JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 7, 560–566 (1968)

Short-Term Memory for Sentences[1]

EDWIN MARTIN, KELYN H. ROBERTS, AND ALLAN M. COLLINS*

University of Michigan, Ann Arbor, Michigan 48104

Short-term memory for active and passive sentences at two levels of grammatical complexity was tested at four retention intervals, 0, 10, 20, and 40 sec. Sentence forgetting was analyzed in terms of differential word-class forgetting. It was hypothesized that *S*s selectively focus on key word classes, with grammatical structure as the guide to selection, and generate recall sentences around retained elements.

The experimental literature on the role of grammatical factors in sentence retention contains no study that relates either amount or form of grammatical structure to retention interval in the short-term-memory situation. The expected returns from such an experiment are considerable. For example, it is of theoretic interest to know the time-course of forgetting of sentences of different kinds (e.g., actives, passives); similarly for sentences of different structural complexity. Of particular interest to the present writers is the possibility of differential forgetting of certain word classes. The reason for this last interest is that recall errors in sentence kind should be predictable from word-class effects. Moreover, since sentences of different structural complexity frequently involve different word classes (e.g., highly complex sentences tend to have more adverbs), it is not unreasonable to expect that differential sentence forgetting according to structural complexity might also be traceable to word-class effects. Thus, the experiment reported here involved short-term memory for sentences of two kinds (active, passive) at two levels of structural complexity. Four retention intervals (0, 10, 20, 40 sec) were used, where the interval activity was counting backwards by threes.

The determination of sentence complexity has been explicated by Martin and Roberts (1966). It is based on a sentence-generation model proposed by Yngve (1960), which can be summarized as follows:

> In order to generate a sentence by means of binary rewrite rules, it is necessary to hold in memory, during each rewrite operation, the node on the right as the node on the left is expanded. Consider the sentence *The new truck has very good brakes.* Listed below, preceded by asterisks, are the rewrite operations required to generate this sentence. Operations not preceded by an asterisk translate word classes into words. To the right of each operation is shown, in square brackets, the specific nodes being held in memory during the operation.[2]

*S	→ NP_1 + VP	[–]
*NP_1	→ T + NP_2	[VP]
T	→ *The*	[NP_2, VP]
*NP_2	→ Adj_1 + N_1	[VP]
Adj_1	→ *new*	[N_1, VP]
N_1	→ *truck*	[VP]
*VP	→ V + NP_3	[–]
V	→ *has*	[NP_3]
*NP_3	→ AdjP + N_2	[–]
*AdjP	→ Adv + Adj_2	[N_2]
Adv	→ *very*	[Adj_2, N_2]
Adj_2	→ *good*	[N_2]
N_2	→ *brakes*	[–]

[1] This research was supported by the Advanced Research Projects Agency, Department of Defense, monitored by the Air Force Office of Scientific Research, under Contract No. AF 49(638)-1235 with the Human Performance Center, Department of Psychology.

* Now at Bolt, Beranek and Newman, Cambridge, Massachusetts.

[2] *S* = sentence; *NP* = noun phrase; *VP* = verb phrase; *T* = article; *Adj* = adjective; *N* = noun; *V* = verb; *AdjP* = adjective phrase; *Adv* = adverb.

Note that the next node to be operated on is, in each case, the node most recently placed into memory.

Our immediate concern is with the number of nodes in memory at the time a terminal word is written out, and how these numbers may be summarized to yield an index of sentence structural complexity. For the seven word classes of this sentence, the number of nodes in memory are 2, 2, 1, 1, 2, 1, and 0, respectively. We have called these numbers "Yngve numbers." The Yngve number of a word class thus represents the depth to which the word class is structurally embedded in the sentence.

Although there are many ways to summarize the foregoing embeddedness vector, we have chosen the arithmetic mean; hence our measure of sentence complexity is mean Yngve number, or sentence *mean depth.* For the illustrative sentence above, its mean depth is

$$\text{Mean depth} = \frac{\sum_{i=1}^{l} y_i}{l} = \frac{9}{7} = 1.29,$$

where y_i is the Yngve number of the ith word and l is the length of the sentence in words.

