Research News

The Workings of Working Memory

The central thesis of cognitive science is that the mind is an information processor; the study of reading gives a unique insight into how that processor works

This is the first in an occasional series of articles on cognitive science and artificial intelligence: the study of the mind as an information processor.

LANGUAGE is such an everyday kind of activity that we tend to take it for granted. After all, no one has to be a genius to read a novel, or to carry on a conversation. But consider what is involved from an information-processing standpoint. For the communicator, language is a process of encoding—taking a complex and highly nonlinear web of ideas, feelings, and associations, and then somehow reducing that web to a linear string of words. For the reader or listener it is just the inverse—taking that linear string of words, decoding it, and then somehow reconstructing a semblance of the original ideas, feelings, and associations in his or her own head.

The more detail one considers, in fact, the more daunting language becomes. Cognitive psychologists have generally concentrated on the comprehension half of language—that is, the decoding process—where they can at least give their experimental subjects a known text as input. But even in that restricted domain, our mental language processors still have to recognize individual words, classify them according to parts of speech, figure out the grammatical role each word is playing, link the words into phrases, clauses, and sentences, keep track of pronouns and what they refer to, and follow chains of inference—all simultaneously.

Indeed, as Carnegie-Mellon University psychologist Marcel A. Just explained at this year's Carnegie-Mellon Symposium on Cognition,* the complexity of language has forced cognitive scientists to confront a major paradox: every time we read a book or carry on a conversation, we do a staggering amount of information processing—and yet we do it with mental information processors that are sharply limited in capacity. The question is how?

To understand more clearly what those limits are, it is important to make a distinction between two kinds of human memory. "Long-term" memory, which is presumably where we store knowledge about grammar and word meanings, along with everything else we know, is essentially infinite in capacity. No one has ever had a head so full of information that he or she could not learn something new. (We leave aside the question of whether he or she *wants* to learn anything new.) Furthermore, long-term memory is remarkably long-lasting, probably because it is based on molecular changes at the neural level. Thus, octogenarians can remember their childhoods vividly.

However, before information can get into long-term memory, it first has to pass though a kind of gateway known as "short-term" memory, which essentially consists of the set of things we are paying attention to at any given time. This is presumably where we store the words we read or hear while we are trying to decode them. The problem is that short-term memory is a very narrow channel. As its name suggests, for example, its contents are highly ephemeral, fading almost as soon as our attention turns elsewhere. Imagine looking up a number in the telephone directory: it barely survives until you dial it.

Worse, short-term memory can hold only a handful of items at any one time, for reasons that no one really understands. True, the definition of "item" is quite elastic; a random string of digits such as 279189471641 would qualify as 12 items, whereas the same digits grouped into more meaningful units—1776, 1492, 1984—would count as only three. Nonetheless, so long as one is counting only meaningful pieces of data, otherwise known as "chunks," the capacity of short-term memo-

AP/Wide World Photo

In the reading laboratory *Carnegie-Mellon University psychologists Patricia Carpenter (left) and Marcel Just have made extensive use of the computerized eye-tracker in the background to analyze the cognitive processes involved in reading.*

*The 21st Carnegie-Mellon Symposium on Cognition, 28 to 30 May 1987, Carnegie-Mellon University, Pittsburgh.

ry is no greater than about seven. Some researchers claim it is closer to four.

So how can language comprehension proceed in such a constrained environment?

A big part of the answer seems to lie in the nature of short-term memory itself, said Just. Whatever else it may be, short-term memory is not just a passive storehouse for data, as envisioned in classical models of memory. During the past decade or so, through a wide variety of psychological experiments and computer models of cognition, researchers have come to see short-term memory as a much more dynamic arena—a kind of blackboard where the mind performs its computations, and where it posts its partial results for later use. Indeed, in a real sense short-term memory *is* the mind's information processor. And for that very reason, said Just, many cognitive scientists now prefer to drop the traditional name in favor of a more descriptive term: "working memory."

In the case of language the dynamism of working memory shows up as what Just calls "immediacy of interpretation." That is, we take the words as they come. We process them on the fly. We make the best interpretation we can as quickly as we can. And then we almost instantly let the unneeded details fall away to make room for whatever comes next.

