JOURNAL OF VERBAL LEARNING AND VERBAL BEHAVIOR 5, 211-218 (1966)

Grammatical Factors in Sentence Retention[1]

EDWIN MARTIN AND KELYN H. ROBERTS

University of Michigan, Ann Arbor, Michigan

A rationale for indexing the structural complexity of sentences was introduced and an experiment reported that demonstrated the relationship between this index and sentence retention. The proposed measure entails a phrase-structure analysis of the sentence and a counting of the grammatical commitments incurred by each word of the sentence. A word is said to be structurally embedded in a sentence to the extent that it determines the structure of those parts of the sentence that follow.

In a six-trial free-learning experiment where sentence complexity and sentence kind were manipulated independently and sentence length held constant, sentences of lesser indexed complexity were recalled significantly more frequently than sentences of greater complexity. The role of sentence kind was found to affect recall, but not in the systematic way predicted by the transformation-grammar model.

The research to be described is an application of a certain conceptualization of sentence structure to sentence-recall behavior. More specifically, a procedure for quantifying the structural complexity of sentences has been developed and an experiment conducted to assess its utility in accounting for recall performance when the to-be-remembered sentences vary according to this complexity measure.

The measure of interest was proposed by Yngve (1960). What it does is to assign a number to each word of a sentence so that, in essence, the more embedded in the sentence a word is, the larger the number assigned. It is intended here to speak of embeddedness only in a structural sense, although it is recognized that a complete separation of syntactic and semantic factors is probably not attainable.

Consider first a plausible psychology of listening to, and reproducing, ordinary English sentences. As a listener receives a grammatical string of words, not only does he perceive each word as it arrives, but in addition, upon the arrival of each, he makes an encoding response a major constituent of which is formation of an expectation as to what is coming next. Such responses are the result of extensive experience with the sentence structure of the language; they are responses that have been acquired in the listener's language community. What particular responses are elicited by each word of the arriving sentence is dictated by (*a*) the speaker's choice of word class for that position in the sentence and his manipulation of pitch, stress, and juncture—in short, the stimulus situation for the listener; and (b) the listener's knowledge of word-word and class-class transition probabilities and of what variations in pitch, stress, and juncture signalize in his language—in other words, the listener's existing habit structures. Thus as a sentence is received, the intonation and prosody of the lengthening string serve as stimuli for anticipatory responses where what is anticipated is a partial ordering of particu-

[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.

lar word classes. The position taken by the present writers is that these responses are a major component of what is placed in memory and that the likelihood of a correct recall depends inversely upon the number of such responses the listener makes in attempting to store the sentence in memory.

In consideration of a speaker as he utters a sentence (as in recall), an essentially parallel analysis is appropriate. "It seems that, as we speak, we incur commitments to finish our sentences in certain ways in order to make them grammatical" (Yngve, 1964, p. 277). As a string of words lengthens, such commitments must exist in the speaker's memory if he is to complete the string in good grammatical form. The parallel, then, between the speaker and the listener is that for both there is a sequence of entries into memory. It will be argued that for each word in a given sentence, the number of expectations elicited in the listener is the same as the number of commitments incurred by the speaker.

Suppose the sentence *The new club member came early* were read to a listener. Upon hearing *The,* the listener responds with the following two anticipations: he expects to hear the rest of the noun phrase just begun with *The,* and he expects a predicate of sort. Correspondingly, in uttering the same sentence, the speaker incurs two commitments upon saying *The:* one of them is to finish the noun phrase just begun with *The,* the other is to provide a predicate. Thus from psychological considerations, *The* may be said to be structurally embedded to a depth of 2, "structurally" because the expectations and commitments are grammatically founded. The next word, *new,* is also assigned a depth of 2; this is because receipt of *new* elicits in the listener an expectation of completion of a new noun phrase and affirms the already-elicited expectation of a predicate. Again, a corresponding statement can be made for the speaker. Similarly, *club* has depth 2. The noun *member,* however, has depth 1 because the only expectation (commitment) in effect is that of a predicate. With the verb *came,* the expectation of (commitment for) a predicate is met, but only partially: its intonation elicits (incurs) in turn an expectation of (commitment for) an adverb and hence is itself embedded to a depth of 1. And finally, the pitch and stress of the adverb *early* indicate that *early* is the terminal word and therefore has depth 0.

