding sentences suggests that performance differences cannot be entirely attributed to the speed of some particular operation, a hypothesis that underlay much of the interesting research on individual differences in cognition in the 1970s (e.g., Hunt, Lunneborg, & Lewis, 1975). For example, suppose that the individual differences in reading times were due only to differences in the speed of lexical access. Then there should be a substantial reading time difference between high span and low span subjects even on the syntactically simpler sentences, which in this particular experiment happen to contain exactly the same words as the more complex sentences. But the large reading time differences occurred only on the more complex object-relative sentences, showing that the speed of lexical access cannot be more than a minor component of the individual differences.

Age-Related Differences

Syntactic constructions that make large demands on the working memory capacity of college students are the very types of constructions that produce age-related performance decrements in elderly people. One reason that working memory is implicated in the age-related changes is that the performance decrements are not general, nor do they appear to be attributed to the loss of some specific linguistic computation. For example, older adults (65–79 years of age) show relatively greater deficits than do younger adults when they must make an inference that requires integrating information across sentences (Cohen, 1979). Making the inference requires the storage of information from previous sentences concurrently with the processing of ensuing sentences, placing a large demand on working memory. The deficit is not general, because the older subjects have much less of a disadvantage when the comprehension test probes for verbatim information from the stimulus sentence.

Working memory was directly implicated in age-related changes in a task that required subjects to repeat sentences of various syntactic types (Kemper, 1986). Elderly adults (aged 70–89) were impaired (compared with young adults aged 30–49) at imitating sentences whose syntax made large demands on working memory. The elderly adults had particular difficulty in imitating sentences containing a long sentence-initial embedded clause, as in "The cookies that I baked yesterday for my grandchildren were delicious." They correctly imitated them only 6% of the time, as compared with 84% for the younger adults. The elderly adults also had some difficulty imitating a sentence containing a long sentence-final embedded clause (42% correct). By contrast, the elderly adults had no difficulty in imitating sentences with short embedded clauses. The age-re-

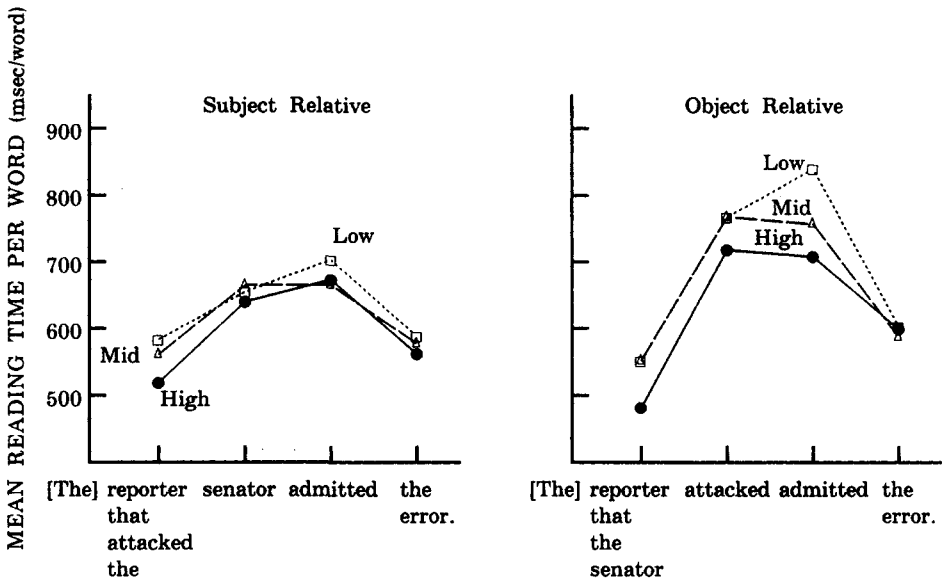

Figure 2. Reading time per word for successive areas of subject- and object-relative sentences, for high, medium (Mid), and low span subjects. (The differences among the span groups are larger for the more difficult object-relative construction, which is the more complex sentence. The differences are particularly large at the verbs, which are points of processing difficulty that are expected to stress working memory capacity. The reading times for parenthesized words are not included in the plotted points.)

lated decline in the ability to imitate sentences is largest in cases in which the processing of the main syntactic constituent is interrupted by the processing of a long embedded constituent. This type of construction requires that the initial portion of the main constituent be retained in working memory while the embedded constituent is processed under the memory load, and then the stored portion must be made accessible again when its final portion is being processed. In addition to this age-related difference in imitation performance, Kemper found a corresponding age-related difference in spontaneous production (Kemper, 1988; Kynette & Kemper, 1986).

Thus, the decline in language performance in the elderly is focused on sentences whose syntax makes large demands on working memory. In general, the individual operations of language processing show little evidence of decline with age when the total processing load is small. However, at times of high demand, the total performance does decline, indicating an age-related decrease in the overall working memory capacity for language.

Syntactic Ambiguity: Single Versus Multiple Representations

Another facet of language that could generate demand for additional resources is syntactic ambiguity, particularly in the absence of a preceding context that selects among the possible interpretations. If comprehenders were to represent more than one interpretation of an ambiguity during the portion of a sentence that is ambiguous, this would clearly demand additional capacity. However, the existing data and the corresponding theories are in disagreement about the processing of syntactic ambiguities. A comprehender encountering an ambiguity might select a single interpretation (Frazier, 1978; Just & Carpenter, 1987; Marcus, 1980), or she or he might retain two alternative interpretations until some later disambiguating information is provided (Gorrell, 1987; Kurtzman, 1985). These two schemes for dealing with syntactic ambiguity have been posed as opposing (and mutually exclusive) alternatives. However, in a series of experiments, we found that both positions could be reconciled by postulating individual differences in the degree to which multiple representations are maintained for a syntactic ambiguity (MacDonald, Just, & Carpenter, in press).

In the model we advance, multiple representations are initially constructed by all comprehenders on first encountering the syntactic ambiguity. Each of the multiple representations is assumed to have an activation level proportional to its frequency, its syntactic complexity, and its pragmatic plausibility. The important new postulate of our theory is that the working memory capacity of the comprehender influences the duration (i.e., intervening text) over which multiple syntactic representations can be maintained. A low span reader does not have sufficient capacity to maintain the two interpretations, and soon abandons the less preferred interpretation, which results in a single-interpretation scheme. In contrast, a high span reader will be able to maintain two interpretations for some period.

The full set of results is too long to present here, because it includes reading times and comprehension rates on unambiguous sentences and two resolutions of ambiguous sentences (MacDonald, Just, & Carpenter, in press, for details). However, we can present the critical data that support the central claim, which makes an unintuitive prediction. In the survey of capacity effects presented above, a greater capacity produces better

performance in all ways that have been measured; the high span readers did not have to trade anything measurable away to read difficult sentences faster and comprehend them better. However, maintaining the multiple interpretations of a syntactic ambiguity is so demanding that it produces a performance deficit, which is shown only by the high span readers.

The critical data do not concern tricky garden path sentences, but on the contrary, they concern the most common syntactic structures in English, such as Sentence 7: "The experienced soldiers warned about the dangers before the midnight raid." This sentence is temporarily ambiguous, as can be demonstrated by considering an alternative resolution, such as Sentence 8: "The experienced soldiers warned about the dangers conducted the midnight raid." The syntactic ambiguity involves the interpretation of "warned" as either a main verb or as a past participle in a reduced relative construction. In Sentence 7 this ambiguity is resolved with the period at the end of the sentence.

The surprising result is that only high span subjects show any effect of the ambiguity in Sentence 7, as evaluated by comparison with the processing time on a control, unambiguous sentence that contains exactly the same words, except for the verb, such as Sentence 9: "The experienced soldiers spoke about the dangers before the midnight raid." Sentence 9 is unambiguous because the verb "spoke" can only be interpreted as a main verb and not as a past participle. For high span subjects, ambiguous sentences like Sentence 7 take longer than their unambiguous counterparts, particularly near or at the end of the sentence, where the ambiguity is resolved. In contrast, low span subjects show a small and generally unreliable effect of the ambiguity. Furthermore, in terms of absolute reading times, the high span subjects take longer than the low spans on reading ambiguous sentences, but not on unambiguous ones like Sentence 9. The reading time differences between the temporarily ambiguous main verb sentences and their unambiguous controls are shown in Figure 3. The sentences were presented in a self-paced, word-by-word, moving window paradigm. In the case of the main verb sentences, the three regions for which the reading time differences are plotted are indicated with brackets for the sample sentence:

"The experienced soldiers

[warned about the dangers] [before the midnight] [raid.]"

Region: 1 2 3

Precisely the same words (with the exception of the verbs) enter into each region for both the ambiguous and unambiguous conditions. Across a series of experiments, the high span readers showed a reliable ambiguity effect at and near the end of the sentence. These results strongly support the contention that high span subjects are maintaining two representations of the syntactic ambiguity, and they pay a concomitant price of slowing down their processing.

We have already commented on one remarkable facet of the results, namely that high span subjects pay a measurable price for maintaining two interpretations. The other remarkable facet is that sentences like Sentence 7, "The experienced soldiers warned about the dangers before the midnight raid," are not garden path sentences. Even though the sentence is temporarily ambiguous, it is ultimately given the more frequent resolution, that "warned" turns out to be the main verb of the sentence. The high span readers have not been led down a garden path. They represented an alternative, less likely path as they walked down the main path. The low span subjects also have not been led down the garden path. They represented just one alternative, and it turned out to be the correct one. They maintained only one interpretation, and thus did not have to pay any significant costs due to the sentence's temporary ambiguity. However, even high span readers do not hold onto multiple interpretations indefinitely; the ambiguity effect for these subjects can also be eliminated if the ambiguous region is greatly lengthened.

Of course, the correct interpretation of the temporary ambiguity could have turned out differently, as in Sentence 8: "The

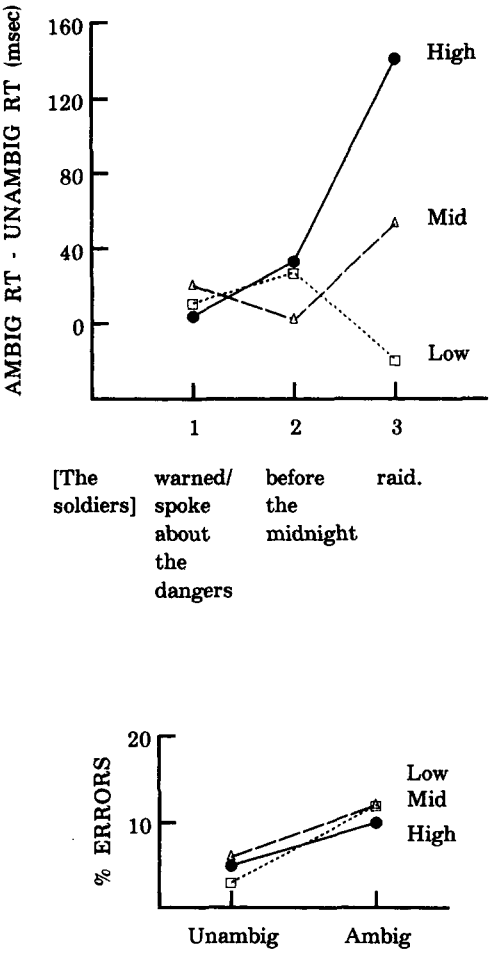

Figure 3. The differences between the ambiguous (AMBIG) and unambiguous (UNAMBIG) sentences in the average reading time (RT) per word for the three regions of the sentences for the high, medium (Mid), and low span subjects. (The high span subjects take additional time on sentences with structurally ambiguous verbs, such as "warned," than those with unambiguous verbs, such as "spoke." This result is consistent with the hypothesis that high span subjects maintain the ambiguity in working memory for a longer period than do medium or low span subjects. The error rates are indicated in the bottom panel.)

experienced soldiers warned about the dangers conducted the midnight raid." In that case, there is a distinct advantage in having represented both alternatives. In this sentence, the correct interpretation is the less likely alternative, the reduced relative interpretation that corresponds to "The experienced soldiers *who were* warned about the dangers conducted the midnight raid." The high span subjects are more likely than the lows to have the correct interpretation available, and this is reflected in their better comprehension of this sentence, as assessed by their accuracy in answering a comprehension question, like "Did someone tell the soldiers about dangers?" Of course, the high span subjects pay the cost of maintaining both interpretations, as indicated by their greater reading time on this sentence than on a control sentence. The low span subjects do not do the extra maintenance for any length of time and have no extra cost to pay as they read along; however, they are led down the garden path, and after encountering an unexpected verb, their resulting comprehension is often near chance level. We will present a more detailed processing account of the results in a later section that describes the capacity constrained simulation model.

