Memory & Cognition
1981, Vol. 9 (5), 478-486

Performance structures in the recall of sentences

JEAN-YVES DOMMERGUES
University of Paris VIII, 93200 St. Denis, France

and

FRANCOIS GROSJEAN
Northeastern University, Boston, Massachusetts 02115

Two experiments were conducted to ascertain whether subjects' recall of sentences reflects more closely their surface structure, as Johnson (1965, 1969) and others have predicted, or their performance structure, as Grosjean, Grosjean, and Lane (1979) have proposed. The results clearly show that, as for pausing and parsing, transitional error probability (TEP) in the rote recall of sentences reflects the product of two, sometimes conflicting, demands on processing: the need to respect the linguistic surface structure of the sentence and the need to balance the length of the constituents. The structures obtained from TEPs were similar to those obtained from other tasks (pausing and parsing), showing that performance structures are not task specific. In addition, the presence of deletable elements in the sentences (such as adjectives and adverbs) was closely associated with high TEPs.

A number of psycholinguistic studies in the 1960s tried to assess the psychological reality and validity of linguistic structures and rules that are proposed by transformational-generative grammarians. For instance, Miller and Isard (1963) showed the effect of syntactic rules in the perception and recall of sentences. Fodor and Bever (1965) found that clicks objectively superimposed on sentences near a major boundary were subjectively located at clause boundaries; this result was taken to confirm the psychological reality of clauses and major constituents. And Johnson (1965) showed that subjects use their knowledge of grammar to break a sentence into functional subunits as they attempt to learn it. In summary, these experiments in recall and perception were aimed at demonstrating that structural descriptions of sentences are psychologically valid.

A few years later, however, E. Martin (1970) found that when subjects were asked to parse sentences into "natural groups," they did not automatically group the verb with the noun phrase (NP) object, as linguistic models would predict; in many cases, they grouped the verb with the NP subject and put the NP object into a separate group. Along the same lines, Levelt (1970) reported that the minor constituents of a sentence were not systematically reflected in the hierarchical structure obtained from errors made in a noise perception study.

This research was supported in part by a Fulbright-Hays grant to the first author and by Grant 768 252 from the National Science Foundation and Grants RR 07143 and 1 Ro1 NS 14923-01A2 from the Department of Health, Education, and Welfare to the second author. The authors would like to thank Harlan Lane for his constructive comments throughout the study. Requests for reprints should be sent to Jean-Yves Dommergues, Département d'anglais, Université de Paris VIII, Rue de la Liberté, 93200 St. Denis, France.

Researchers have recently studied this lack of correspondence between linguistic surface structures and the performance structures obtained from experimental data (Grosjean, Grosjean, & Lane, 1979; E. Martin, 1970). For instance, Grosjean et al. (1979) found that pause durations yielded reliable performance structures. Grosjean et al. characterized performance pause structures as the product of two (sometimes conflicting) demands on the speaker: the need to respect the linguistic structure of the sentence and the need to balance the length of the constituents in the output. A simple cyclical model, combining, for each pause location, an index of linguistic complexity (based on the surface structure of the sentence) and a measure of the distance to the midpoint of the segment, accounted for 72% of the pause-time variance, as opposed to 56% for the linguistic index alone. The generality of the model was shown by its good prediction of a number of dependent variables, such as pause durations, indexes of relatedness, and parsing, in unrelated studies in American Sign Language and English (Grosjean et al., 1979; Grosjean, Lane, Battison, & Teuber, 1981).

In order to test further the generality of performance structures, several other measures (apart from the ones used by Grosjean et al., 1979) are available: for example, probe latency, click location, and transitional error probability (TEP), used by Johnson (1965, 1969, for instance). We decided to ascertain whether the last measure (TEP) could produce performance structures similar to those obtained from pausing and parsing or whether the TEP structures best reflect linguistic surface structures.

Johnson (1965) tested the hypothesis that there are particular word-to-word transitions within sentences in which the probability of a transitional error is sig-

0090-502X/81/050478-09$ 01.15/0

nificantly greater than it is for other transitions. His hypothesis was that these points should occur at both major and minor constituent breaks. To verify this, he used sentences of the following type: (1) "The tall boy saved the dying woman." (2) "The house across the street is burning." The subjects' task in the recall experiment was to learn an eight-item paired associate list in which the digits from 1 to 8 were the stimuli and the eight sentences were the responses. The sentences were presented to subjects on a memory drum at a 4.4-sec rate, with a 4-sec intertrial interval. The instructions emphasized that the subjects should report as much of each sentence as they could remember. The probability of a transitional error within a sequence was computed by counting the frequency with which a particular word was wrong, given that the preceding word was correct, and dividing that frequency by the total frequency that the preceding word was correct (i.e., dividing the transitional error frequency by the number of opportunities for an error). A TEP was then attributed to all word boundaries in each sentence, as is illustrated in the following example:

(1a) The tall boy saved the dying woman.
.11 .05 .12 .07 .03 .02

For this type of sentence, the Kendall tau correlation between the level at which a linguistic division occurs (a measure obtained with hierarchical node values) and TEP was .64. Thus, Johnson (1965) concluded that, as the conditional probabilities were predictable from the linguistic surface structure of the sentences, subjects actively use linguistic structure in their learning and recall of sentences and, hence, that linguistic structures have psychological reality.

In view of the results obtained by Grosjean et al. (1979), we hypothesize that the way subjects chunk sentences in recall corresponds more to the way performance structures are built and produced than to the way linguistic structures are constructed and described. In the first experiment, therefore, we take each of the 14 sentences used by Grosjean et al. and obtain TEPs for each word boundary. We then correlate these results, first, with two sets of experimental data obtained by Grosjean et al. (pause durations and parsing values) and, second, with two predictors: syntactic complexity indexes (CIs), based on the surface structure of the sentences, and theoretical performance structure indexes (PIs), obtained from the model proposed by Grosjean et al. (1979) to account for their pausing and parsing data. Based on previous studies that attempted to use linguistic models to predict performance data (Grosjean et al., 1979, 1981; E. Martin, 1970), we expect higher coefficients of correlation between TEP and pausing, parsing, and the performance structure predictor than between TEP and the syntactic complexity index. In a word, TEPs should provide better reflections of the performance structures than of the linguistic surface structures of sentences.

EXPERIMENT 1

Method

Subjects. Thirteen undergraduate students served individually in an experiment that lasted 45 min.

Materials. The 14 experimental sentences were taken from Grosjean et al. (1979), who adapted them from Bever, Lackner, and Kirk (1969). They varied in length from 11 to 13 words; five were simple sentences and nine were complex, containing subordinate clauses or embedded relative clauses. The sentences were printed randomly on cards in sets of four.

Procedure. The subjects were given 20 sec to read and learn four written sentences, each with an associated digit. They were then given a digit at random and had to recall the corresponding sentence. Subjects were given seven trials on each set. Sentences and sets were randomized across subjects, and all responses were tape-recorded (on a Tandberg 1600X).

Data analysis. The following errors were used to calculate TEPs: omissions, additions, and substitutions. TEPs were calculated for each sentence and each subject in the manner described by Johnson (1965): The TEP at a word boundary was the number of times a particular word was wrong, given that the preceding word was correct, divided by the number of times the preceding word was correct. The probabilities for each sentence were pooled across subjects and were then correlated with the corresponding pause durations, parsing indexes, CIs, and PIs reported by Grosjean et al. (1979) for the same 14 sentences. The pause durations were obtained by Grosjean et al., who asked subjects to read the sentences at five different rates. The pauses found at each word boundary were pooled, and the mean duration was computed and expressed as a percentage of the total pause duration in that sentence. The parsing values were also obtained by Grosjean et al. (1979) from "linguistically naive" subjects who were asked to parse the same 14 sentences according to the following procedure: "Find the main break in the sentence and put a slash with a number 1 on top; then consider the two parts of the sentence independently and divide them up in turn with slash 2, and continue dividing up each part until every word boundary has a slash and number indicating its importance" (Grosjean et al., 1979, p. 65). Parsing indexes were pooled across subjects, and means were computed for each word boundary, following the procedure detailed in Grosjean et al. (1979).

CIs based on the surface structure of the sentences were obtained for each word boundary by counting the number of nodes dominated by the word boundary node, including in the count the word boundary node itself. Thus, in Figure 1 (top structure), the CI between "Reynolds" and "the" is 9 because the boundary node dominates five nodes on the left and three nodes on the right and is itself counted. The CI between "lawyer" and "called" is 4, as the boundary node dominates one node on the left and two nodes on the right and is itself counted, and so on.