The sentence *The new club member came early* can thus be characterized by the following set of numbers: 2, 2, 2, 1, 1, 0. These numbers reflect the structural involvement of each word in the sentence from the point of view of both a listener and a speaker. If the minimal psychology of listening and reproducing just outlined is tenable, these numbers also should serve to index how much of a load on memory is imposed by the sentence.

A formal procedure for determining such a characterizing set of numbers for any sentence (in any language, actually) has been detailed by Yngve (1960). It consists of drawing up a binary phrase-marker tree for the sentence in question and then counting the number of left branches leading to each word. The mean of this set of what might be called Yngve numbers may be taken as a measure of the structural complexity of the sentence as a whole. For a given sentence, then, let its mean depth be formally defined as the mean of its Yngve numbers.

The hypothesis under scrutiny in the present research is that the likelihood of recall of a sentence is inversely related to the mean depth of that sentence. Accordingly, one of the independent variables was sentence complexity as indexed by the mean-depth measure.

The second grammatical factor of concern here is that of sentence kind. Considerable attention has been directed to the idea that the corpus of English sentences can be partitioned into kernels and nonkernels. A kernel is an active, affirmative sentence. Negative, passive, passive-negative, and interrogative sentences exemplify nonkernels. Formally, Chomsky (1957) proffers the notion that the basic rules of grammar should apply to kernels and that a set of transformation rules should be used to derive the nonkernels from the kernels; psychologically, Miller (1962) suggests that a nonkernel sentence is remembered by first transforming it to its underlying kernel and then remembering the kernel plus a transformation tag of some sort. On either level, the passive sentence, say, *Assignments were arbitrarily made by the foreman* would be treated (analyzed, remembered) as the kernel *The foreman arbitrarily made assignments* plus a transformation (grammatical rule, memory tag) to the passive. The grammatical utility of such a view is clear (Bach, 1964; Chomsky, 1965); but whether or not such processes occur within the human information handler is not clear. Mehler (1963), Miller (1962), and Miller and McKean (1964) have pressed persuasively, with data, for recognition of automatic transformations among sentence kinds as a cognitive process. However, the possibility exists that distinctions like kernel vs. nonkernel do not as such bear upon recall performance, but that the effects so far observed and attributed to transformation processes can more plausibly, and more parsimoniously, be explained in terms of a structural index like sentence mean depth. Therefore, the second independent variable was sentence kind.

Method

On each of six free-learning trials, the same six sentences were read to *S*, who then proceeded immediately to recall them as best he could. Two factors were manipulated orthogonally: there were six sentence kinds within *S*s and two levels of sentence mean depth between *S*s.

Materials. Twelve sets of six sentences each were constructed such that each set contained one each of the following six sentence kinds: kernel (K), passive (P), truncated passive (P_T),[2] negative (N), passive-negative (PN), and truncated passive-negative (P_TN). Of the 12 sets of sentences, six were made up of sentences all of which had Yngve numbers 1, 3, 2, 1, 1, 1, 0, and hence mean depth 1.29; the other six sets had sentences with Yngve numbers 1, 4, 3, 2, 1, 1, 0, and hence mean depth 1.71. For example, *They were not prepared for rainy weather* is a P_TN with mean depth 1.29, while *Children are not allowed out after dark* is a P_TN with mean depth 1.71. All sentences were exactly seven words long. The sentences within each set of six were completely unrelated in the judgment of the investigators.

Subjects. The *S*s were 120 University of Michigan undergraduates who volunteered for paid participation. As a rule, they were run in groups of five; however, occasionally only one or two were run at a time in order to fill out a group. Ten *S*s were assigned to each of the 12 sets of sentences.

Procedure. Five *S*s at a time were seated in standard classroom desk-chairs in an otherwise empty room (except for *E*'s table). They were each given one sheet of blank paper and told that they would hear six ordinary English sentences read aloud by *E* in immediate succession and in a normal speaking voice; and that after the last sentence had been read, *E* would say "start," at which time they were to begin writing out as many of the just-heard sentences as they could remember. As soon as an *S* finished writing, *E* picked up the sheet on which the recalled sentences had been written. After all five *S*s were finished, fresh sheets of paper were distributed and *E* read the same six sentences again, in a new order, and the *S*s again attempted to reproduce them. In all, six such trials were effected. Every sentence occurred once in each ordinal position over the six trials.