The comprehension errors converge with this analysis, showing that the capacity constrained on-line processing is also a determinant of the ultimate comprehension of the sentence. The comprehension error rates are higher for ambiguous sentences, even if they are resolved with the main verb interpretation. Of course, the error rates are higher for the less likely relative clause interpretation.

A further control study showed that the ambiguity effect persists even when relative clause sentences were not included, so that the effect could not be attributed to a strategy that developed in the presence of this construction. Moreover, the ambiguity effect cannot be attributed to differences between the verbs in the ambiguous and unambiguous conditions, rather than to the syntactic ambiguity. The ambiguity effect is essentially eliminated if the same verbs are used, but proper nouns replace the common nouns, as in Sentence 10: "Captain Thompson warned about the dangers before the midnight raid." The proper noun eliminates the reduced relative interpretation and, consequently, the ambiguity effect. However, in the same control experiment, the ambiguity effect was replicated for sentences like Sentence 7. Finally, the ambiguity effect occurs not only when the sentences are presented in isolation, but also when the sentences are embedded in short paragraph contexts.

These findings demonstrate that individual differences in working memory capacity can produce different parsing outcomes for syntactically ambiguous sentences, which in turn can lead to differences in comprehension. The model unifies the previously disparate single and multiple representation models of parsing and points to the adaptability of the parsing mechanisms to the availability of memory resources.

Extrinsic Memory Load

The availability of working memory resources has often been experimentally manipulated through the introduction of an extrinsic memory load, such as a series of words or digits that are to be retained during comprehension. The extrinsic load could consume resources simply by virtue of being maintained in working memory or by virtue of rehearsal and recoding processes that compete for resources (Klapp, Marshburn, & Lester, 1983). As the extrinsic load increases, one or more facets of performance degrades such as the reading rate or the ability to recall the load items (Baddeley & Lewis, reported in Baddeley, 1986; Baddeley, Eldridge, Lewis, & Thomson, 1984).

The maintenance of an extrinsic load interferes with sentence comprehension, which suggests that they compete for the same resources. An extrinsic load condition was included in the syntactic complexity experiment described earlier (King & Just, 1991), involving sentences with subject-relative and object-relative clauses, such as "The reporter that the senator attacked admitted the error." When the subjects were required to retain one or two unrelated words while reading the sentence, their ability to answer a subsequent comprehension question was less than in a condition in which there was no extrinsic load.

Figure 4 shows the comprehension accuracy of the high and low span subjects when there was no additional load, compared with a condition in which the subjects were maintaining either one or two words. A comparison of the overall performance for the two panels confirms that, as expected, accuracy is generally higher for the subject-relative sentences (left-hand panel) than for the linguistically more complex object-relative sentences (right-hand panel). The accuracy is particularly impaired for the low span readers, who have less capacity for the complex computations entailed in processing the object relatives. Indeed, King and Just (1991) found that half of the low span subjects had comprehension rates for the object-relative sentences that were indistinguishable from chance (66% in this study). Given the low level of comprehension in the absence of a load, the extrinsic load only slightly decreases comprehension

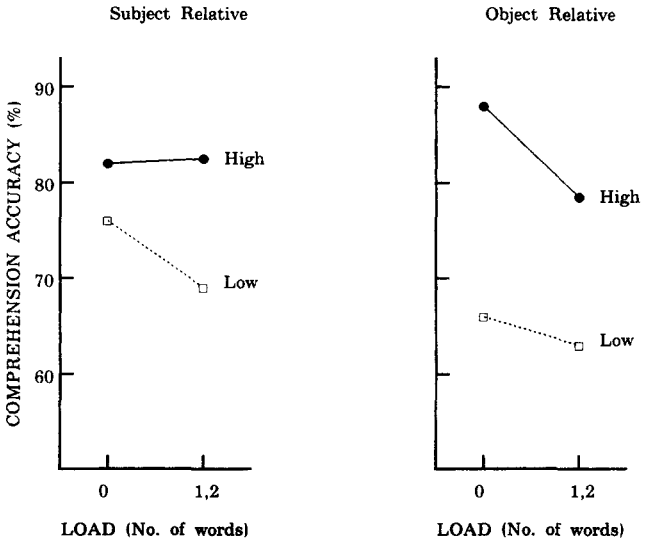

Figure 4. Comprehension accuracy for subject-relative and object-relative sentences in the absence or presence of an extrinsic memory load for high and low span reading groups. (Comprehension accuracy is lower for the low span subjects, for the more complex object-relative clause sentences, and when subjects are simultaneously retaining a load.)

accuracy. The effect of the load is much more apparent for the high span subjects who, in the absence of a load, have sufficient capacity to comprehend the object-relative sentences with much greater accuracy than the low span subjects. In sum, comprehension accuracy decreases both when there is less capacity because of a subject variable (as in the span contrast), because of additional linguistic complexity (as in the contrast between the two sentence types), or because of an additional extrinsic memory load (as in the contrast between a load of 0 words and 1 or 2 words). Such data are consistent with the hypothesis that some aspect of maintaining an extrinsic load competes for the resources used in the comprehension process.

Not only does load maintenance interfere with comprehension, but comprehension can reciprocally interfere with the maintenance of an extrinsic load. This effect was demonstrated in a modified version of the Reading Span task. The sentence-final words in this task constitute a concurrent load because the words from previous sentences must be maintained during comprehension of the current sentence. Thus, it is possible to manipulate the complexity of the sentences and examine its effect on recalling the sentence-final words. If the sentences in the set are linguistically more difficult, then the recall of the sentence-final words decreases (Carpenter & Just, 1989). The easy sentences (e.g., Sentence 11: "I thought the gift would be a nice surprise, but he thought it was very strange") primarily differed from the more difficult sentences (e.g., Sentence 12: "Citizens divulged that dispatching their first born was a traumatic event to face") in the presence of more common and concrete words, as well as in syntactic complexity. The subjects, all college students, were given sets of two, three, or four sentences to read and then asked to recall the sentence-final words.

The number of sentence-final words recalled was generally lower if the subjects had read a set of difficult sentences than if they had read easy sentences. This was particularly the case if the number of sentence-final words exceeded the reader's span, as assessed by the Reading Span task. For example, high span readers could retain and recall almost all of the words of a four-sentence set of easy sentences, but they could recall only three items if the set was composed of hard sentences. However, the difficulty of the sentences had little influence on their recall for sets of three or fewer sentences. Similarly, medium span readers could recall almost all of the words of a three-sentence set if the sentences in the sets were easy, but they could recall only two items if the set was composed of hard sentences. The pattern of recall results, shown in Figure 5, demonstrates that the processing of difficult sentences interferes with the ability to retain and recall an extrinsic load, particularly if the subjects are operating close to their working memory capacity limit.

These overall effects do not indicate which comprehension processes may be more or less vulnerable to the effect of an extrinsic load. There is some evidence that low-level processes may be less vulnerable to the effect of an extrinsic load than more conceptual processes. The time attributed to word encoding is not affected by the presence or absence of a memory load (Carpenter & Just, 1989), nor is the speed of retrieving and judging whether an instance (e.g., pine) is a member of a prespecified category (e.g., tree) (Baddeley, 1986). However, even as low-level a process as lexical access may be sensitive to working memory constraints. An attribute of gaze duration that is related to lexical access (an increase in gaze duration with decreasing normative word frequency) showed sensitivity to extrinsic load, at least for individuals with larger working memory capacities (Carpenter & Just, 1989). It is commonly assumed in the literature on automaticity that perceptual processes and highly practiced processes are less dependent on attention and may be less vulnerable to any kind of interference from a competing task (Cohen, Dunbar, & McClelland, 1990). If this assumption is applied to language processing, one might correspondingly predict that higher level comprehension processes would be much more disrupted by a competing task, such as maintaining an extrinsic load, than would lower level processes. However, as yet, there are inadequate data against which to examine this hypothesis. What the data do suggest is that comprehending a sentence and storing an extrinsic load draw on shared resources, so that performance declines when the two tasks conjointly exhaust the resources.

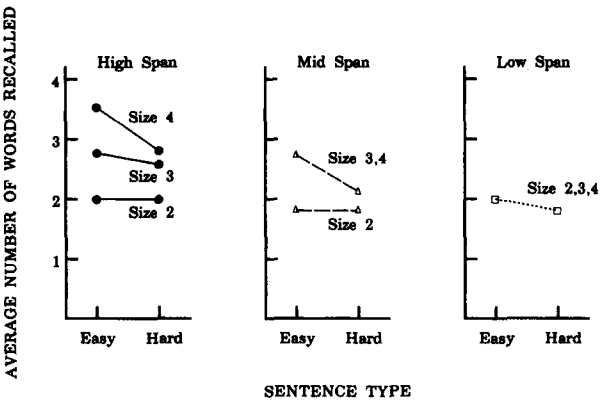

Figure 5. The average number of load words that are recalled after a subject reads a set of sentences that are either hard, because they contain difficult vocabulary, or easy. (Recall primarily decreases when the subject is at or near his or her capacity and the sentence is hard.)

Distance Effects

An intrinsic part of language comprehension is the ability to interrelate information that comes from different constituents, such as clauses or sentences. Consequently, there is a need to retain information over time and intervening computations. Working memory provides the resources to store information from preceding constituents while simultaneously providing the computational resources to process ensuing constituents. The greater the distance between the two constituents to be related, the larger the probability of error and the longer the duration of the integration processes (e.g., Chang, 1980; Jarvella, 1971).

Text distance effects. A text distance effect has been found with a variety of constructions and relations. For example, sentences that contain pronouns take longer to read if other clauses or sentences intervene between the sentence and the earlier referent, presumably because of increased time to search for the referent (Clark & Sengul, 1979). Sentences that contain a term referring to a member (e.g., trombonist) of a category that

is specified earlier (e.g., musician) take less time to read if the sentences are successive than if one or more sentences intervene between the sentence containing the superordinate term and the one containing the subordinate (Carpenter & Just, 1977; Lesgold, Roth, & Curtis, 1979). Similar distance effects have been found for sentences that are related by causality (Keenan, Baillet, & Brown, 1984), as well as pronominal reference, adverbial reference, and connectives (Fischer & Glanzer, 1986; Glanzer, Dorfman, & Kaplan, 1981; Glanzer et al., 1984). These distance effects have been interpreted as suggesting that comprehension involves representing the relation between the current phrase or clause and earlier information (Just & Carpenter, 1980; Kintsch & vanDijk, 1978). This relating takes less time if the earlier, relevant information is still available in working memory. In contrast, if the earlier, relevant information is no longer activated, then the relating process will require either searches of long-term memory and more constructive inferences, or there will be a failure to relate the new information to the earlier information.

Readers with larger working memory capacities are able to maintain more information in an activated state, and hence are better at interconstituent integration, as will be described below. In particular, there is a strong relation between a subject's reading span and the text distance over which he or she can successfully find an antecedent for a pronoun (Daneman & Carpenter, 1980). This result was found in two experiments that manipulated the number of sentences that intervened between the last mention of the referent and the pronoun. Readers with larger reading spans were more accurate at answering comprehension questions that asked the identity of the person referred to by the pronoun. More precisely, the maximal distance across which a reader could correctly assign the pronoun was well predicted by his or her reading span, as shown in Figure 6. One explanation that can be ruled out is that higher span readers are simply more skilled at selecting important referents for storage. If this were the only factor operating, then the performance should not decline monotonically with distance. Intervening text that does not reinstate a referent causes forgetting, with more forgetting by low span than high span subjects. Thus, working memory capacity and individual differences in this capacity are clearly implicated in the integration processes that are used in constructing a coherent referential representation of the text.