After more than a decade of research into the cognitive aspects of reading, said Just, he, his colleague Patricia A. Carpenter, and their associates consider this on-the-fly processing strategy to be one of their central discoveries. It is by no means obvious. For example, many researchers in the past have tacitly assumed that we follow a much safer and more rational wait-and-see strategy, postponing the interpretation of a phrase or sentence until we can put each word into context with the words around it. And yet, said Just, the empirical evidence for immediacy is now quite substantial. For example:

■ Once we have finished reading, say, a detective novel, we can generally recall only the gist of it—a broad outline of the plot, perhaps, the names of some of the characters, and who did the deed. We almost never remember such details as the exact wording of the sentences; this is relatively low-level information that we use for processing the text as we go along, and then throw away.

Generally speaking, said Just, the lower the level of information, the more quickly it seems to fade from working memory. At the very lowest level, for example, is information about the typeface of the individual letters, and whether they are upper or lower case. To demonstrate precisely how fast this information decays, researchers from the University of Illinois recently presented subjects with a computer display in which the cases alternated from letter to letter: *In ThE eStUaRiEs Of ThE fLoRiDa EvErGlAdEs ThE rEd MaNgRoVe.* . . . Meanwhile, they monitored exactly where the subjects were focusing at each instant using a standard eye fixation apparatus, which reflects a beam of infrared light from the cornea.

We do a staggering amount of information processing—with mental information processors that are sharply limited in capacity. The question is how?

Now, in previous work with this kind of apparatus, said Just, he and Carpenter had already shown that people tend to scan written text in a rather jerky fashion. Readers will fixate on one point in the sentence for a few tenths of a second, and then move their eyes very rapidly to the next fixation point one or two words away. Since these fixations tend to occur at the most important words of the text—*estuaries*, *Florida*, *Everglades*, and so on in the above example—the presumption is that they correlate with the reader's mental information processing; the longer the fixation on a given word, the heavier the computational load.

What the Illinois experimenters did was to use this fact to play a trick, said Just. As expected, the subjects' fixate-and-move pattern was unaffected by the alternating type cases. But sometimes, when a subject was in the middle of a eye movement and presumably not encoding anything, the computer would suddenly change all the upper case letters to lower case, and vice versa: *iN tHe EsTuArIeS oF tHe FlOrIdA eVeRgLaDeS tHe ReD mAnGrOvE.* . . . The surprise was that the subjects never even noticed. Moreover, their eye fixations gave no evidence of extra computation. The low-level information about type case had essentially been lost from their working memories within about 15 milliseconds, the average time it takes to flick the eye from one fixation to the next. (To a bystander, by the way, the changes were very visible indeed; anyone not synchronized with the display would usually be caught in the middle of his or her own eye fixation and would thus be able to see the transformation clearly.)

At a somewhat higher level, said Just, word order seems to stay in working memory much longer than type case information. This is exactly what one would expect, since a reader needs to keep the word information available to compute such things as grammatical dependencies or pronoun reference. But even here, he said, the memory fades within seconds. In fact, it is possible to show that word order begins to fade just as soon as a grammatical structure is complete. In one recent experiment, for example, subjects were stopped in the middle of reading a passage and asked to recall the precise wording of a phrase from the preceding sentence. They were much less successful than subjects who read a slightly rearranged text where the phrase was part of the current sentence.

■ Sentences that are "grammatical," but hard to understand, generally turn out to be those that put an extra burden on the reader's working memory. As an example, said Just, consider a center-embedded sentence: *The salesman that the doctor met departed.* One of the reasons this is so awkward is that the outer clause (*The salesman . . . departed*) is interrupted in midstream. This means that the reader first has to suspend processing of the initial two words when he or she gets to the central clause (*the doctor met* [*the salesman*]). Then he or she has to remember those two words—thereby taking up valuable space in working memory—while processing the entire central clause; then retrieve the words when the central clause is complete; and finally, associate them correctly with the remainder of the outer clause.

The result is that experimental subjects are only about 85% successful in paraphrasing center-embedded sentences. When presented with doubly center-embedded sentences such as *The salesman that the doctor that the nurse despised met departed*, their performance is little better than random. Moreover, eye fixation studies show that the subjects tend to spend most of their processing time at the most difficult points in the sentences—in this case, at the transition between *met* and *departed*.