Finally, Grosjean et al. (1979) obtained PIs from the model they developed. The model assigns to each word boundary a number that is the product of the height of the boundary in the surface structure tree and the proximity of that boundary to the bisection point in a symmetrical tree. Below, we outline the performance model and the steps followed to obtain the PIs for the sentence in Figure 1 (bottom structure):

> Step 1. Starting with the largest constituent that has not been analyzed (at the beginning, the whole sentence), compute a CI for every word boundary, based on the surface structure tree of the constituent.
>
> *When* 5 *the* 1 *new* 1 *lawyer* 4 *called* 1 *up* 2 *Reynolds* 9 *the* 1 *plan* 2 *was* 2 *discussed* 1 *thoroughly.*
>
> Step 2. Continuing with this same constituent (at first the whole sentence), compute for each word boundary a relative proximity index of that boundary to the bisection point: the number of words from the start (or end) of the

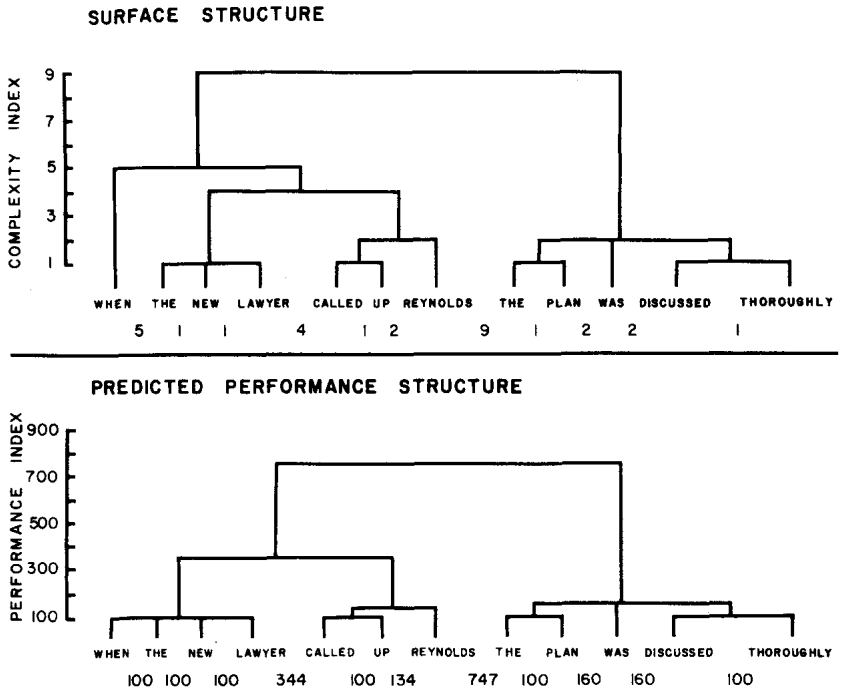

Figure 1. A sentence from the study by Grosjean, Grosjean, and Lane (1979). The top structure represents the surface structure of the sentence and its complexity indexes (CIs). The bottom structure represents the predicted performance structure and its performance indexes (PIs). The CIs and PIs are used to give height to the nodes of the respective structure along a ratio scale.

constituent to the boundary (whichever is less) divided by half the number of words in the constituent, expressed as a %: %RP.

When 17% *the* 33% *new* 50% *lawyer* 67% *called* 83% *up* 100% *Reynolds* 83% *the* 67% *plan* 50% *was* 33% *discussed* 17% *thoroughly.*

Step 3. Multiply the values assigned to each word boundary: The boundary with the largest product (747) is the constituent break and retains its product. No other product is retained.

When 85 *the* 33 *new* 50 *lawyer* 268 *called* 83 *up* 200 *Reynolds* 747 *the* 67 *plan* 100 *was* 66 *discussed* 17 *thoroughly.*

Step 4. Take each of the two constituents just created ("When the new lawyer called up Reynolds" and "the plan was discussed thoroughly"), build a new surface structure for each, calculate CIs and %RP for each word boundary, multiply these values, find the largest product, and ignore all others (i.e., repeat Steps 1-3 for each constituent). Thus, for the first constituent,

	When	*the*	*new*	*lawyer*	*called*	*up*	*Reynolds*
CI:	5	1	1	4	1	2	
%RP:	29	57	86	86	57	29	
Product:	145	57	86	*344*	57	58	

The operation produces two smaller constituents on either side of the largest product (344), "when the new lawyer" and "called up Reynolds," which in turn pass through the process: CIs and %RPs are calculated for each, their products are obtained, and so on, iteratively. Thus, for the first constituent, we have the following new values:

	When	*the*	*new*	*lawyer*
CI:	2	1	1	
%RP:	50	100	50	
Product:	*100*	*100*	50	

In this case, we have a tie for the largest product, and so both values of 100 are retained. The constituent "new lawyer" now passes through the process, to give

	new	*lawyer*
CI:	1	
%RP:	100	
Product:	*100*	

This procedure is applied to all other constituents (which increase in number but decrease in size after each cycle) until all word boundaries obtain a "largest product." The final values for this sentence are:

When 100 *the* 100 *new* 100 *lawyer* 344 *called* 100 *up* 134 *Reynolds* 747 *the* 100 *plan* 160 *was* 160 *discussed* 100 *thoroughly.*

Thus, for the same set of 14 sentences and for each word boundary within each sentence, there was a TEP obtained in this study and a pause duration, a parsing value, a CI, and a PI obtained by Grosjean et al. (1979). With this data, we computed Pearson product-moment correlations between the TEPs pooled across all 14 sentences (N = 154) and the corresponding pause durations, parsing values, CIs, and PIs.[1]

Results and Discussion

The coefficients of correlation obtained between TEP and the other experimental data (pausing and parsing) and between TEP and predictor values (CIs and PIs) are presented in Table 1 (Column 1). Examining first the correlations between TEP and the experimental data, we note that these are not very high: .42 between TEP and parsing, and .41 between TEP and pausing. This is in sharp contrast with the .92 correlation obtained between pausing and parsing by Grosjean et al. (1979). Next, we note that our measure of linguistic complexity

Table 1
Pearson Product-Moment Correlations Between TEPs and Other Experimental Data (Parsing and Pausing Values) and Between TEPs and Predictor Values (the Linguistic Complexity Indexes, CIs, and the Performance Structure Indexes, PIs)

	All Breaks in All Sentences (N = 154)	All Breaks in the Four Sentences With One or No Deletables (N = 44)
	TEP and Experimental Data	
TEP-Parsing	.42	.70
TEP-Pausing	.41	.64
	TEP and Predictors	
TEP-CI	.39	.56
TEP-PI	.44	.79

Note–For $N = 154$, $p < .01$ when $r \geq .25$; for $N = 44$, $p < .01$ when $r \geq .37$.

of the sentence (CI) is a rather poor predictor of the TEP data (r = .39). This is again in contrast with the .75 correlation reported by Grosjean et al. (1979) between CI and pausing, and especially the .64 correlation reported by Johnson (1965) between TEP and his measure of linguistic complexity. And finally, we note that the performance model proposed by Grosjean et al. is not a very good predictor of the TEPs: r = .44, as compared with r = .85 for pausing in the Grosjean et al. study.

Two factors may explain these findings. The first is that many of our sentences were not balanced, in that the NP subject did not always contain the same number of words as the verb phrase (VP). In Sentence 3, for instance, the surface NP is 1 word long and the VP is 11 words long: (3) "John asked the strange young man to be quick on the task." This may well affect the recall of sentences. Indirect evidence for this comes from the learning of strings of letters. Marmurek and Johnson (1978), for instance, asked subjects to learn sets of permutations of a base sequence of letters of the type ABCDEFGH. When a set of permutations defined a balanced hierarchical organization for the base sequence, recall was better than when the organization was unbalanced. Although this result was obtained with letter strings, we believe the balance factor applies equally well to verbal material. It is interesting to note that in his earlier studies using sentences, Johnson (1965, 1969) used fairly balanced structures. For example, in the sentence (1) "The tall boy saved the dying woman," the NP is three words long and the VP is four words long. This could explain in part Johnson's rather high correlation (.64) between TEP and his measure of structural complexity.

A second factor concerns the number of deletable elements that are contained in the sentence. These elements are adjectives, adverbs, and prepositional phrases, such as "strange," "young," and "on the task" in (3) "John asked the strange young man to be quick on the task." Johnson (1969) noted that the TEPs obtained with sentences containing such elements (he mentions only adjectives and adverbs) were not predicted as well as those obtained with sentences devoid of them. As it happens, most of our sentences contained deletables. Omitting them in recall does not make the sentence ungrammatical or anomalous, but it does increase the TEPs considerably. For instance, in the sentence (4) "Our disappointed woman lost her optimism since the prospects were too limited," the TEP between "our" and "disappointed" tended to be quite high, because subjects often left out the adjective in the recall of the NP.

We thus returned to our 14 sentences and isolated 4 that contained only one or no deletables and recomputed a new set of correlations on the reduced set of data (N = 44). We reasoned that if deletables were a factor in the low correlations between TEPs and the experimental data (pausing and parsing) and between TEPs and the predictors (CIs and PIs), then all correlations based on this subset would be higher, since sentences with more than one deletable had been excluded. In addition, having controlled for deletables, we now expected a difference between the TEP-CI and TEP-PI correlations, as one predictor (PI) takes into account the need of subjects to store and output constituents of equal length, whereas the other (CI) does not.