Anomaly detection over a text distance by children. Greater facility in integrating information over distances in a text also characterized young readers (7 and 8 years old) who had larger working memory spans, as assessed by a modified version of the Reading Span task, involving digit repetition and recall (Yuill, Oakhill, & Parkin, 1990). Two groups of children were selected who had similar levels of word decoding skill (accuracy) and vocabulary knowledge, but differed in overall language comprehension measures. They were given a comprehension test that required them to integrate two pieces of information in a story. For example, one piece of information was an adult's reaction that was inconsistent with a norm, such as blaming a boy for sharing sweets with his little brother. The resolution (for example, information that the little brother was on a diet) was either adjacent or separated from the anomaly by two sentences. The two groups of children were similar in their ability to comprehend passages when the information was adjacent, indicating that both groups of children understood the task and could perform it comparably in the absence of a large working memory demand. However, the children with smaller working memory capacities performed worse than those with larger capacities when the two sources of information were separated by some text.

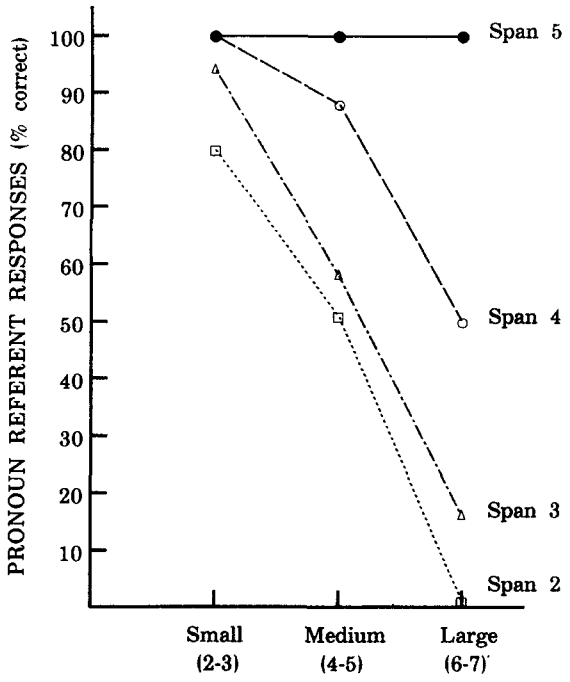

Figure 6. Correct interpretation of a pronoun as a function of its distance from the preceding anaphoric reference for high, medium, and low span reading groups. (The subjects with the highest span [5 words] have no difficulty interpreting a pronoun, even if its referent occurred six or seven sentences before. By contrast, the subjects with the lowest span [2 words] cannot interpret such pronouns and are only moderately successful if the referent occurred two or three sentences earlier.)

The results for children converge with the individual differences data for adults to suggest that working memory capacity is related to the ability to retain text information that facilitates the comprehension of subsequent sentences. These studies indicate that individuals with larger capacities are more successful in integrating information over a distance in a text.

Summary of Results

Across these five aspects of comprehension, we have described qualitative differences among readers (in the permeability of their syntactic processing to pragmatic information, and in their representing one versus two interpretations of a syntactic ambiguity) as well as quantitative differences (in the time course of comprehension and in the accuracy of comprehension). Comprehension performance generally declines with an

intrinsic memory load (such as retaining information across successive sentences of a text) or an extrinsic one, with greater declines for lower span readers. Reading slows down at just that point in a sentence that introduces a computational demand, and slows down more for low span than high span subjects. Although lower span readers typically show a disadvantage compared with high span readers in both reading time and errors, there are also situations in which high span readers show an apparent disadvantage in the costs associated with maintaining two representations of a syntactic ambiguity. The constraints on every person's capacity limit the open-ended facets of comprehension, so that a reader or listener cannot generate every possible forward inference, represent every interpretation of every ambiguity, or take into consideration every potentially relevant cue to an interpretation.

Simulation Model

To examine the theoretical sufficiency of the capacity hypothesis, we have simulated aspects of the experiments described above using the CAPS architecture, which is a hybrid of a production system and a connectionist system. As in a connectionist system, activation is propagated from source (condition) elements to target (action) elements, but the propagation is performed by the productions. The productions can operate in parallel with each other and propagate activation reiteratively over several processing cycles, until the target element reaches some threshold. The constraint on capacity is imposed by limiting the total amount of activation that the system has available for maintaining elements in working memory and for propagating activation to other elements in the course of processing. Before describing how the activation constraint is applied, we will describe some of the key properties of CAPS/READER.

1. Associated with each working memory element is a real number called its activation level, which represents the element's strength. An element satisfies a production's condition side only if its activation level lies above some threshold specified by the production or by convention.

2. Most working memory elements are propositions of the form (concept :relation concept) or (concept [implicit :isa] concept). These elements can form a network.

3. Production firings direct the flow of activation from one working memory element (called the *source*) multiplied by a factor (called the *weight*) to another working memory element (called the *target*).

4. One processing cycle is defined as the matching of all productions against working memory and the consequent parallel firing of all satisfied productions.

5. Long-term knowledge consists of a declarative database separate from working memory.

One unconventional aspect of the parser, which we call CC READER, is that the number of processing cycles that it takes to process each word depends on the amount of activation that is available. If activation is plentiful, then a given production may be able to activate its action elements to threshold in a single cycle. But if storage or processing demands conjointly exceed the activation maximum, then the production would increment the activation level of its action elements rather gradually, until the level reached threshold. For example, in a conventional production system parser, there might be a production like, "If the word *the* occurs in a sentence, then a noun phrase is beginning at that word." In CC READER, the corresponding production would be, "If the word *the* occurs in a sentence, then increment the activation level of the proposition stating that a noun phrase is beginning at that word." If there were a shortage of activation, then the CC READER production would fire reiteratively over several cycles until the proposition's activation level reached a threshold. CC READER's smaller grain size of processing permits evidence (activation) for a particular result to accumulate gradually, with potential for input from several sources.

The constraint on the total amount of activation in the system conjointly limits how much activation can be propagated per cycle and how many elements can be maintained in working memory. Given a certain maximum on the amount of activation that is available, there are many points along an isoactivation curve that can be adopted. At one extreme, the brunt of an activation shortage could be borne by the storage function, so that at times of peak demand there would be a lot of forgetting of partial and final products in working memory, but processing (particularly the time it takes to compute something) would remain unchanged. At the other extreme, the brunt could be borne by the processing function, so that at times of peak demand, processing would slow down but there would be no forgetting. Limiting the amount of activation available per cycle slows down processing by requiring more cycles of propagation to raise an element's activation to a given level. The current implementation chooses an intermediate point in this trading relation. Any shortfall of activation is assessed against both the storage and processing, in proportion to the amount of activation they are currently consuming. For example, if the quota is a total of 36 units of activation, and the old elements are consuming 32 for maintenance, and the current processing cycle requires 16 for propagation (for a total budget request of 48 units), then the shortfall of 12 units is assessed proportionally. Thus, maintenance and storage each receive $^{36}/_{48}$ or $^{3}/_{4}$ of their need, or 24 units of activation for maintenance and 12 for propagation by the productions. The effect of exceeding the quota is both forgetting and a slowing down of the processing. In this scheme, the degree of constraint on processing (e.g., comprehension) depends on the maintenance demands of the moment, and vice versa.

There are several implications of this scheme. First, decrementing the activation levels of old elements is a form of continuous forgetting by displacement. This is unlike conventional displacement models, which posit one element displacing another from a limited number of storage locations. In contrast, this model posits that the activation that is used to maintain old elements is drawn away by the action of the productions' incrementing the activation levels of other elements. Thus, when the activation demands exceed the constraint, there will be gradual forgetting of old information (or old partial products) with each new cycle of processing that exceeds the activation quota. Second, when the number of partial products is small (e.g., early in a task) then the forgetting due to displacement will be less than when there are many partial products (e.g., late in a task that requires storage of partial products). Furthermore, the constraint does not apply until the demand for activation exceeds

the quota. Thus the effects of the constraint will arise at different times for different people. The manifestation of the constraint is some combination of forgetting (decrease in activation of old elements) and a slowdown of processing (more cycles of activation will be required for an element's activation level to reach a given threshold).

Negative activation (e.g., lateral inhibition or suppression) does not play as large a role in this system as it does in a conventional network model. In a conventional network, many elements initially receive some activation increment, with a gradual selection of the correct elements among them, using mechanisms like lateral inhibition. In CC READER, a form of lateral inhibition occurs when a production increments the activation level of one element and simultaneously decrements the level of a collateral element.[2] In addition, much of the decrementing of the unselected elements occurs through the application of the activation maximum. As activation is propagated to the selected elements, there is that much less activation left for maintaining unselected elements, and so the activation level of the unselected elements is decremented. What negative activation is propagated is not counted in toward the fixed activation quota. Also, the activation constraint does not apply to long-term memory elements that are in working memory, or to instances of long-term knowledge. For example, the knowledge that *the* is a determiner or the knowledge that a particular instance of the word *the* is a determiner is considered long-term knowledge, and is not subject to the activation constraint.

CC READER deals primarily with the syntactic level of processing of sentences containing embedded clauses. The motivation for this focus is that many of the new results we have recently obtained on individual differences and capacity constraints pertain to the syntactic level of processing. We do not mean to imply that the syntactic level is in any sense the most important or central level of language understanding, even though it plays a central role in the reported experiments and hence in this simulation model. The model also involves some semantic analysis (assigning case roles) and lexical access (activating word meanings to threshold). CC READER parses only single sentences. As a matter of expedience, we have not inserted a referential level of processing, so that the model does not construct a referential representation of what it reads, and hence has no way of relating one sentence to another. The main goal of the simulation model is to demonstrate the sufficiency of a psychologically plausible understanding system whose performance varies under different capacity constraints. The variation in the model's performance is intended to resemble the variation among individuals of different capacities and the variation within individuals comprehending under different processing loads.

CC READER can parse each of the sentence types shown in the Appendix (which includes reduced relative sentences, sentences with an embedded object-relative clause, and passive sentences), as well as similar sentence structures that can be formed by adding combinations of adverbs, adjectives, and prepositional phrases to each of the sentences in the table. As CC READER parses a sentence word by word, it incrementally constructs a representation of the sentence that is sufficient for answering *wh-* and *yes–no* questions. Moreover, the parser simulates the real-time processing profiles of high or low span subjects (depending on its activation quota) as they read the successive words of the various sentences. For example, in the case of syntactic ambiguities, the parser performs either like a high span or a low span subject, depending on how much activation it has available to propagate per cycle. Thus it demonstrates the sufficiency of our theoretical account, including the capacity constrained individual differences.

The parser consists of 57 productions which may be sorted into four categories (initialization, lexical access, syntactic parsing, semantic analysis). The syntactic productions can be further subdivided, as indicated in Table 2. Some of the categorizations are somewhat arbitrary, because some productions perform functions that fall under two or more of the above categories. A brief description of the function of the productions follows.

One production initializes the entire system before a sentence is read, and a second production reinitializes before each new word of the sentence is input. A third production, which roughly corresponds to the perceptual encoding of a word form, fetches the next word of the sentence when all the processing of the preceding word is completed (i.e., when all of the new propositions engendered by the preceding word have been activated to their target level).

Lexical Access

When the next word in the sentence is fetched, then one of the lexical access productions inserts a token (a copy) of the corresponding lexical entry from the lexicon into working memory, as well as inserting a flag indicating that lexical access is taking place. The lexicon, a database that is considered a part of long-term memory, contains information about words, such as their possible parts of speech. The base activation level of an entry in the lexicon is proportional to the logarithm of the corresponding word's normative frequency in the language; when a lexical entry is copied into working memory, it is given an initial activation level equal to its parent activation level in the lexicon.

A second production reiteratively increments that token's activation level until it reaches a threshold. Because the initial activation level is proportional to the logarithm of the word's frequency, the number of iterations required to activate the

[2] In our theory, the decrementing of the activation of collateral elements is a type of suppression, the inverse of activation for maintenance (Tipper, 1985; Tipper & Cranston, 1985). Deactivation plays an important role in many activation-based computational schemes in which several alternative competing elements are activated for some purpose, and only one of them is ultimately selected. This is also the way deactivation is used in CC READER. The significance of suppression for a capacity theory is that people who are more efficient at suppression might effectively gain capacity by freeing up activation. A recent probe response-time study has provided some suggestive evidence in the domain of lexical ambiguity. This study found that almost a second after reading a sentence that ended with an ambiguous word, better comprehenders had suppressed the word's irrelevant meaning, whereas poorer comprehenders showed evidence that it was still active, even though the context made it clear that it was irrelevant (Gernsbacher, Varner, & Faust, 1990). In this way, efficient suppression mechanisms could increase capacity.