Results

Correct Recalls. For each combination of sentence mean depth and sentence kind, the mean number of correct recalls over the six trials per *S* was computed. These data are summarized in Table 1, where each mean represents 60 *S*s. Trial data are not presented because for each sentence kind the difference between mean depths 1.29 and 1.71 is essentially constant over trials; the magnitude of the difference apparent in the over-trials

[2] A passive is truncated when the agent is not specified. *We were met by our two children* is not truncated because the agent (our two children) is named; however, *The power was turned off at five* is truncated because who turned the power off is not given.

TABLE 1
MEAN NUMBER CORRECT RECALLS PER *S* IN SIX TRIALS AND MEAN THORNDIKE-LORGE COUNT

Sentence kind	Sentence mean depth			
	1.29		1.71	
	Recall	T-L[a]	Recall	T-L[a]
K	3.27	80.2	2.17	86.8
P	3.99	73.3	2.97	80.2
$\mathrm{P_T}$	3.27	84.0	2.50	80.3
N	4.71	91.0	3.29	91.0
PN	3.94	84.5	4.04	88.0
$\mathrm{P_TN}$	3.97	90.3	3.46	89.9
Mean	3.86	83.9	3.07	86.0

[a] In the computation of means, A = 50, AA = 100.

means of Table 1 appeared on the very first trial and remained intact as the likelihood of correct recall increased over the succeeding five trials.

A 2 × 6 analysis of variance indicates that sentence mean depth is a highly significant factor, $F(1,118) = 18.22$, $p < .001$; that sentence kind is similarly highly significant, $F(5,590) = 15.45$, $p < .001$; and that the interaction between depth and kind is significant, $F(5,590) = 3.62$, $p < .01$.

Subsequent to conduct of the experiment, the average Thorndike-Lorge (1944) count for each sentence was computed. The means for the 12 combinations of sentence mean depth and sentence kind are shown in Table 1. The product-moment correlation between the total number correct recalls and the Thorndike-Lorge count is $r = .15$ over the 72 sentences used, which for $df = 70$ is not statistically significant. A 2 × 6 analysis of covariance does not change the picture revealed by the 2 × 6 ordinary analysis of variance.

In view of the significant interaction between sentence mean depth and sentence kind, the difference between mean depths 1.29 and 1.71 was examined for each sentence kind. On the basis of the t test with $df = 118$, for kinds K and P the differences are significant beyond the .005 level; for $\mathrm{P_T}$, beyond the .05 level; for N, beyond the .001 level. For kinds PN and $\mathrm{P_TN}$ the differences are not significant ($p > .30$ and $p > .10$, respectively). Again, the picture does not change under covariance analysis with Thorndike-Lorge count as the covariate.

Errors. In Table 2 are listed the number and kinds of errors arising from each combination of sentence mean depth and sentence kind. For example, in the first column one can see that out of the (60 *S*s) × (6 trials) = 360 response opportunities for recalling a K of mean depth 1.29, there occurred 164 errors, of which 61 were omissions. Of the remaining 103 overt errors, 71 were Ks, 3 were Ps, 5 were $\mathrm{P_T}$s, etc. From the bottom row of the column, one reads that the mean depth of the 96 grammatical complete-sentence errors is 1.28.

Aside from omissions, the most frequent type error was a recall error of the same sentence kind as the presented sentence. These are entered on the main diagonal and are attributable chiefly to substitutions of adverbs and adjectives. Those (grammatical complete-sentence) errors off the main diagonal are transformations of the presented sentence. The proportion of grammatical complete-sentence errors that are transformations is .30 for presented sentences of mean depth 1.29, and .26 for presented sentences of mean depth 1.71. This difference is not statistically significant ($z = .47$, $p = .64$). Of the total number of off-diagonal errors (transformation errors), the proportion that are Ks is .32 for presented sentences of mean depth 1.29, and .15 for presented sentences of mean depth 1.71. This difference is highly significant statistically ($z = 3.56$, $p < .001$).

The average mean depth of grammatical complete-sentence errors over the six kinds is 1.26 for presented sentences of mean depth 1.29, and 1.56 for presented sentences of mean depth 1.71. This means that presented sentences of indexed complexity 1.29 give rise to errors of similar complexity, whereas for more complex presented sentences, those with mean depth 1.71, resulting errors are simpler structurally.