Much the same analysis can be applied to other types of awkward usage, said Just. Garden-path sentences, for example: *Mary dropped the cup and the saucer accidentally landed on the rug*. When readers reach the conjunction *and*, they of course try to interpret it on the fly; as a result they assume that it links *cup* and *saucer*. When they read further and realize that *saucer* is in fact the subject of a whole new sentence, they have to back up and undo some of their previous processing. Thus the awkwardness.

■ Individual differences in reading ability seem to arise mainly from differences in the efficiency and capacity of working memory. The key to this finding is the reading span task, which was pioneered in 1980 by Car-

negie-Mellon's Carpenter and Meredyth Daneman: subjects are simply asked to read an ordinary text while remembering the last word of each sentence as they complete it. Thus, someone reading the first paragraph of this article would have to remember the sequence *granted*, *conversation*, *standpoint*, *words*, and *head*.

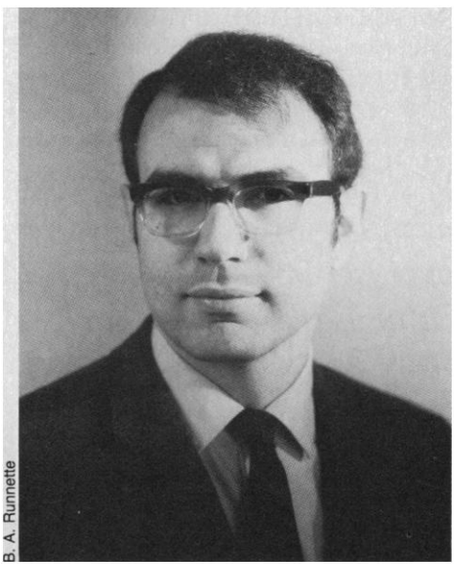
Needless to say, nobody is very good at this. But performance per se is not the point. What Carpenter and Daneman were after was a measure of overall "operational capacity" in working memory, a quantity that would encompass both storage capacity *and* processing power. And indeed, subjects do show distinct variations in performance in the task. Some people can hold as many as five words in memory, while others never do better than two or three.

Intriguingly enough, said Just, the high-span subjects also read faster than their low-span brethren. And they consistently score higher on tests of reading comprehension, such as recognizing what a given pronoun refers to. In fact, said Just, the reading span task has turned out to be one of the best predictors of reading comprehension ever found; the more traditional tests of mental ability, such as remembering strings of nonsense syllables, show much lower correlation.

The Carnegie-Mellon group has accordingly spent a great deal of time during the past several years trying to find out exactly what is going on here. One example pointed out by Just was the recent work of graduate student Jonathan King, who combined the reading span task with the study of awkward sentences. King asked his subjects to remember a sequence of random digits while they read a center-embedded sentence: *The cat that the dog bit ate the mouse*. Meanwhile, he measured the amount of time they spent reading each word.

To no one's surprise, the subjects tended to dwell longest at the point of greatest difficulty, between *bit* and *ate*. In addition, they tended to spend a greater and greater fraction of their time at that point as the number of remembered digits went up. Once again, however, there were significant individual differences. So long as the digit sequences were short, for example, the high-span subjects were able to breeze through the sentence with barely a pause. They only started to slow down at the point of confusion when the number of digits began to exceed their span. And even then they were still the fastest readers.

For the medium-span readers the story was much the same, except that they began to slow down with fewer digits—exactly as one might expect if they had a smaller working memory capacity. And for the low-span readers, as one might also expect, the sentence was difficult even without any digits; they already seemed to be working at maximum capacity.



B. A. Runnette

William Chase. *The late Carnegie-Mellon University psychologist and his colleagues pioneered the theory of skilled memory.*

The obvious conclusion from all this is that the high-span subjects have more raw operational capacity in their working memories, said Just. Now, what about individual differences in processing power?

One way to approach that question is to couple the reading span task with eye fixation studies, as Just and Carpenter have recently done. It turns out that the amount of time the subjects spend on each word can largely be explained by two factors, he said. The first is word length: the longer the word, the longer the fixation. The Carnegie-Mellon group attributes this factor to the encoding process—recognizing that a particular pattern of letters is a word. As it happens, the low-span subjects do just as well on this task as the high-span subjects.

The second factor, however, is word frequency: the more frequent the word, the shorter the fixation. The researchers attribute this to lexical access—retrieving the meaning of a word from long-term memory. And here the high-span subjects do perform better. In fact, said Just, this seems to be one of the main reasons that they read faster than the low-span subjects: they can simply comprehend the words faster.