Table 2
Categories of Productions in CC Reader

Category	Number	Subcategory
Initialization	3	Initializing system and getting next word
Lexical access	3	Accessing lexicon
Syntactic parsing	10	Parsing noun phrases
	16	Parsing verb phrases
	7	Handling transitions between subjects, predicates, and direct objects
	12	Handling embedded clauses
Semantic analysis	6	Handling verb voice and agent–patient assignment

entry to threshold is logarithmically related to word frequency, replicating the word frequency effect in human gaze durations (Carpenter & Just, 1983). A paraphrase of this production is "If you are doing lexical access of the current word, then propagate a fixed increment of activation to the corresponding word token." A third production removes the lexical access flag when the token's activation level reaches a threshold.

Syntactic Parsing

These productions implement a connectionist, activation-based parser. The grammar that the parser uses is conventional (largely adapted from Winograd, 1983), and fairly limited, with a focus on handling embedded clauses that modify the sentence subject. Figure 7 graphically depicts some of the main parsing paths that are taken in the processing of the sentences in some of the experiments described in this article.

The grammar and the productions that operate on it can be usefully compared to an augmented transition network (ATN), another procedure for parsing sentences. In an ATN, the nodes correspond to syntactic constituents and the arcs linking the nodes correspond to the syntactic and sequential properties of constituents. An ATN's syntactic parsing of a sentence consists of tracing a single path through the network, moving from one node to the next by choosing the arc whose conditions are satisfied by the next word or words. Although the current grammar resembles an ATN, there are several important differences between them. First, in CC READER's grammar, more than one arc can be traversed when leaving a node because two interpretations may be maintained concurrently if capacity permits. Second, traversing an arc can require several iterations of an action (to activate the next node to threshold), rather than a one-shot arc traversal. Third, nonsyntactic information can influence a syntactic decision if capacity permits.

The main syntactic constituents that CC READER parses include noun phrases, verb phrases, prepositional phrases, and clauses. Its general goal at the syntactic level is to recognize instances of these types of constituents, the transitions among them, and the relations among them. In addition, the syntactic productions attempt to recognize which of the constituents make up the subjects and predicates of clauses.

When the beginning of a new constituent is encountered, the goal of constructing a representation of that constituent is created. Similarly, if one constituent indicates that a mating constituent will be forthcoming (e.g., the presence of a subject indicates that a predicate will be encountered), then the goal of representing the mate will be activated. The activation level to which a goal is initially incremented is proportional to its a priori importance, so that an important goal is not likely to be completely forgotten even if the activation constraint applies. For example, the goal of representing a subject or a predicate of a clause, which might have to be maintained during the processing of many intervening words, is activated to a high level because of its importance in the syntactic representation of a sentence. In contrast, the goal of representing a syntactically less crucial constituent, such as a prepositional phrase, is given a lower activational level.

The activation of a constituent is incremented to its target level over cycles by using a form of the delta rule used in McClelland's (1979) cascade model. Each activation increment is an increasing function of the difference of the proposition's current level and its target level. Thus, early increments are large, but as the difference between current and target activation levels decreases, the size of the increments also decreases.

Semantic Analysis

These productions do a limited case role analysis, focusing on agents and recipients of actions as they occur in active or passive clauses. For example, one of the productions can be paraphrased as "If you see a prepositional phrase starting with *by* containing an animate head noun, and modifying a passive verb, then activate the proposition that the head noun of the prepositional phrase is the agent of the verb." These case-role productions operate in collaboration with those that determine which constituents are the subjects and predicates.

CC READER is fairly word oriented, using word boundaries to segment its major processing episodes. As each ensuing word of a sentence is encoded, all the enabled productions continue to fire until they have all run their course. Then the next word is encoded, and so on. This scheme uses immediacy of processing in that the interpretation of each new word proceeds as far as possible when the word is first encountered. The notable exception to immediacy occurs in the processing of syntactic ambiguity under conditions of high capacity, as described.

The performance of CC READER can be compared to that of human subjects in a number of the situations that bear on the processing of syntactic complexity, syntactic ambiguity, and syntactic modularity (permeability to nonsyntactic information). The parsing model generally performs very much like the human subjects. It slows down in similar places, and has similar error distributions. The model's most novel property is that its performance changes with its activation maximum, and the model's variation in performance mirrors the differences between low and high span subjects. In the simulations described later, CC READER is given a low activation quota (constant across the three studies) to simulate low span subjects, and a high activation quota (again constant across the three studies) to simulate high span subjects.

Syntactic Ambiguity

The total amount of activation that is available determines whether CC READER will maintain one or two interpretations

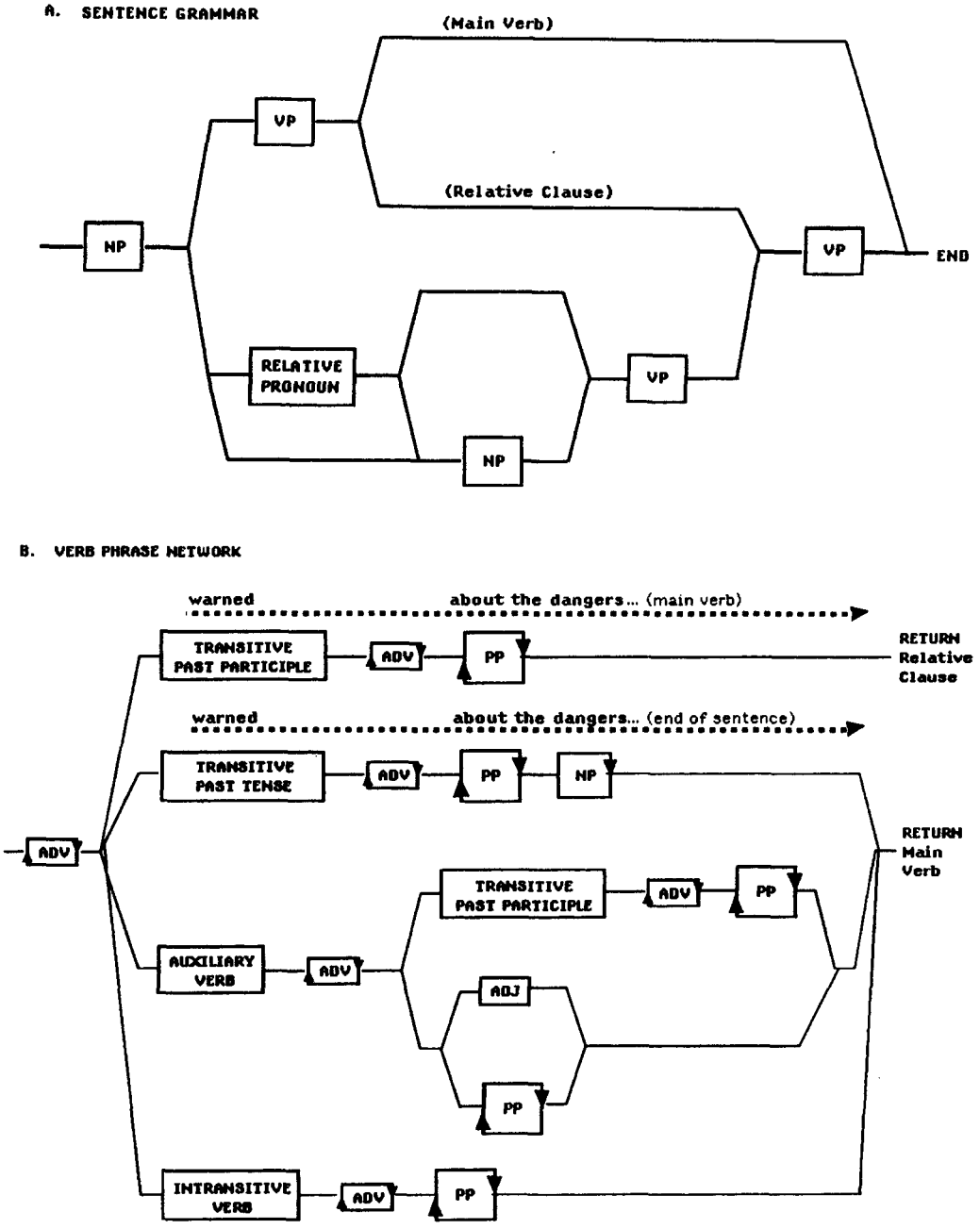

Figure 7. The sentence grammar, on the top, and the verb phrase network, on the bottom, that are used by the syntactic productions. (The ambiguity between main verb sentence and the reduced relative clause sentence is indicated in the verb phrase network by indicating that the phrase "warned about the dangers" is compatible with two branches of the network. VP = verb phrase; NP = noun phrase; ADV = adverb; PP = prepositional phrase; ADJ = adjective.)

in the ambiguous region of a syntactic ambiguity. First consider the case of the model processing a fragment, like "The experienced soldiers warned. . .," under a high activation maximum (simulating a high capacity reader). Both representations of the syntactic ambiguity are activated when the ambiguity is first encountered, each with an activation level proportional to its relative normative frequency. The two interpretations of the fragment correspond to two alternative paths (indicated by dashed lines) in the grammar depicted in Figure 7, namely the two upper paths in the verb phrase network. In the case of a verb form like "warned," which can be either a transitive past participle or a transitive past tense, both representations are activated. However, the past tense interpretation is activated to a higher level than the past participle, by virtue of the former's

higher normative frequency. As subsequent words of the sentence are encountered, they are interpreted in terms of both paths, as long as they fit both paths. In effect, the high capacity permits some degree of nonimmediacy, as the model is permitted to wait and see which of the two interpretations turns out to be correct. Recall that the disambiguation in the case of the main verb sentences used in the experiment is the end of the sentence. When the disambiguation is encountered, it fits only one path in the grammar. At that point, extra cycles relative to the unambiguous case are consumed in incrementing the activation level of the appropriate representation while deactivating the inappropriate interpretation. If the verb is unambiguous (e.g., "spoke"), then only one path (the second from the top in the verb phrase network) is followed.

Initially, the simulation of the low span subjects (low maximum activation) resembles the high span simulation, in that both interpretations are represented. However, in the case of the low maximum, the supply of activation is inadequate to maintain the secondary interpretation for more than one or two words beyond the ambiguity, so its activation level declines to a level below which it no longer satisfies the condition sides of productions. In the absence of a competing interpretation, the supply of activation is adequate to continue to process the main verb interpretation without further effect of the ambiguity up to and including the end of the sentence.

Thus, in the case of the low span simulation, there is little or no difference in the number of cycles per word between ambiguous and unambiguous sentences, either in the ambiguous region or at the encounter with the disambiguation, as shown in Figure 8. CC READER simulates the difference between processing ambiguous and unambiguous sentences by low and high span subjects, depending on its activation maximum.

Syntactic Complexity

In processing a center-embedded sentence, CC READER's performance profile (the number of cycles spent on each of the four sentence regions) resembles that of the human subjects in several respects, as shown in Figure 9. The most important similarity is that the model spends more time in the demanding regions if its activation maximum is smaller than if it is greater, simulating low and high span subjects respectively. Furthermore, the model's cycles (as a function of its activation maximum) differ more in the demanding regions of the sentence than in the undemanding regions, as is the case for the human subjects of different capacities. Additional points of similarity are that the model spends more cycles on the main verb and on the last word of the subordinate clause than on other parts of the sentence, and that like the subjects, the model takes longer on object relatives than on subject relatives.