There are several basic properties of Yngve numbers worth noting. First, they carry only structural information; such factors as word-class uncertainty are not reflected. This is because Yngve numbers are assigned to word classes on the basis of structural relations between word classes. Second, an Yngve number must always be a non-negative integer: it is zero only for the final position in the sentence; if y_i is the Yngve number of the ith position, then y_{i-1} cannot be larger than $y_i + 1$; and if there are l words in the sentence, the largest Yngve number cannot be larger than $l - 1$. Given these properties, mean depth of a sentence of length l is bounded by

$$\text{Min(mean depth)} = \frac{l-1}{l},$$

and

$$\text{Max(mean depth)} = \frac{l-1}{2}.$$

Owing to the grammatical structure of sentences, however, the maximum mean depth is never realized in usage. [See Yngve (1960) for a detailed discussion of this last point.]

The psychological relevance of Yngve's model is rather easy to state. It is not implausible, as a first approximation, to assume that the phrase-structure model of a given sentence represents the hierarchy of habits a language user exercises both in understanding and in generating that sentence. Thus, for the illustrative sentence above, perception of the first word, *The*, is seen to elicit in the listener, by virtue of its intonation and word-class membership, the expectation of two further constituents; namely, a noun phrase (NP) and a verb phrase (VP) of some sort. In this sense, the psychological interpretation of the Yngve number of that word is number of expected constituents. In general, the constituents in square brackets to the right of a given word (see the list of operations above) are assumed to be the constituents elicited in expectation by that word. Since the learned constituent expectations experienced by the listener match the constituent constraints a speaker must follow, communication is possible. Should the speaker violate the constraints attaching to his choice of intonation and word class, the listener may fail to comprehend, and hence the probability that a meaningful message will be relayed is reduced. The Yngve numbers of a sentence thus index learned constraints that are common to speaker and listener. We have shown elsewhere (Martin and Roberts, 1966) that sentence complexity, as measured by the mean of these Yngve numbers, significantly affects sentence retention in the free learning situation.

METHOD

Materials. Thirty-two seven-word sentences were constructed such that there were eight sentences for each of the four combinations of two kinds, active (A) and passive (P), and two mean depths (or levels of complexity), 1.00 and 1.86. All A and P sentences of mean depth 1.00 had Yngve numbers 1, 2, 1, 1, 1, 1, 0; all A and P sentences of mean depth 1.86 had Yngve numbers 1, 4, 3, 2, 2, 1, 0. The word classes of the sentences of each of the four kind-mean depth combinations are given in Table 1. The following are experimental exemplars of the four combinations:

A-1.00, *Soldiers will appreciate the short training session;* A-1.86, *Doctors are now adopting the practice slowly;* P-1.00, *Traps were set by the Indian hunters;* P-1.86, *Laws are often passed by Congress quickly.*[3]

[3] A list of the sentences is available from the author or from the American Documentation Institute. To obtain a copy, order Document No. 9910 from the Chief, Photoduplication Service, Library of Congress, Washington 25, D.C., Auxiliary Publications Project, remitting $1.25 for 33 mm. microfilm or $1.25 for 6 × 8 in photo copies.

Four different random orders of the 32 sentences were constructed. A retention interval of 0, 10, 20, or 40 sec was then assigned to each sentence such that (a) over the four orders every sentence was assigned each of the intervals once, and (b) within blocks of four sentences in a given ordering all four intervals occurred once. Thus, over the four orders every sentence was tested at every retention interval, and within an ordering the several retention intervals were distributed rectangularly.

Each sentence was assigned a different three-digit number. It was from this number that the *S*s counted backwards by threes for the duration of the retention interval.

The four orders, together with (a) a warning buzzer that preceded each sentence by approximately 2 sec, (b) the assigned number that immediately succeeded each sentence, and (c) a buzzer that signaled the end of the retention interval, were recorded on magnetic tape in a circular fashion so that following the last sentence of the ordering, the ordering began again. This last was done so that each of the *S*s assigned to a particular ordering could begin with a different sentence.