In his talk at the Carnegie-Mellon Symposium, and in later conversations with *Science*, Just was the first to admit that the story of language and working memory is far from complete. For example, how do these individual differences in processing ability arise? No one really knows, although the answer may well have major implications for educational practice.

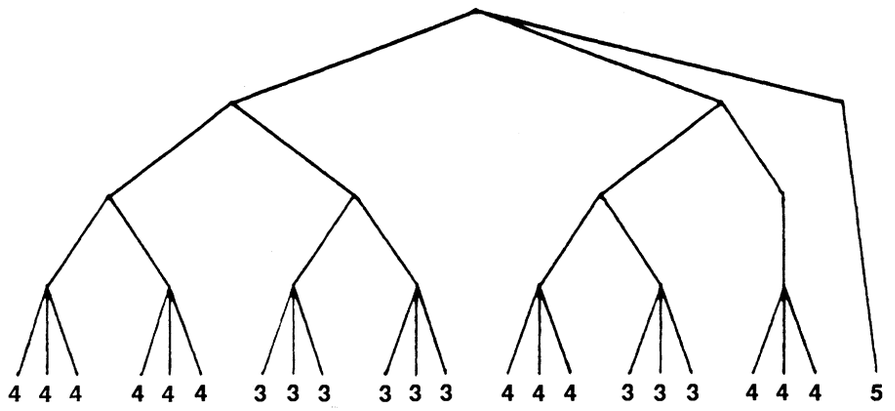
Or consider the question of working memory capacity. The data seem contradictory, said Just. There is abundant evidence that the limits are very tight, with storage space for no more than seven chunks of data at any one time. Indeed, that is the whole point of our on-the-fly processing strategy. And yet, said Just, computer models of language understanding suggest that working memory actually needs to hold dozens of chunks during processing. "The notion of 'seven chunks' of capacity in working memory is incompatible with what we have to do in comprehension," he said. So what is going on?

One possible answer, he suggested, is that the whole concept of "capacity" in working memory is more subtle than it seems: "The total of seven only seems to make sense when you give someone seven homogeneous chunks." he said. "Maybe it doesn't make sense when you get into more complex, dynamic tasks."

Another answer—or perhaps just a more detailed way of giving the same answer—is that working memory is somehow being augmented with long-term memory, which has a nearly infinite capacity. In fact, said Just, there is already a fairly detailed account of how such an augmentation might work: the theory of "skilled memory," which was developed in the early 1980s by the late Carnegie-Mellon psychologist William Chase and his colleagues.

As it happens, skilled memory theory was discussed at the Carnegie-Mellon Symposium by Chase's former student, K. Anders Ericsson of the University of Colorado in Boulder, and in later conversations with *Science* by another of Chase's students, James Staszewski of Carnegie-Mellon. The theory is an outgrowth of cognitive scientists' long-standing interest in the performance of human experts, they explained. In particular, it addresses the fact that some people can perform prodigious feats of memory with very little apparent effort. There are waiters, for example, who can listen to dozens of dinner orders in an evening and keep them all straight—without taking any notes. And there are chess masters who can play dozens of games simultaneously—blindfolded.

Chase and his colleagues concluded that these feats are possible because the memory experts themselves have learned to build elaborate cognitive structures for storing the information. These structures then allow them to shuffle information in and out of their long-term memories so fast that their long-term memories effectively become extensions of their working memories.



How to memorize 80 digits. *This diagram shows the cognitive structure devised by the subject SF. He would first chunk the individual digits into groups of three or four, denoted here by numerals. Then he would keep these groups straight by collecting them into larger structures such as "two groups of four," and so on. The result was an extended hierarchy.*

Perhaps the most dramatic illustration of this principle was provided in the late 1970s by the researchers' experiment with SF, an undergraduate who learned to memorize strings of random digits. SF was no better than average when he started out; like most of us, he reached his limit at about six digits. Nor was he given any mnemonic tricks for the experiment. He simply came to a 1-hour practice session three to five times per week, listened to the digits being read at the rate of one per second, and tried to memorize as many digits in each new string as he could. Within a week his span had started to improve. After 18 months it was up by an order of magnitude. Taking in digits by the dozens, he could rattle them off again with very few mistakes. He could start anywhere in the middle of the string and do the same thing again. He even could start in the middle and go backward. By the time the experiment ended SF could reliably do 84 digits and might well have gone further. (Tragically, the experiment had to end: SF contracted leukemia and later died.)