Each of the effects in the model's performance can be explained in terms of the demand for activation and the consequences of the demand not being met. For example, the reason that the model takes extra time at the two verbs of object-relative sentences is that more productions fire during the computations on the verb than at other points in the sentences, because the verbs are grammatically related to a large number of other constituents, such as subjects and direct objects. The larger number of productions invoked at the verbs than at other

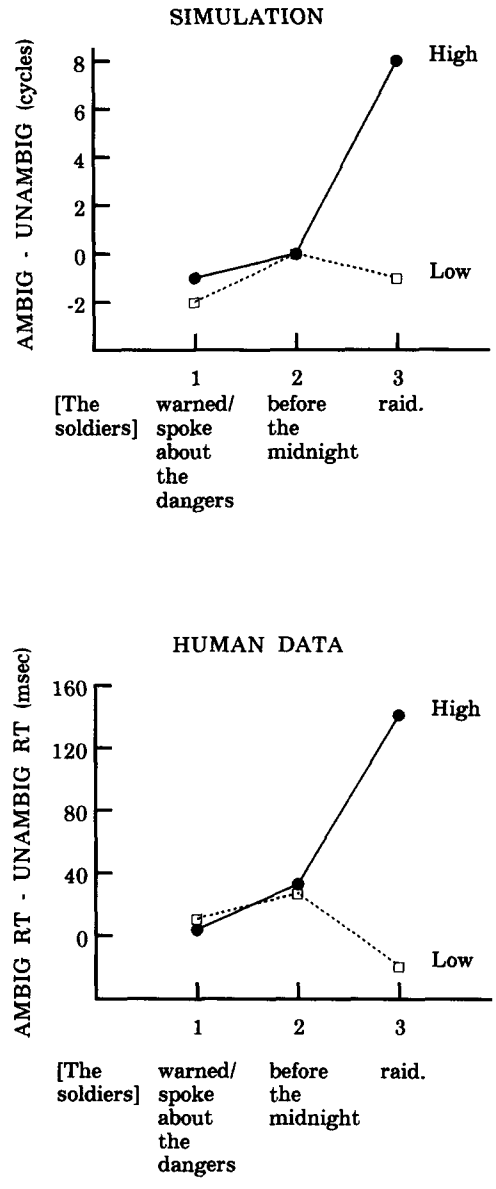

Figure 8. The top graph presents the difference in the number of cycles needed to process the ambiguous (AMBIG) and unambiguous (UNAMBIG) sentences when the simulation, CC READER, is operating with more working memory capacity, to simulate the high span subjects, or less, to simulate the low span readers. (The bottom graph presents the human data for comparison with the simulation. RT = reading time.)

points in the sentence demand extra activation, and moreover the activation pool is likely to have depleted somewhat by the time the sixth or seventh word of the sentence is encountered. Thus the productions that operate on the verbs require more cycles to accomplish their function. In the case of the object-relative sentences, the agent–patient computations that occur at the second verb require that a new agent be associated with the embedded clause, putting another additional demand on activation. The activation shortage is exacerbated in the case of the

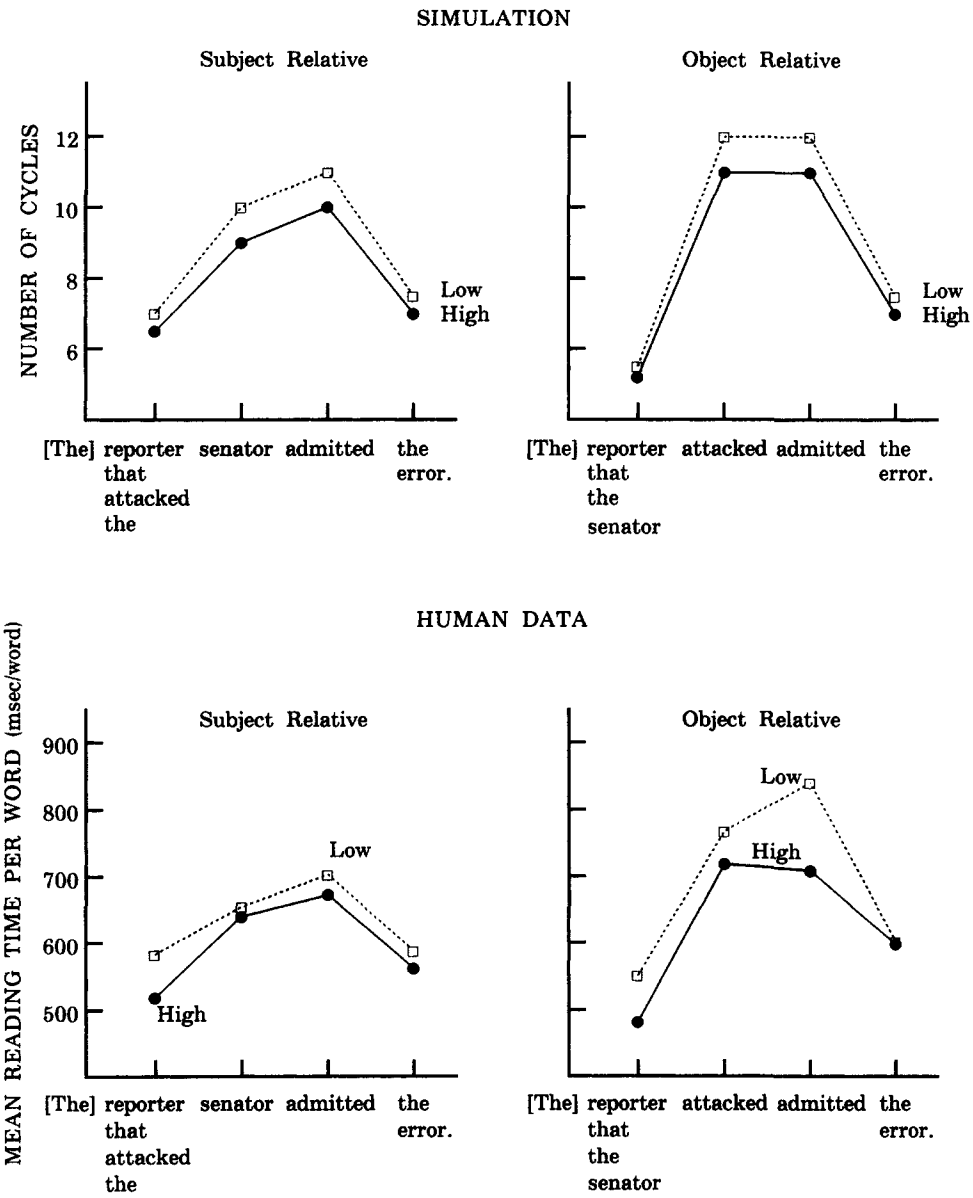

Figure 9. The number of cycles expended on various parts of the subject-relative sentences (on the left) and object-relative sentences (on the right) when the simulation, CC READER, is operating with more or less working memory capacity. (The bottom graph presents the human data for comparison with the simulation.)

low span simulation, which has a smaller activation maximum. The words that follow the verbs evoke fewer productions, so even though the activation maximum applies during their firing, they complete their execution in a smaller number of cycles (compared with the verb processing), and the high–low difference becomes smaller.

In summary, a simulation that varies the amount of activation available for simultaneously computing and maintaining information accounts for the reading time differences between high and low span subjects dealing with syntactic complexity provided by center-embedded clauses.

Pragmatic Influence on Syntactic Processing

The simulation demonstrates how the contribution of a pragmatic cue to syntactic analysis depends on an adequate supply of activation. First consider the simulation of the high span subjects (in which the activation maximum is relatively high) in processing the sentences containing reduced relative clauses. The inanimacy information encoded with the subject noun (e.g., "evidence") is still in an activated state when the verb is being processed, and this information is used to select between the two interpretations of the verb (past tense vs. past partici-

ple) that are initially activated. The inanimacy of the subject noun favors the selection of the past participle and the deactivation of the past tense interpretation. From that point on, the sentence fragment is no longer ambiguous, so only one interpretation is maintained and it is the appropriate one. There is no unexpected resolution of an ambiguity at the *by* phrase for the high-capacity simulation that has been provided with the pragmatic cue.

Consider the case with an animate subject noun (e.g., "defendant"), which provides no information to resolve the ambiguity of the verb. Because the high span simulation has adequate capacity, both interpretations of the sentence fragment can be maintained in an activated state. But, as previously described, there is a cost associated with maintaining two interpretations, and this cost is incurred primarily at the disambiguation at the *by* phrase. Thus the number of cycles spent on the *by* phrase is greater in the animate reduced than in the inanimate reduced condition, as shown by the solid line in the right-hand side of the upper panel of Figure 10. Human high span subjects behaved similarly, as shown in the bottom panel of Figure 10.

Now consider the simulation of the low span subjects in processing the sentences with reduced relative clauses. Even if the subject noun is inanimate, the activation shortfall prevents the inanimacy information from being maintained long enough to be of use in disambiguating the verb. Furthermore, there is inadequate activation to maintain both interpretations of the verb, so only the more frequent main verb interpretation is maintained. The encounter with the *by* phrase reveals the inconsistency with the main verb representation, requiring a cycle-consuming reparsing of the ambiguous verb. So the number of cycles consumed on the *by* phrase is relatively large in the inanimate reduced condition. Similarly, the number is equally large in the animate reduced condition (in which there is no inanimacy cue), as shown by the dashed line in the right-hand side of the upper panel of Figure 10. The human low span subjects behaved similarly, as shown by the dashed line in the bottom panel of Figure 10.

The simulation model also provides a reasonable fit to the processing of the unreduced sentences, in both the high span and low span cases. The simulation of the low span subjects accurately predicts the absence of an inanimacy effect in the unreduced condition, and a lower cycle count than in the reduced conditions, as shown by the dashed lines in the left-hand sides of Figure 10. The simulation of the high span subjects benefits from the inanimacy information even in the unreduced condition, as do the human high span subjects.

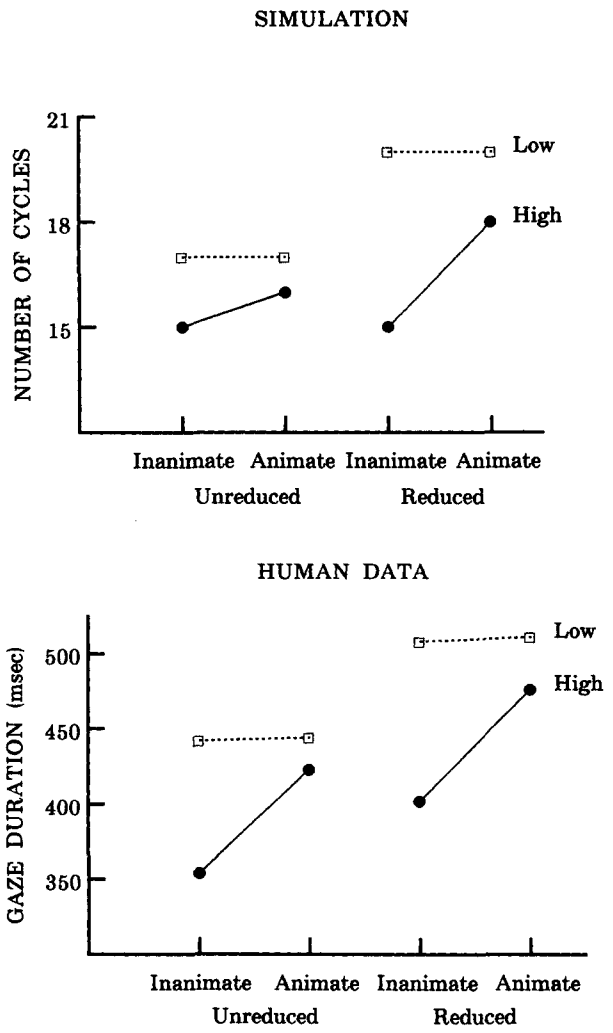

Figure 10. The number of cycles expended on the *by* phrase for sentences containing inanimate and animate grammatical subjects when the simulation is operating with more or less working memory capacity. (The bottom graph presents the human data for comparison with the simulation.)

Explorations of the Model

In this section, we will describe two aspects of the simulation model in more detail: (a) the total amount of activation being consumed after the processing of each successive word of a sentence, and (b) the consequences of alternative allocation schemes when the demands exceed the supply of activation.

Activation consumption over time. In addition to the number of cycles spent on each word of a sentence, CC READER offers another index of resource consumption—the total amount of activation being consumed at the completion of each word of a sentence. In the simulation runs we have done, each sentence was treated as text initial, as though working memory were previously unused. Thus, the processing began with some total amount of available activation, a quota that differed for the simulations of low, medium, and high capacity subjects. Then, as each successive word was read, some of the total activation was consumed by the partial and final products that were generated in processing that word. These products are part of the orthographic, lexical, semantic, and syntactic representation of the word and sentence. A more complete model would also generate a referential level of representation.

In general, the total amount of activation that is consumed increases as the successive words of the sentence are processed, up to the point at which the maximum activation is reached, as shown in Table 3. The table shows the activation consumption of a simulation of a low capacity subject (using an activation maximum of 29 units) and a high capacity subject (53 units).