The rehearsed intonation and rate of enunciation of the recorded sentences were those of ordinary, conversational discourse. The materials were presented to the *S* via earphones. The *S*'s recalls were both scored on the spot by the *E* and recorded on magnetic tape for later checking and, in the case of recall errors, transcription.

Subjects. The *S*s were 32 University of Michigan undergraduates who volunteered for paid participation. Eight *S*s were randomly assigned to each of the four orderings.

Procedure. Each *S* was told that his memory for English sentences was being tested. The ready signal (buzzer) preceded presentation of each sentence by approximately 2 sec. The sentence was then presented, followed immediately by the three-digit number. The *S* began counting backwards by threes until stopped by the buzzer, whereupon he attempted to recall the sentence as it was given. For the 0-sec retention interval, the recall signal was sounded immediately after *S* heard the number, thereby making the actual retention interval approximately 1 sec. This was done on the view that if all sentences were followed by a number, less variability in "set" might result. Delay of the recall signal for the other retention intervals (10, 20, and 40 sec) was measured from the end of the sentence.

The *S*s had unlimited recall time. Most *S*s indicated verbally they could not remember the sentence when such was the case. The *E* started the next trial sequence when it was clear that the preceding recall attempt was finished.

The attention of the *S* was drawn to the counting activity that filled the retention interval by offering him a bonus directly related to the accuracy and speed with which he counted backwards. He was told that his base pay was $1.00, but that he could earn up to $1.75. The 32 recall events, plus instructions, required approximately 30 min of the *S*'s time.

Results

Whole-Sentence Correct Recall. By "whole-sentence correct recall" is meant a sentence in recall that matches exactly the presented sentence. For each combination of sentence kind, mean depth, and retention interval, there were (32 *S*s) × (2 sentences) = 64 experimental events. The proportion of these that were correct recalls are shown in Table 2 for each of the 16 combinations of sentence kind, mean depth, and retention interval.

TABLE 1

Word-Class of Each Position for the Four Kind-Mean Depth Combinations

Word	Sentence kind-mean depth				Legend
position	A-1.00	A-1.86	P-1.00	P-1.86	for Fig. 1
1	Agt	Agt	Obj	Obj	● —— ●
2	Aux	Aux	Aux	Aux	○ - - - ○
3	MV[a]	Adv$_1$	MV	Adv$_1$	▲ —— ▲
4	T	MV	by	MV	△ - - - △
5	Adj$_1$	T	T	by	■ —— ■
6	Adj$_2$	Obj	Adj	Agt	□ - - - □
7	Obj	Adv$_2$	Agt	Adv$_2$	× - - - - ×

[a] MV = main verb.

TABLE 2

PROPORTION CORRECT SENTENCE RECALLS OUT OF 64 OPPORTUNITIES

Sentence kind-mean-depth	Retention interval			
	0	10	20	40
A-1.00	.55	.39	.41	.22
A-1.86	.73	.58	.31	.34
P-1.00	.77	.70	.59	.42
P-1.86	.59	.44	.25	.25

As can be seen, there is a strong interval effect, $F(3, 93) = 26.06, p < .001$: with lengthening retention interval, proportion of correct recalls declines regularly. The main effect of mean depth is also statistically significant, $F(1, 31) = 4.80, p < .05$, and takes the form of poorer recall of presented sentences of mean depth 1.86. More important than this latter however, is the interaction between kind and mean depth, $F(1, 31) = 54.74, p < .001$: for A sentences, presented sentences of mean depth 1.86 were recalled slightly better than presented sentences of mean depth 1.00, $t(62) = 1.96$, $p > .05$; but for P sentences, presented sentences of mean depth 1.86 were recalled significantly more poorly than presented sentences of mean depth 1.00, $t(62) = 5.04$, $p < .001$.