It was SF's performance, as studied in great detail by Chase and his colleagues, that laid the first empirical foundation for skilled memory theory. Lest that performance seem like a fluke, moreover, Staszewski has recently repeated the experiment using another subject, DD. (The experiment had originally begun under Chase.) By the time the experiment terminated—in this case because DD graduated and moved away from Pittsburgh—he was up to 106 digits. The obvious conclusion is that almost anyone can achieve this kind of memory capacity, given enough time, practice, and motivation.

Perhaps the most intriguing thing about SF's performance, however, was the way he went about learning to memorize the digits. As it happens, he was a cross-country runner. So within the first week of the experiment he hit upon the idea of taking the numbers in groups of three or four, and associating them with racing times. The sequence 3492 thus became "3 minutes and 49.2 seconds—nearly a world record time for the mile." Sequences that did not fit in well with this scheme became ages ("89.3—a very old man"), or even dates ("1944—near the end of World War II"). In other words, SF intuitively rediscovered a classic trick of magicians and performing mnemonists: take the data in small chunks and give each chunk a meaningful, or "semantic" label.

SF continued to use this approach throughout the experiment. Later, Chase deliberately chose DD for the follow-up experiment because DD, too, was a runner; one of the reasons he was able to progress to 106 digits was that he was taught SF's semantic memory system at the beginning. Of course, there is a down side to semantic encoding: it tends to ensure that memory expertise is quite narrow. When SF and DD were tested on their ability to memorize strings of letters, for example, they were still quite ordinary. Their skill with numbers did not transfer at all. Nonetheless, Chase and his colleagues saw semantic encoding as a prerequisite for skilled memory performance.

Learning how to group the numbers semantically took SF well beyond his early limit of six or seven digits. However, he quickly hit another plateau: as soon as he began to accumulate more than three or four groups, he could not keep them straight anymore. The next important advance came only when he began to segment the groups into larger units—"three groups of four," and so forth—so that he could remember the order. This trick in turn took him to yet another plateau where the larger units started to get jumbled. With practice, however, he was able to advance again by introducing yet another level of structure.

And so it went. The eventual result was the hierarchical retrieval structure shown on this page. (DD produced a very similar hierarchy.)

A hierarchy, of course, is one of the simplest ways to organize things. And yet, as Chase and his colleagues were quick to point out, SF's hierarchy had at least one intriguing feature: it never had more than three or four nodes at any given level. This was consistent with even the most constrained models of working memory, since it implied that SF would never have to hold more than three or four chunks in memory at any given time. But it also explained how he could transcend those constraints: what he held in working memory was not the data per se but a set of pointers to the data, which was actually held in long-term memory.

So what does all this have to do with reading? To begin with, said Just, reading is clearly an expert skill: assuming that a child reads for 1 hour per day on the average, then by high-school graduation he or she will have scanned some 20 million words. "Language is one of the few things that almost everyone is an expert at," said Just. "I think that the skill is at least as impressive as that of a chess master."

In addition, he said, the hierarchical memory structures developed by SF and DD are suggestive of the kind of mental structures people seem to build as they are reading. The analogy is not perfect, of course. "In the digit-span task, the subjects learn a certain number of structures for packing in the information quickly and getting it back quickly," said Just. "A reader is also packing in information quickly. But he doesn't know what the high-level structure is ahead of time; his task is to discover and construct it."

On the other hand, he said, written text is full of cues that play a similar role. Topic sentences, parallel construction, emphasis on the first and last words of a paragraph, conventional plot devices—in short, the whole corpus of rhetorical tricks that have been taught by generations of English teachers, all seem to be designed to help the reader structure the information in the text and to guide it smoothly into his mind. ■

M. MITCHELL WALDROP

ADDITIONAL READING

M. A. Just and P. A. Carpenter, "A theory of reading: from eye fixations to comprehension," *Psychol. Rev.* **87**, 329 (1980).

M. Daneman and P. A. Carpenter, "Individual differences in working memory and reading," *J. Verbal Learn. Verbal Behav.* **19**, 450 (1980).

K. A. Ericsson, W. G. Chase, S. Faloon, "Acquisition of a memory skill," *Science* **208**, 1181 (1980).