Table 3
Total Consumption of Activation After Successive Words of a Sentence

Reading capacity	Units of activation consumed								
	The	reporter	that	the	senator	attacked	admitted	the	error.
Low	9.2	16.2	16.3	29.0	29.0	29.0	29.0	29.0	29.0
High	9.2	16.2	16.3	31.0	38.1	53.0	53.0	53.0	53.0

Once the consumption reaches the maximum level (the quota), the consumption generally remains at the maximum throughout the rest of the sentence. The higher the maximum capacity, the later in the sentence the maximum consumption is reached.

The time course of the consumption has a number of potentially interesting implications. First, any sentence that is long enough or complex enough can bring a reader to his or her maximum consumption. Second, the partial products from a completed sentence must be purged if the processing of an ensuing sentence is to start out at a consumption level far below the maximum. Presumably, any such purging would spare the highest levels of representation.

One counterintuitive prediction of the model is that under some circumstances a preceding context could slow the processing of an ensuing sentence, particularly in the case of a low capacity subject. These circumstances would arise if the costs (in terms of consumption of activation) of storing some of the representation of the preceding context outweighed the benefits. As we pointed out in the introduction, the costs may be minimized by mechanisms that select only the most central and most recent propositions for retention while deactivating less central propositions. The benefits are that the preceding context could preactivate relevant concepts that facilitate the integration of the new sentence. If the costs outweigh the benefits, then the processing of an ensuing sentence could be slower when it is preceded by a supporting context sentence.

Alternative allocation schemes. The simulation results described above were obtained with a budget allocation scheme that was evenhanded when the activation demands exceeded the supply. That is, both processing and storage demands were scaled back by the same proportion to stay within the maximum activation limit, imposing an across-the-board budget cut. We will now briefly describe two other allocation schemes that were explored. One scheme favors processing, charging less of the projected activation deficit against the processing demands and charging more against the storage demands. Thus, if the activation quota is about to be exceeded, the amount of activation that is propagated by a production is only slightly less than it would be otherwise, but the activation levels of old elements are decremented more severely. The second scheme favors storage. If the activation quota is about to be exceeded, the amount of activation that is propagated by a production is substantially less than it would be otherwise, but the activation levels of old elements are decremented only slightly. These two schemes are implemented by changing the value of a bias parameter in the model. These two budget allocation schemes, as well as the evenhanded one used in the main simulation, come into play only when the quota is reached, so they do not make any difference early in a text-initial sentence.

The allocation scheme that favors processing makes the processing faster after the quota is reached; that is, fewer cycles are spent per word relative to the evenhanded allocation scheme. This is because there is more activation available to propagate on each cycle, requiring fewer cycles to increment an element's activation to a given level. An occasional negative consequence of this scheme is that an important intermediate product can be forgotten, especially if the maximum quota is a low one to begin with. If this occurs, the parser essentially fails to complete any processing that is dependent on the forgotten intermediate product.

The allocation scheme that favors storage makes the processing slower after the quota is reached; that is, more cycles are spent per word. The favoring of storage means that, relative to an evenhanded scheme, more elements can be maintained in an activated state, or that a similar number of elements can be maintained in a more activated state. If this scheme were applied to the processing of the syntactically ambiguous sentences, then the representations of both interpretations would be maintained for a longer time than with an evenhanded allocation.

These different schemes, in combination with a quota that is intermediate between the high and low, suggest a mechanism for a trading relation between speed and accuracy of processing, in which accuracy refers to maintaining and later using partial products appropriately. If the activation quota is intermediate, then an evenhanded allocation scheme produces performance that is intermediate between the performance with the high and low quotas. The biased allocations schemes applied to the intermediate quota can produce less predictable results. Favoring processing can make the processing as fast as high-quota, evenhanded allocation, at the expense of an occasional comprehension error resulting from the forgetting of an important partial product. Favoring storage can make the processing as slow as in the low-quota, evenhanded allocation situation, but the probability of errors due to forgetting is lower. Thus, the allocation bias parameter of the model provides a possible mechanism for a trading relation between speed and accuracy.

In summary, the simulations capture the critical empirical phenomena in several experiments related to individual differences in working memory. Although the fit of the simulations is imperfect, it nevertheless demonstrates how the same basic comprehension strategy can produce different kinds of performance when different amounts of resources are available.

General Discussion

A capacity theory shifts the scientific schema within which cognition is studied to focus on the intensity and dynamics of thought in addition to its structural aspects. This focus provides the basis of our account of individual differences in comprehension, and it also provides a new perspective on a number of issues. In this section, we will discuss three of them: (a) the implications of the capacity theory for cognitive theories in areas other than language; (b) the implications of capacity theory for resource allocation policy and processing efficiency

differences; and (c) a comparison of the present analysis of individual differences to other approaches.

Dynamics of Cognition

A capacity theory deals centrally with the resources underlying thought. Like a structural theory, it assumes an underlying architecture, which in this case consists of a working memory, procedural knowledge (in the form of productions), and declarative knowledge (stored in a declarative knowledge base and in productions). The capacity theory further encompasses the dynamic aspects of processing and storage in this architecture, reflecting the moment-to-moment modulation in the resource demands. To account for the performance differences among individuals, the new theory proposes a dynamic allocation of a constrained capacity. Individual differences in the amount of capacity to be allocated or resultant differences in allocation policy can account for systematic differences in performance without postulating differences in the underlying architecture.

The capacity theory turned out to have a surprising implication for functional system architecture: The degree of interaction between subsystems (modules) may be dependent on capacity. Interaction between modules has previously been viewed as being either architecturally permitted or not permitted. Capacity theory makes the useful point that architectural permission may be a necessary condition for interaction between subsystems, but it is not sufficient. Like every other aspect of cognitive processing, interaction requires resources, and in the absence of the resource, the interaction cannot occur, even if it is architecturally permitted. Thus, the question of whether a module of a system is autonomous or interactive may depend on the capacity to sustain the interaction.

More generally, the term *architecture* calls forth an analogy to the design of a house, with an emphasis on the partitioning of a larger structure into functionally distinguishable modules, such as kitchens and sleeping areas, and traffic patterns among them. In contrast, a focus on resource-dependent activity calls forth an analogy to a dynamic system, such as a river. Even though a river can sometimes be partitioned, the partitions are not its main features. Rather, the main features are the flow and the hydraulic structures, such as a waterfall or a standing wave or a whirlpool, which are influenced by the variation in the river's volume. Although both waterfalls and standing waves have the appearance of permanent attributes, they differ in that waterfalls are fairly permanent consequences of basic structure, whereas standing waves are transient attributes that can come and go with changes in the flow. To fully characterize a river, it is necessary to provide some account of its dynamics, and it is not sufficient to specify only its structure, as an aerial photograph might. A capacity theory of language provides some of the beginnings of a theory of cognitive dynamics.

Other cognitive domains. The implications of capacity theory may be examined in cognitive domains other than language, such as problem solving, complex decision making, and higher visual information processing. These domains seem amenable to analysis within a capacity theory because, like language, they involve sequential symbol manipulation.

One implication of capacity theory is that some of the performance differences among individuals within a task domain will be explained in large part in terms of working memory capacity. When the task demands are high enough to strain capacity, individuals with a smaller working memory capacity should be less able to perform computations quickly or store intermediate products. Some of these implications have recently been confirmed in the area of complex problem solving, specifically in solving the problems in the Raven Progressive Matrices Test. Subjects with lower scores in the Raven test were less able to store intermediate goals, as indicated in an independent task, and as indicated by their disadvantage on Raven test items that required the storage of a large number of subgoals and partial solutions (Carpenter, Just, & Shell, 1990).

A related implication is that within any task domain large performance differences among individuals will emerge, primarily when the task demands consume sufficient capacity to exhaust some subjects' resources. In the domain of language comprehension, capacity limitations are more evident when the linguistic construction is more complex or when there is an extrinsic load, as we have described. The current research extends this approach beyond finding larger individual differences in harder tasks, to finding larger individual differences in the harder parts of a single task. For example, within a given reading task, the reading times vary with the transient demand as the reader progresses through a sentence. Similarly, capacity use in other domains should vary with the ongoing computational and storage demands. Researchers in the domains of perception, motor control, and problem solving have shown that more capacity is required for more difficult tasks (e.g., Hirst, Spelke, Reaves, Caharack, & Neisser, 1980; Norman & Bobrow, 1975; Schneider & Shiffrin, 1977; Shallice, 1982). The research reported in this article suggests that individual differences could be used as an avenue for examining the effects of capacity constraints within tasks in these domains as well.

A third implication of capacity theory is that there is an intensity dimension of thought, in addition to the correctness and speed dimensions. Moreover, the intensity of thought varies in magnitude throughout the performance of a given task. Phenomenologically, one can feel oneself being more or less engaged in one's thought processes, or feel more or less concentration of thought. More objectively, pupillometry studies (Ahern & Beatty, 1979; Beatty, 1982) and physiological studies of glucose metabolism (Haier et al., 1988) have confirmed that these measures of intensity are sensibly related to the subject and task characteristics. For example, the pupillometry results indicate that the effort expended is greater if the subject is less skilled (as indicated by psychometric scores) or the task is more difficult (e.g., varying in arithmetic difficulty). Similarly, the glucose metabolism studies indicate that less skilled subjects expend greater effort at solving difficult problems from the Raven test. These findings are consistent with the hypothesis that individuals differ systematically in the effort that they have to expend to perform a task and that different tasks consume different amounts of resources in several domains besides language comprehension.

A fourth implication of the theory is specific to the domain of attention, the birthplace of capacity theory. In particular, Kahneman's (1973) capacity theory of attention laid a foundation for most of its successors, including the comprehension theory described in this article. Capacity theories of attention

account for performance decrements that occur when the resource demands of the task exceed the available supply (Navon & Gopher, 1979; Wickens, 1984). Because detailed capacity theories of attention already exist, they are able to immediately benefit from the specific proposal made in the comprehension theory, that capacity be defined in terms of the activation available for information maintenance and computation. Attention theories often refer to some underlying commodity that enables performance, usually labeling it *capacity* or *resources* but failing to specify its nature (Navon, 1984). The activation definition provided by the current theory sharpens the concept of resources and their allocation, and may ultimately provide better accounts of capacity constraints in the types of situations addressed by theories of attention.

Finally, the theory implies that in areas besides language processing, interactions between component processes in a complex task will occur only if they are architecturally permissible and if there are adequate resources to support the interaction. Admittedly, language is the domain in which modules of processing are most frequently postulated. Nevertheless, it is often possible to decompose the performance of a complex task into subprocesses whose interaction can be usefully examined from this new perspective.

One capacity or many capacities? In this analysis of working memory capacity for language, we have assumed that there is a single capacity that encompasses various facets of language comprehension processing, including lexical access, syntactic analysis, and referential processing. This assumption is supported by several results. Perhaps the most important is the finding that comprehension deteriorated similarly when the demand on capacity was increased by a diversity of factors, including a syntactic embedding, a syntactic ambiguity, or the presence of additional sentences between a noun and a subsequent pronominal reference to it. In addition, the presence of an extrinsic load degraded comprehension similarly to a syntactic complexity (a center-embedded clause), suggesting that both effects occurred because of processes that drew on shared resources. The results of the animacy study, in which pragmatic information influenced the syntactic analysis for high span subjects, suggest that syntactic processing draws on the same capacity that supports the maintenance and use of pragmatic information. Finally, the Reading Span task measure correlated with comprehension performance across all of the various studies, implying that a common capacity (the one measured in the Reading Span test) mediated performance in all of the studies, regardless of whether a syntactic or a pragmatic factor was manipulated to increase the processing demand. The total set of results is most easily explained by a common capacity that underlies language comprehension.

Although the results from comprehension tasks indicate a common working memory capacity for language comprehension, it would be incorrect to assume that all language processes draw on a single capacity. In particular, there is evidence that language production may draw on somewhat different resources. Individual differences in a word-generation task (generating an appropriate completion for an incomplete sentence) do not correlate with individual differences in the Reading Span task after motor production speed has been partialled out (Daneman & Green, 1986). Another locale in which comprehension and production are distinguished is in language development. In a longitudinal study of language development in children, separate clusters of production and comprehension skills developed with different time courses, again supporting the suggestion that these are different types of skills (Bates, Bretherton, & Snyder, 1988).