If the .05 level is accepted as criterion, the main effect of sentence kind was not statistically significant, $F(1, 31) = 3.13, p > .05$. The only additional effect not mentioned so far that was of statistical significance is the interaction between mean depth and interval, $F(3, 93) = 3.60$, $p < .01$, which is traceable to the depressed proportions for the A-1.86 and P-1.86 combinations relative to the A-1.00. combination at the 20-sec retention interval.

Frequency of Error Types. Collapsing over the 16 combinations of presented-sentence kind and mean-depth and retention-interval there were (32 *S*s) × (32 sentences) = 1024 experimental events. There were 483 completely correct recalls. Of the 541 errors, 432 were complete-sentence errors of the same syntactic kind as their present versions, 22 were complete-sentence errors of a different syntactic kind than their presented versions, 34 were incomplete (partial) recalls, and 53 were complete omissions. Thus, combining the frequencies of completely-correct recalls (483) and recall errors that were of the same syntactic kind (A or P) as their presented versions (432), it is seen that in 89.4% of the recalls the grammatical form of the presented sentence was retained. By comparison, the incidence of "transformation" errors (A → P or P → A) is quite low: only 22, or 2.1%, of the 1024 recall events.

In general, all four error types increased somewhat with retention interval. The rates of increase, however, were not distinctive over the four combinations of presented sentence kind and mean depth. The most marked effect of interval was on omission errors: collapsing over presented-sentence kind and mean depth, the number of omissions for the intervals 0, 10, 20, and 40 sec were 1, 3, 5, and 34, respectively.

Mean Depth of Complete-Sentence Errors. As noted above, there were 454 complete-sentence errors. Each of these sentences was analyzed according to phrase-structure grammar and a mean-depth value assigned. It was found (a) that error mean depth is not a function of retention interval and (b) that error mean depth regresses toward a value somewhere between the two mean-depth levels (1.00 and 1.86) of the presented sentences. [That the latter is not a within-*S*s-design effect is attested to by the fact that the same phenomenon was observed by Martin and Roberts (1966), where presented-sentence mean depth was a between-*S*s factor.] For the four combinations of presented-sentence kind and mean depth, the error mean depths varied randomly and very closely about the following mean depths: 1.06 for the A-1.00 combination; 1.14 for P-1.00; 1.61 for A-1.86; and 1.63 for P-1.86.

Correct Recall of Words. For each of the word classes shown in Table 1, the proportion of the assigned words that occurred in recall was determined for each of the retention intervals. A word was considered retained if it occurred in recall, regardless of other errors in the sentence; a word was considered not-retained if it did not appear in recall, which includes the case of omitting the entire sentence. These data are shown in Fig. 1, where each point is based on 64 recall events.

From Table 2, it is clear that P-1.00 sentences were best recalled at all retention intervals. The word-class effects of Fig. 1 accordingly show no marked differential word-class effects. This is to be contrasted with the other combinations of sentence kind and mean depth, where poorer whole-sentence recall (Table 2) is associated with clear word-class effects (Fig. 1).

Consider the difference in mean depth for A sentences. From Table 2, there is superiority of A-1.86 sentences over A-1.00 sentences at all retention intervals except the 20-sec interval. Note the corresponding word-class effects in Fig. 1. The word classes Adj_1 for A-1.00 sentences and Adv_1 for A-1.86 sentences are the most labile and show large, divergent deletion effects at the 20-sec interval. This is less true of the next most labile word classes for the two kind-mean depth combinations [Adj_2 for A-1.00 sentences, *Aux* (auxiliary verb) for A-1.86 sentences]. It is also of interest to note that for the more structurally complex (A-1.86) of the two kind-mean depth combinations under consideration