TABLE 2
ERRORS IN RECALL

Type of error	Presented sentence											
	K		P		P_T		N		PN		P_TN	
	1.29	1.71	1.29	1.71	1.29	1.71	1.29	1.71	1.29	1.71	1.29	1.71
K	71	68	9	7	15	9	10	8	1	2	7	2
P	3	1	47	105	10	3	0	0	0	11	0	0
P_T	5	58	7	9	70	118	1	0	1	0	6	3
N	16	14	1	1	3	2	32	85	5	13	6	20
PN	0	0	3	8	0	0	0	0	43	48	2	0
P_TN	1	8	1	1	9	6	2	0	8	0	38	83
Incomplete	6	7	6	10	3	5	3	7	5	4	2	3
Ungrammatical	1	0	0	0	0	0	0	1	0	0	0	0
Omissions	61	74	47	41	54	67	30	62	61	40	61	42
Total	164	230	121	182	164	210	78	163	124	118	122	153
Mean depth[a]	1.28	1.47	1.19	1.56	1.23	1.53	1.24	1.63	1.34	1.65	1.30	1.62

[a] Incomplete and ungrammatical sentences do not contribute to these means.

Discussion

The experiment reported supports two major conclusions: first, that structural complexity as indexed by the sentence mean-depth measure is a definitive factor in sentence retention; second, that when sentence complexity and sentence length are controlled, the role of sentence kind in explaining recall behavior becomes marginal.

The data indicate that irrespective of sentence kind, likelihood of correct recall follows inversely sentence mean depth. If some aspect of memory fails and recall is attempted, the resulting error is very likely to be a sentence of the same kind as the presented sentence but with a lesser mean depth. These findings are in accordance with the analysis of listening and reproducing given at the outset; they are not, however, explicable by the transformation-grammar model.

Consider first the conclusion that sentence kind is of minimal interest when sentence mean depth and sentence length are controlled. In applying the transformation-grammar model to the human information processor, one is required to argue that a nonkernel is stored in memory as the corresponding kernel plus a transformation tag. This means that retention of kernels should be superior to retention of nonkernels: representation in memory of a nonkernel is susceptible both to loss of the kernel component and to loss of the transformation tag that must accompany the kernel component. Thus sentence kind is given a central role with kernels clearly seen as easiest to remember. The present data indicate that recall of kernels is uniformly inferior to recall of nonkernels, and that except for this disconfirming regularity there is no consistent effect on recall attributable to the remaining sentence kinds.

It is not inappropriate to question the foregoing conclusion on the grounds that because sentence kind was manipulated within *S*s, and because there are, for example, four sentences out of the six that are of the passive form, there may exist an induced response bias toward the passive form and therefore away from the kernel. If so, then (a) recall performance should be superior for passives relative to nonpassives, and (b) grammatical complete-sentence errors for nonpassive presented sentences should tend toward the passive form. Regarding the first possibility, from Table 1 it is apparent that the passives tend to be recalled slightly more frequently than the nonpassives. By *t* test, the total number of recalls for each of the 48 passive sentences were compared with the total number recalls for each of the 24 nonpassive sentences. The result was $t(70) = .56$, $p > .50$. With respect to the frequency of passive errors to nonpassive presented sentences, from Table 2 one can calculate that of the 127 complete-sentence transformation errors given to the two nonpassive sentence kinds (K and N), 79 were of one of the four passive kinds, or, on the average, 19.8 per kind, and 48 were of the remaining nonpassive kind. Thus neither the correct-recall data nor the error data indicate a response bias against nonpassive kinds.

Consider now the conclusion that sentence mean depth is a significant factor in sentence retention. Current applications of the transformation-grammar model to sentence-recall behavior (Mehler, 1963; Miller, 1962) do not provide for prediction of differential recallability of two sentences of the same kind. The present data indicate that such differences clearly exist and can, to some extent, be quantified on a structural-complexity basis: except for the passive-negatives, those sentences of a given kind with mean depth 1.29 were recalled with greater frequency than those of the same kind with mean depth 1.71. It might be argued in return that when two simple sentences of the same kind differ in mean depth, appropriate grammatical analysis will reveal that the more complex of the two involves additional grammatical rules that deal with aspects of the sentence other than simply sentence kind. But such additional grammatical rules cannot be transformation rules since transformation rules are the basis of sentence kind. Therefore these additional rules must be the rules of phrase-structure grammar, and in the introduction it was submitted that the generation of sentences and the "understanding" and retention of sentences by human language users corresponds to a phrase-structure grammar of the language: the number of left branches of a phrase-marker tree leading to a given word defines the depth (structural embeddedness) of that word, and hence the number of grammatical commitments incurred by the speaker and the number of expectations elicited in the listener. Thus the transformation-grammar model cannot predict differences in retention of two sentences of the same kind without resorting to phrase-structure grammar, which is tantamount to admitting the basic rationale of Yngve's original idea.