In tasks that do not involve any overt language use at all (such as arithmetic or spatial tasks), the required resources overlap only partially with those used in language comprehension. Performance in the reading span task is sometimes positively correlated with nonlanguage tasks, but the correlations are generally much lower than with language comprehension tasks (Baddeley et al., 1985; Daneman & Tardif, 1987), indicating that only a subset of the resources is shared across diverse tasks (but see Turner & Engle, 1989, for an alternative view).

In sum, we cannot conclude that the working memory capacity used for language comprehension is the single cognitive capacity. Rather, it is likely that there is a large set of processing resources, only a subset of which is used for a given task domain. It remains to be seen whether a capacity theory within a domain other than language will be equally effective.

Resource Allocation and Processing Efficiency

The capacity theory also helps to crystallize a number of questions for which we cannot offer definitive answers, but which nevertheless benefit by being considered within this theoretical framework. In this section, we will describe two such issues: resource allocation and the role of processing efficiency.

Resource allocation schemes. The different performance characteristics of individuals as a function of their working memory capacity indicate the existence of an implicit allocation policy when demands exceed capacity. If the demand for resources exceeds supply, what factors influence the allocation of activation? We have already briefly explored the consequences of allocation schemes that favor storage or favor processing, but we have not discussed the question of whether some processes might be favored over others. For example, those processes that are less demanding of resources might be favored at time of inadequate resource supply. The less demanding processes include lower level processes (e.g., perceptual recognition) and automatic processes. In contrast, the higher level processes in comprehension, such as those that construct the referential level of representation in the case of a syntactic ambiguity, may be not executed or not executed fully in times of inadequate resource supply.

Processes that require a greater variety of inputs also may be less likely to be executed than those that require only one type of input. For example, if one of the inputs to the process is not yet computed because it depends on other intermediate products, then this process will be less likely to be executed to completion before a deadline is reached. Such a mechanism may account for the lack of effect of pragmatic information on first-pass syntactic processing in low span subjects. Thus, even in a parallel system, interactive processes might be subject to capacity constraints more than noninteractive processes. These general allocation heuristics account for the present studies and suggest that language comprehension may be a useful domain

in which to study the issue that arises in other domains when the person's resources cannot meet the task's demands.

Total capacity and efficiency. As we discussed earlier, the individual differences reported here may reflect differences in total capacity, differences in processing efficiency, or both. Choosing between these two explanations of individual differences is not necessarily a matter of deciding which one is correct, but rather of deciding which one best accounts for a phenomenon, given some assumptions about total capacity and processing efficiency. One assumption is that capacity limitations affect performance only if the resource demands of the task exceed the available supply. One consequence of the assumption is that individual differences should be most evident during periods of high demand. If individuals differ in the amount of available activation, then performance differences should be more apparent if the task's demands exceed the available resources of lower ability subjects. Consistent with this assumption, our studies documented that performance differences among college student readers of different working memory capacities are smaller when the comprehension task is easy, and larger when it is demanding.

A second assumption is that changes in total capacity affect the execution of a wide range of processes in a wide range of tasks. Consequently, the types of generalized performance changes induced by fatigue, extreme age, or concentration can usually be interpreted as consequences of changes in total capacity. In contrast, a change in processing efficiency is assumed to be more specific to a particular process. Thus, changes in the efficiency of a process are often assumed to result from practice or some instructional intervention. Indeed, intensive practice in several simple tasks, such as Stroop-type tasks, induces large changes in the speed of responding that are typically interpreted in terms of changes in efficiency of underlying processes (Cohen, Dunbar, & McClelland, 1990; Schneider & Shiffrin, 1977). Intensive practice in reading might similarly induce greater efficiency in some component processes of comprehension; the time spent in out-of-school reading is correlated with reading skill in fifth-grade students, accounting for approximately 9% of the variance in one study (Anderson, Wilson, & Fielding, 1988). Total capacity is assumed to be less susceptible to such intervention. Although it is possible to explain practice effects in terms of a change in total capacity, the explanation would have to be more complex. In order to increase total capacity, the practice would have to recruit more activation or additional processes and structures, quite apart from producing any efficiency gain in the originally targeted processes.

The total capacity and processing efficiency accounts are not mutually exclusive. Like the particle–wave duality that is used to explain the nature of light, total capacity and process efficiency may conjointly explain differences in effective working memory among individuals. However, we believe that the new phenomena reported here are better explained in terms of total capacity than process efficiency.

Alternative Approaches to Individual Differences

The capacity explanation of individual differences can be related to several other explanations of the nature of individual differences in language processing. In this section, we will examine three such explanations.

Identifying a particular process as a source of differences. A number of earlier studies have started with the widely held assumption that language comprehension consists of a set of component processes, and that it should be possible to identify a particular process or set of processes as a source of individual differences in reading comprehension.

Word perception processes, including word decoding and lexical access, are frequently implicated as a source of individual differences (Frederiksen, 1981; Jackson & McClelland, 1979; Perfetti, 1985; Perfetti & Lesgold, 1977). One of the originators of this approach, Hunt (1978; Hunt, Lunneborg, & Lewis, 1975), found a highly replicable correlation (about .3) between the speed of retrieving complex codes and a global measure of verbal ability (such as SAT verbal tests). This finding suggested that retrieval-intensive operations, such as lexical access, might be a source of differences. Other researchers have found that the speed of word decoding, even among college students, is a source of individual differences (Cunningham, Stanovich, & Wilson, 1990; Frederiksen, 1981). Another major source of individual differences is *higher level comprehension processes,* such as syntactic, semantic, and referential-level processes. These higher level processes may be the source of the high correlation between reading and listening comprehension performance (e.g., Curtis, 1980; Sticht, 1977). For example, Jackson and McClelland (1979) found that the best predictor of good reading (operationalized as a combination of speed and accuracy) was good listening comprehension, with speed of access to letter names accounting for an additional but small proportion of the individual differences (see also Palmer, MacLeod, Hunt, & Davidson, 1985). The implicit assumption underlying this approach is that a few component processes in language comprehension are responsible for individual differences; by virtue of being particularly slow or errorful, these processes supply degraded information to other comprehension processes, thereby degrading the entire performance.

A somewhat different assumption, more in keeping with the capacity theory we have proposed, is that a slow or errorful component process robs other processes not only of good data, but also of resources (Frederiksen, 1981). For example, if someone were particularly slow at lexical access, that process might consume so much time that an insufficient amount of time was left for other processes, such as syntactic analysis, to execute properly (as suggested by Perfetti & Lesgold, 1977). Capacity theory predicts that individual differences in a single component process could generate differences in total capacity, and thus can account for the results. The capacity account is further supported by the common finding that individual differences in overall reading ability are not uniquely associated with a single component process, such as lexical access, but with a variety of different component processes of comprehension (Frederiksen, 1981), all of which may be fast or slow because of an overall capacity difference.

Vocabulary size. Another componential approach to explaining individual differences in comprehension focuses on vocabulary size. There is a strong correlation between vocabulary size and reading comprehension (Jensen, 1980; Marshalek, 1981; Sternberg, 1985; Sternberg & Powell, 1983). The conven-

tional explanation of this correlation is that the processes that make comprehension more efficient in some people than in others are approximately the same processes that lead to vocabulary acquisition. Supporting this explanation, several studies have shown that people who are high in comprehension ability are better able to induce the meaning of a word from context, a process thought to lie at the heart of vocabulary acquisition (Sternberg & Powell, 1983; van Daalen-Kapteijns & Elshout-Mohr, 1981; Werner & Kaplan, 1952). According to capacity theory, people with a larger working memory for language would not only have an advantage in normal comprehension, but their extra capacity could also provide the resources to permit better induction of word meanings and hence better vocabulary acquisition. Consistent with this hypothesis, Daneman and Green (1986) found a high correlation (.69) between reading span and the ability to provide a definition of a novel word in a passage. Moreover, the correlation was high (.53) even if general vocabulary knowledge was partialled out. In contrast, vocabulary knowledge itself did not significantly correlate with the ability to induce word meanings when reading span was partialled out. In sum, working memory capacity plays a role in vocabulary acquisition, and therefore in accounting for differences in vocabulary knowledge, by virtue of its role in comprehension.

Motivational differences. Performance differences are sometimes ascribed to motivational rather than cognitive factors. Such a hypothesis might postulate that lower span individuals simply do not try as hard as higher span individuals either in the Reading Span task or in other comprehension tasks, and that consequently they recall fewer words and correctly answer fewer comprehension questions. Although a motivational explanation could account for such general differences, there are two features of the overall results that strongly favor the capacity theory explanation over a motivational explanation. First, the motivational explanation has difficulty accounting for those cases in which lower span readers expend more effort than higher span readers, if effort is assessed by the time spent trying to comprehend a sentence. In particular, low span readers spent more time than high span readers on the most complex parts of the sentences containing a center-embedded clause. A more subtle reason for rejecting the motivational account is that such an explanation is at a molar level; it might explain group differences, but it would give little insight into the precise form of performance degradation for the more fine-grained manipulations of several of the experiments. For example, in the experiment that varied the distance between a pronoun and its prior referent, low span readers and medium span readers showed systematic degradation in performance as a function of the distance, not just poorer performance than high span readers. Similarly, the effects of span on reaction time for object-relative sentences occurred on those portions of the sentence that are most capacity demanding, according to independent linguistic criteria; the effects are not generalized reading differences. The specificity and orderliness of such effects are well accounted for by the capacity theory but would not be illuminated by a general motivational factor.

The theory of capacity constrained comprehension not only provides an account of individual differences, but also suggests an important new perspective to complement the structural analysis that has dominated much of recent analyses of cognitive architecture. It also serves to bring language comprehension closer to the analysis of other types of cognition.

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with Scholastic Aptitude Test scores. *Science, 205,* 1289–1292.

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

Anderson, R. C., Wilson, P. T., & Fielding, L. G. (1988). Growth in reading and how children spend their time outside of school. *Reading Research Quarterly, 23,* 285–303.

Baddeley, A. D. (1986). *Working memory.* New York: Oxford University Press.

Baddeley, A. D., Eldridge, M., Lewis, V., & Thomson, N. (1984). Attention and retrieval from long-term memory. *Journal of Experimental Psychology: General, 113,* 518–540.

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). San Diego, CA: Academic Press.

Baddeley, A. D., Logie, R., Nimmo-Smith, I., & Brereton, N. (1985). Components of fluent reading. *Journal of Memory and Language, 24,* 119–131.

Bates, E., Bretherton, I., & Snyder, L. (1988). *From first words to grammar: Individual differences and dissociable mechanisms.* Cambridge, England: Cambridge University Press.

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin, 91,* 276–292.

Bever, T. G. (1970). The cognitive basis for linguistic structures. In J. R. Hayes (Ed.), *Cognition and the development of language* (pp. 279–362). New York: Wiley.

Carpenter, P. A., & Just, M. A. (1977). Integrative processes in comprehension. In D. LaBerge & J. Samuels (Eds.), *Basic processes in reading: Perception and comprehension* (pp. 217–241). Hillsdale, NJ: Erlbaum.

Carpenter, P. A., & Just, M. A. (1983). What your eyes do while your mind is reading. In K. Rayner (Ed.), *Eye movements in reading: Perceptual and language processes* (pp. 275–307). San Diego, CA: Academic Press.

Carpenter, P. A., & Just, M. A. (1989). The role of working memory in language comprehension. In D. Klahr & K. Kotovsky (Eds.), *Complex information processing: The impact of Herbert A. Simon* (pp. 31–68). Hillsdale, NJ: Erlbaum.

Carpenter, P. A., Just, M. A., & Shell, P. (1990). What one intelligence test measures: A theoretical account of the processing in the Raven Progressive Matrices Test. *Psychological Review, 97,* 404–431.

Chang, F. R. (1980). Active memory processes in visual sentence comprehension: Clause effects and pronominal reference. *Memory & Cognition, 8,* 58–64.

Clark, H. H., & Sengul, C. J. (1979). In search of referents for nouns and pronouns. *Memory & Cognition, 7,* 35–41.

Cohen, G. (1979). Language comprehension in old age. *Cognitive Psychology, 11,* 412–429.

Cohen, J. D., Dunbar, K., & McClelland, J. L. (1990). On the control of automatic processes: A parallel distributed processing account of the Stroop effect. *Psychological Review, 97,* 332–361.

Cottrell, G. W. (1989). *A connectionist approach to word sense disambiguation.* London: Pitman.