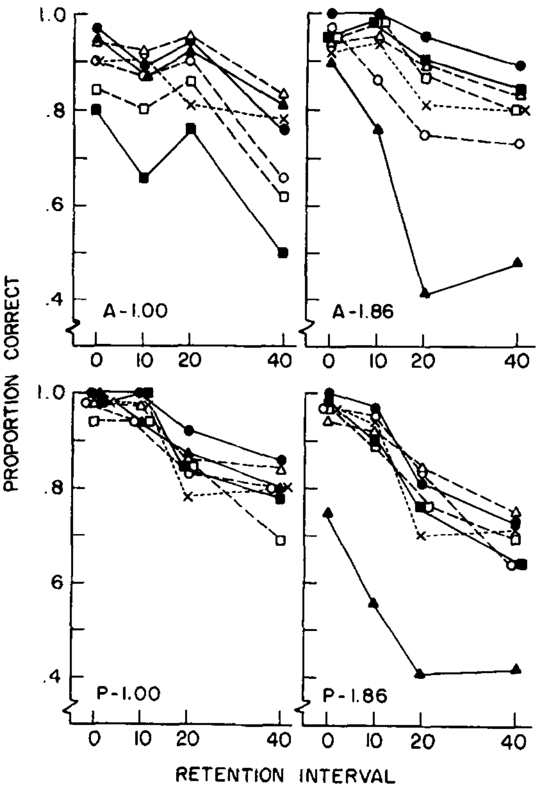

FIG. 1. Proportion of words of a given word class correctly recalled as a function of retention interval for the four combinations of presented-sentence kind and mean depth. See Table 1 for legend.

(A-1.00 and A-1.86), both *Agt* (agent, i.e., logical subject) and *Obj* (logical object) are better retained. This last emphasizes the idea that grammatical structure serves to focus the *S*'s processing (attentional) facilities on key elements, to the detriment of less critical elements.

Consider now the difference in mean depth for P sentences. From Table 2, there is uniform superiority of P-1.00 sentences over P-1.86 sentences at all retention intervals. This effect is traceable primarily to the lability of Adv_1 in P-1.86 sentences. Although there is a general inferiority of word classes for P-1.86 sentences at the 10-, 20-, and 40-sec intervals, the magnitude of this inferiority is not sufficient to account for the large difference between proportion whole-sentence correct recalls for P-1.86 and P-1.00 sentences as shown in Table 2. As with A sentences, then, it appears that poorer whole-sentence recall is traceable primarily to certain word-class effects.

Discussion

The principal results are the following: (a) Ordinary English sentences were forgotten over retention intervals up to 40 sec in somewhat the same manner as more isolable verbal units (e.g., consonants, trigrams, words). A noticeable difference is that, by comparison with these more isolable units, immediate sentence recall is rather poor: the proportion of correct recalls for the four combinations of sentence kind and mean depth ranged between .55 and .77 for the 0-sec interval. (b) Mean depth interacted significantly with sentence kind. Among the passive sentences, less structural complexity (mean depth = 1.00) allowed better recall than more structural complexity (mean depth = 1.86); while the reverse was true among active sentences. (c) The foregoing two results covaried with several distinct types of errors, the most intriguing of which were differential word-class errors. The most labile of the word classes were *Adv*, *Adj*, and *Aux*, and these were differentially represented in the four kind-mean depth combinations. (d) The structural complexity (mean depth) of complete-sentence errors was not related to retention interval. Although frequency of complete-sentence errors increased with retention interval, the apparent strength of tendencies to simplify initially complicated sentences and to complicate initially simple sentences remained invariant. These same tendencies were reported by Martin and Roberts (1966), where the free-learning technique was used.

Probably the most important result of the present experiment is the relation between the proportion of correct sentence recalls over the four combinations of sentence kind and mean depth and differential forgetting over word classes. From Table 2 it is clear that P-1.00 sentences were best recalled. From Fig. 1 it is clear that the least variability in retention over word classes occurred among P-1.00 sentences. Note further from Table 2 that the proportion of correct recalls of A-1.86 sentences at the 0-sec retention interval approximated that of P-1.00 sentences at the same retention interval. At longer retention intervals, recall of A-1.86 sentences was decidely poorer than recall of P-1.00 sentences. Concomitant with this is the fact that among A-1.86 sentences, variability in retention over word classes was small at the 0-sec retention interval, but large at longer retention intervals. What is suggested by these considerations is that there exists a relation between variability in word-class retention effects and likelihood of correct sentence recall. If the ranges of proportion of words correct over word classes for each of the 16 combinations of kind, mean depth, and retention interval (see Fig. 1) are examined for covariation with proportion of correct sentence recalls for these 16 combinations (see Table 2), it turns out that the product-moment correlation is $-.80$ ($df = 14, p < .01$). Thus the more variably do *S*s process different word classes, the less likely are they to generate in recall an exact replica of the presented sentence.