Further evidence on these matters comes from a study by the present writers in which *S*s made judgments as to which of two seven-word sentences they thought would be easier to remember if they had to relay one of them to someone else after a delay of several minutes. Since the sentence pairs were in writing in front of them, and since their judgments were not paced, *S*s need not have stored any information in memory; only their intuitions regarding ordinary English sentences presumably were in effect. Twelve Ks were paired with 12 Ns and with 12 Ps, thus giving 144 K-N and 144 K-P sentence pairs. Of the 12 sentences of each kind, there were two each of mean depths 1.00, 1.14, 1.29, 1.43, 1.57, and 1.71. Thus the difference in mean depths for pairs of sentences ranged from —.71 to .71; a difference of —.71 obtained when, for example, the K member of a K-N pair had mean depth 1.00 and the N member had mean depth 1.71. Which kind of sentence occurred first in a pair was balanced over the magnitudes of difference in mean depth. The judgments were made immediately following the free-learning experiment, and by the same *S*s. The sentences of the two experiments were completely unrelated semantically.

With respect to the resulting judgments of retainability as a function of sentence kind, Ns were chosen over Ks 54% of the time and Ps were chosen over Ks 50% of the time. As for judgments as a function of difference in sentence mean depths, *S*s consistently chose the sentence of lesser mean depth as easier to remember: for K-N pairs, Kendall's *tau* between proportion of choices of K over N and the —.71 to .71 range of mean-depth differences is .54 ($p = .01$); for K-P pairs, *tau* is .85 ($p < .001$). In a similar study comparing Ps and $\mathrm{P_T}$s, Ps were judged easier to remember than $\mathrm{P_T}$s 51% of the time; the corresponding *tau* is .93 ($p < .001$).

Thus, in a situation where *S*s are asked to make judgments regarding storage and subsequent reproduction, that is, to exercise their knowledge of the language without an actual test of ability, one again finds that it is structural complexity and not sentence kind *per se* that accounts for behavior.

The final bit of evidence to be adduced in support of the position taken here is provided by Mehler (1963). His is a free-learning study very similar to the one reported here. If the Yngve numbers for the examples he gives of the K, P, N, and PN sentences used in his experiment are determined and sentence mean depths figured, the relationship that emerges between total number correct recalls and sentence mean depth is a remarkably strong one: the mean depth values are 1.17, 1.38, 1.43, and 1.67 and the total number correct recalls are 300, 243, 234, and 191, respectively, for his K, P, N, and PN sentences. Although Mehler gives a transformation-grammar interpretation of these results, it can be seen that a phrase-structure (sentence-complexity) analysis induces a near-perfect linear orderliness.

References

Bach, E. *An introduction to transformational grammars.* New York: Holt, Rinehart & Winston, 1964.

Chomsky, N. *Syntactic structures.* 's-Gravenhage: Mouton, 1957.

Chomsky, N. *Aspects of the theory of syntax.* Cambridge: M. I. T. Press, 1965.

Mehler, J. Some effects of grammatical transformations on the recall of English sentences. *J. verb. Learn. verb. Behav.*, 1963, **2**, 346-351.

Miller, G. A. Some psychological studies of grammar. *Amer. Psychologist*, 1962, **17**, 748-762.

Miller, G. A., and McKean, K. O. A chronometric study of some relations between sentences. *Quart. J. exp. Psychol.*, 1964, **16**, 297-308.

Thorndike, E. L., and Lorge, I. *The teacher's word* *book of 30,000 words*. New York: Bur. Publ., Teacher's Coll., Columbia Univ., 1944.

YNGVE, V. H. A model and an hypothesis for language structure. *Proc. Amer. Phil. Soc.*, 1960, **104**, 444-466.

YNGVE, V. H. Implications of mechanical translation research. *Proc. Amer. Phil. Soc.*, 1964, **108**, 275-281.

(Received January 4, 1965)