Cunningham, A. E., Stanovich, K. E., & Wilson, M. R. (1990). Cognitive variation in adult college students differing in reading ability. In

T. H. Carr & B. A. Levy (Eds.), *Reading and its development* (pp. 129–159). San Diego, CA: Academic Press.

Curtis, M. E. (1980). Development of components of reading skill. *Journal of Educational Psychology, 72,* 656–669.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior, 19,* 450–466.

Daneman, M., & Green, I. (1986). Individual differences in comprehending and producing words in context. *Journal of Memory and Language, 25,* 1–18.

Daneman, M., & Tardif, T. (1987). Working memory and reading skill re-examined. In M. Coltheart (Ed.), *Attention and performance XII* (pp. 491–508). London: Erlbaum.

Ellis, N. C., & Hennelley, R. A. (1980). A bilingual word-length effect: Implications for intelligence testing and the relative ease of mental calculation in Welsh and English. *British Journal of Psychology, 71,* 43–52.

Ferreira, F., & Clifton, C. (1986). The independence of syntactic processing. *Journal of Memory and Language, 25,* 348–368.

Fischer, B., & Glanzer, M. (1986). Short-term storage and the processing of cohesion during reading. *The Quarterly Journal of Experimental Psychology, 38A,* 431–460.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: Bradford.

Frazier, L. (1978). *On comprehending sentences: Syntactic parsing strategies.* Bloomington: Indiana University Linguistics Club.

Frederiksen, J. R. (1981). *A componential theory of reading skills and their interactions* (ONR Final Report No. 4649). Cambridge, MA: Bolt Beranek & Newman.

Garfield, J. (Ed.). (1989). *Modularity in knowledge representation and natural-language understanding.* Cambridge, MA: MIT Press.

Gernsbacher, M. A., Varner, K. R., & Faust, M. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16,* 430–445.

Glanzer, M., Dorfman, D., & Kaplan, B. (1981). Short-term storage in the processing of text. *Journal of Verbal Learning and Verbal Behavior, 20,* 656–670.

Glanzer, M., Fischer, B., & Dorfman, D. (1984). Short-term storage in reading. *Journal of Verbal Learning and Verbal Behavior, 23,* 467–486.

Gorrell, P. G. (1987). *Studies of human syntactic processing: Ranked-parallel versus serial models.* Unpublished doctoral dissertation, University of Connecticut, Storrs.

Haier, R. J., Siegel, B. V., Nuechterlein, K. H., Hazlett, E., Wu, J. C., Paek, J., Browning, H. L., & Buchsbaum, M. S. (1988). Cortical glucose metabolic rate correlates of abstract reasoning and attention studied with positron emission tomography. *Intelligence, 12,* 199–217.

Hirst, W., Spelke, E. S., Reaves, C. C., Caharack, G., & Neisser, R. (1980). Dividing attention without alternation or automaticity. *Journal of Experimental Psychology: General, 109,* 98–117.

Hitch, G. J., & Baddeley, A. D. (1976). Verbal reasoning and working memory. *Quarterly Journal of Experimental Psychology, 28,* 603–621.

Holmes, V. M., & O'Regan, J. K. (1981). Eye fixation patterns during the reading of relative clause sentences. *Journal of Verbal Learning and Verbal Behavior, 20,* 417–430.

Huey, E. B. (1908). *The psychology and pedagogy of reading.* New York: Macmillan.

Hunt, E. B. (1978). Mechanics of verbal ability. *Psychological Review, 85,* 199–230.

Hunt, E. B., Lunneborg, C., & Lewis, J. (1975). What does it mean to be high verbal? *Cognitive Psychology, 2,* 194–227.

Jackson, M. D., & McClelland, J. L. (1979). Processing determinants of reading speed. *Journal of Experimental Psychology: General, 108,* 151–181.

Jarvella, R. J. (1971). Syntactic processing of connected speech. *Journal of Verbal Learning and Verbal Behavior, 10,* 409–416.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

Just, M. A., & Carpenter, P. A. (1980). A theory of reading: From eye fixations to comprehension. *Psychological Review, 87,* 329–354.

Just, M. A., & Carpenter, P. A. (1987). *The psychology of reading and language comprehension.* Newton, MA: Allyn & Bacon.

Just, M. A., Carpenter, P. A., & Woolley, J. D. (1982). Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General, 111,* 228–238.

Kahneman, D. (1973). *Attention and effort.* Englewood Cliffs, NJ: Prentice-Hall.

Keenan, J. M., Baillet, S. D., & Brown, P. (1984). The effects of causal cohesion on comprehension and memory. *Journal of Verbal Learning and Verbal Behavior, 23,* 115–126.

Kemper, S. (1986). Imitation of complex syntactic constructions by elderly adults. *Applied Psycholinguistics, 7,* 277–287.

Kemper, S. (1988). Geriatric psycholinguistics: Syntactic limitations of oral and written language. In L. Light & D. Burke (Eds.), *Language, memory, and aging* (pp. 58–76). New York: Cambridge University Press.

King, J., & Just, M. A. (1991). Individual differences in syntactic processing: The role of working memory. *Journal of Memory and Language, 30,* 580–602.

Kintsch, W. (1988). The use of knowledge in discourse processing: A construction-integration model. *Psychological Review, 95,* 163–182.

Kintsch, W., & vanDijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85,* 363–394.

Klapp, S. T., Marshburn, E. A., & Lester, P. T. (1983). Short-term memory does not involve the "working memory" of information processing: The demise of a common assumption. *Journal of Experimental Psychology: General, 112,* 240–264.

Kurtzman, H. (1985). *Studies in syntactic ambiguity resolution.* Doctoral dissertation, MIT, Cambridge, MA. Distributed by Indiana University Linguistics Club, Bloomington.

Kynette, D., & Kemper, S. (1986). Aging and the loss of grammatical forms: A cross-sectional study of language performance. *Language and Communication, 6,* 65–72.

Larkin, W., & Burns, D. (1977). Sentence comprehension and memory for embedded structure. *Memory & Cognition, 5,* 17–22.

Lesgold, A. M., Roth, S. F., & Curtis, M. E. (1979). Foregrounding effects in discourse comprehension. *Journal of Verbal Learning and Verbal Behavior, 18,* 291–308.

Lyon, D. (1977). Individual differences in immediate serial recall: A matter of mnemonics? *Cognitive Psychology, 9,* 403–411.

MacDonald, M. C., Just, M. A., & Carpenter, P. A. (in press). Working memory constraints on the processing of syntactic ambiguity. *Cognitive Psychology.*

Marcus, M. P. (1980). *A theory of syntactic recognition for natural language.* Cambridge, MA: MIT Press.

Marshalek, B. (1981). *Trait and process aspects of vocabulary knowledge and verbal ability* (NR154-376 ONR Technical Report No. 15). Stanford, CA: Stanford University, School of Education.

Masson, M. E. J., & Miller, J. A. (1983). Working memory and individual differences in comprehension and memory of text. *Journal of Educational Psychology, 75,* 314–318.

McClelland, J. L. (1979). On the time relations of mental processes: An examination of systems of processes in cascade. *Psychological Review, 86,* 287–330.

McClelland, J. L., & Rumelhart, D. E. (1988). *Explorations in parallel distributed processing: A handbook of models, programs, and exercises.* Cambridge, MA: MIT Press.

Navon, D. (1984). Resources—A theoretical soup stone? *Psychological Review, 91,* 216–234.

Navon, D., & Gopher, D. (1979). On the economy of the human processing system. *Psychological Review, 86,* 214–255.

Norman, D. A., & Bobrow, D. G. (1975). On data-limited and resource-limited processes. *Cognitive Psychology, 7,* 44–64.

Palmer, J., MacLeod, C. M., Hunt, E., & Davidson, J. E. (1985). Information processing correlates of reading. *Journal of Memory and Language, 24,* 59–88.

Perfetti, C. A. (1985). *Reading ability.* New York: Oxford University Press.

Perfetti, C. A., & Goldman, S. R. (1976). Discourse memory and reading comprehension skill. *Journal of Verbal Learning and Verbal Behavior, 14,* 33–42.

Perfetti, C. A., & Lesgold, A. M. (1977). Discourse comprehension and sources of individual differences. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 141–183). Hillsdale, NJ: Erlbaum.

Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception and Psychophysics, 2,* 437–442.

Sanford, A. J., & Garrod, S. C. (1981). *Understanding written language: Explorations in comprehension beyond the sentence.* New York: Wiley.

Schneider, W., & Shiffrin, R. M. (1977). Controlled and automatic human information processing: I. Detection, search, and attention. *Psychological Review, 84,* 1–66.

Shallice, T. (1982). Specific impairments of planning. *Philosophical Transactions of the Royal Society of London B, 298,* 199–209.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge, England: Cambridge University Press.

Sharkey, N. E., & Mitchell, D. C. (1985). Word recognition in a functional context: The use of scripts in reading. *Journal of Memory and Language, 24,* 253–270.

Sheldon, A. (1974). The role of parallel function in the acquisition of relative clauses in English. *Journal of Verbal Learning and Verbal Behavior, 13,* 272–281.

St. John, M. F., & McClelland, J. L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence, 46,* 217–257.

Sternberg, R. J. (1985). *Beyond IQ: A triarchic theory of human intelligence.* New York: Cambridge University Press.

Sternberg, R. J., & Powell, J. S. (1983). Comprehending verbal comprehension. *American Psychologist, 38,* 878–893.

Sticht, T. G. (1977). Comprehending reading at work. In M. A. Just & P. A. Carpenter (Eds.), *Cognitive processes in comprehension* (pp. 221–246). Hillsdale, NJ: Erlbaum.

Thibadeau, R., Just, M. A., & Carpenter, P. A. (1982). A model of the time course and content of reading. *Cognitive Science, 6,* 157–203.

Tipper, S. P. (1985). The negative priming effect: Inhibitory priming by ignored objects. *Quarterly Journal of Experimental Psychology, 37A,* 571–590.

Tipper, D. P., & Cranston, M. (1985). Selective attention and priming: Inhibitory and facilitatory effects of ignored primes. *Quarterly Journal of Experimental Psychology, 37A,* 591–611.

Turner, M. L., & Engle, R. W. (1989). Is working memory capacity task dependent? *Journal of Memory and Language, 28,* 127–154.

van Daalen-Kapteijns, M. M., & Elshout-Mohr, M. (1981). The acquisition of word meanings as a cognitive learning process. *Journal of Verbal Learning and Verbal Behavior, 20,* 386–399.

vanDijk, T. A., & Kintsch, W. (1983). *Strategies of discourse comprehension.* San Diego, CA: Academic Press.

Waltz, D. L., & Pollack, J. B. (1985). Massively parallel parsing: A strongly interactive model of natural language interpretation. *Cognitive Science, 9,* 51–74.

Werner, H., & Kaplan, E. (1952). The acquisition of word meanings: A developmental study. *Monographs of the Society for Research in Child Development, 15,* 190–200.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention* (pp. 63–102). San Diego, CA: Academic Press.

Winograd, T. (1983). *Language as a cognitive process.* Reading, MA: Addison-Wesley.

Yuill, N., Oakhill, J., & Parkin, A. (1990). *Working memory, comprehension ability and the resolution of text anomaly.* Unpublished manuscript, University of Sussex, Brighton, East Sussex, England.

Appendix

Examples of Sentences That the Model Can Parse

Unambiguous simple sentences
- The senator spoke the truth.
- The senator was attacked by the reporter.

Unambiguous embedded sentences
- Subject relatives
  - The senator who attacked the reporter admitted the error.
  - The senator who attacked the reporter was warned by the policeman.
  - The senator who was attacked by the reporter was warned by the policeman.
  - The senator who was attacked by the reporter was from Illinois.
- Object relatives
  - The senator who the reporter attacked told the policeman.
  - The senator who the reporter attacked was warned by the policeman.
  - The senator the reporter attacked told the policeman.
  - The senator the reporter attacked was from Illinois.
  - The senator the reporter attacked cried.

Ambiguous sentences
- Main verb interpretation
  - The senator attacked with the bat.
  - The senator attacked the reporter.
- Relative clause interpretation
  - The senator attacked with the bat admitted the error.
  - The senator attacked by the reporter admitted the error.

Received September 14, 1990
Revision received January 28, 1991
Accepted March 18, 1991 ■