If word classes are differentially processed into memory, then it must be that the rules governing this selective processing are the rules of grammar. In many instances, the functional class of a given word is not determined until it becomes embedded in a sentence. Simpson (1965) has shown that with increasing approximations to English, the proportion of total errors in serial learning attributable to the forgetting of adverbs (*Adv*) increases steadily, while the proportion attributable to the forgetting of nouns (*Agt*, *Obj*) decreases steadily. Thus as more grammatical structure is introduced, there arises greater variability, or increased selectivity, in the processing of different word classes.

The results of the present experiment indicate that word-class analyses may be a productive approach to the problems of behavioral linguistics. A reasonable corresponding hypothesis is that word classes, when in a sentential context, are differentially attended. The view that occurs to the present writers is that *S*s selectively focus on key elements of the input string, with grammatical structure acting as the functional stimulus that directs his selection. Recall, then, would consist of generating a grammatical English sentence that incorporates specificially the key elements and less specifically any non-key elements that were retained.

Apropos of the foregoing comment are results recently obtained by Roberts (1966). In a short-term-memory experiment, A and P sentences were constructed so that N_1 (*Agt* for A sentences, *Obj* for P sentences) and N_2 (*Obj* for A sentences, *Agt* for P sentences) were meaningfully interchangeable. For example, a P sentence might be *The solider was watched by the man.* An equally likely P sentence may be obtained by exchanging the nouns *soldier* and *man.* Roberts also manipulated sentence complexity, using mean depths 1.14 and 1.86. Presented-sentence kind was a within-*S*s variable, mean depth a between-*S*s variable. For each mean-depth level, 32 *S*s were tested on eight sentences, four of each of the two kinds. The retention interval was fixed at 20 sec, and the interval was filled with a counting activity.

Of the (32 *S*s) × (4 sentences) = 128 retention tests for each of the four combinations of sentence kind and mean depth, the following numbers of complete-sentence errors that were transformations of syntactic kind were observed: for presented A-1.14 sentences, 13 P errors; for A-1.86, 9 P errors; for P-1.14, 7 A errors; and for P-1.86, 9 A errors. The data of interest are the frequencies with which these errors in kind ("transformation" errors) did not involve the appropriate reversal in position of N_1 and N_2. These frequencies are 11, 6, 4, and 7 for the combinations A-1.14, A-1.86, P-1.14, and P-1.86, respectively. The point to be made is that out of the 38 errors in syntactic kind, 28, or 74%, involved failure to reverse appropriately N_1 and N_2. In contrast, among the 175 complete-sentence errors that were of the same syntactic kind as their presented versions, only 27, or 15%, involved inappropriate reversal of N_1 and N_2. These data support the idea that in processing sentential messages, the *S*s focus on key elements, to the detriment of other, sometimes informative elements, and generate in recall an essentially "new" sentence based on the elements retained.

References

Martin, E., and Roberts, K. H. Grammatical factors in sentence retention. *J. verb. Learn. verb. Behav.*, 1966, **5**, 211–218.

Roberts, K. H. The interaction of normative associations and grammatical factors in sentence retention. Paper read at Midwestern Psychological Association, 1966.

Simpson, W. E. Effects of approximations to sentence word-order and grammatical class upon the serial learning of word lists. *J. verb. Behav. verb. Learn.*, 1965, **4**, 510–514.

Yngve, V. A model and hypothesis for language structure. *Proc. Amer. Phil. Soc.*, 1960, **104**, 444–466.

(Received March 3, 1967)