As can be seen in the second column of Table 1, our expectations were confirmed. First, the correlations between TEP and the experimental data increased substantially: r = .70, as opposed to r = .42, for TEP and parsing; r = .64, as compared with r = .41, for TEP and pausing. Second, the CI is now a better predictor of the TEP data: r = .56, a coefficient that is closer to Johnson's (1965) correlation of .64. And third, the Grosjean et al. (1979) model of performance structures proves to be a better predictor of the TEP than the CI: r = .79, as compared with r = .56 for the latter predictor. This difference is significant at the .001 level (t = 3.15; test for the equality of two correlation coefficients for related samples, Weinberg & Goldberg, 1979). This corroborates results published by Grosjean et al. (1979, 1981), who found that performance data obtained from such diverse tasks as reading at slow rates, parsing, making relatedness judgments, and recalling sentences in speech and sign language are better correlated with a model that takes into account the surface structure of the sentence and the need to store and output constituents of equal length than with the sole surface structure of the sentence.

In this first experiment, therefore, we have obtained additional evidence that TEPs are influenced by the presence or absence of deletable elements in the sentence (when adjectives, adverbs, and prepositional phrases are dropped in recall, higher TEPs are obtained at unimportant breaks), by the syntactic structure of

the sentence (thus confirming the results obtained by Johnson, 1965), and by the length of the main constituents in the sentence (if the NP and VP constituents are of unequal length, for example, the main TEP may not be found at the constituent break but within a constituent). In Experiment 2, we will attempt to confirm these results.

EXPERIMENT 2

A first aim of this experiment is to demonstrate experimentally the influence of deletable elements and of unbalanced sentences on TEPs. To do this, we will use especially designed balanced and unbalanced sentences that contain few or many deletables. Based on Experiment 1, we first expect that the two dependent variables under study, TEP and parsing, will be highly correlated with each other in both balanced and unbalanced sentences when these contain few deletables; this should not be true, however, with sentences containing many deletables, as TEPs are influenced by such elements, whereas parsing values are not. Second, we expect higher correlations between TEP and CI when sentences are balanced than when they are unbalanced and better correlations for sentences that have few deletables. And third, we expect that the performance model proposed by Grosjean et al. (1979) (which takes into account the surface structure of the sentence and the length of each constituent) is overall a better predictor of TEPs than is the surface structure of the sentence (CI) by itself: It is as good a predictor for balanced sentences, but a much better predictor of TEPs in unbalanced sentences. Neither the CI nor the performance model, however, is expected to be a very good predictor of the TEPs obtained from sentences containing many deletable elements.

Method

Subjects. Forty undergraduate students with no reported speech or hearing defects served individually in the recall experiment, which lasted 30 min. Twenty other undergraduates parsed the experimental sentences in a group session.

Materials. Four types of sentences were used in the experiment; each type was represented by two exemplars, which made a total of eight experimental sentences. Type 1 sentences were balanced, in that the numbers of words in the NP and VP were identical. They were: (5) "Many children in Sweden put on warm clothing," and (6) "Most tourists from England look up elderly relatives." This type of sentence contained two deletable elements ("many" and "warm" in Sentence 5), but the first was not judged to be critical, as its omission would not affect the computation of the TEP between the first and second words of the sentence.

Type 2 sentences were unbalanced, in that the NP was one word long and the VP was seven words long. They were: (7) "She questioned Mary after the really noisy concert," and (8) "We visited Rome between the most tiring flights." These sentences contained two adjacent deletables in the prepositional phrase ("really" and "noisy" in Sentence 7), but only one was judged as being critical for the computation of TEPs ("really").

Type 3 sentences were balanced and contained three critical deletables, for example: (9) "The cold winter there alarmed the poor farmers," and (10) "The small gadget here attracted the young shopkeepers." The critical deletables in Sentence 9 are "cold," "there," and "poor."

Type 4 sentences were unbalanced and contained two critical deletables, for example: (11) "Robert asked him to quickly send new orders," and (12) "Jill told them to finally forget past quarrels." In Sentence 11, the critical deletables are "quickly" and "new."

For practical purposes, we will refer to sentences with only one critical deletable as "sentences with few deletables" (Types 1 and 2) and refer to sentences with two or more deletables as "sentences with many deletables" (Types 3 and 4).

Procedure. In the recall experiment, the subjects learned and recalled the sentences as in Experiment 1. Sentences were presented in blocks of four, each sentence being of a different type. The order of presentation was random, as before; subjects were given seven trials. The answers were tape-recorded.

In the parsing experiment, the Grosjean et al. (1979) procedure used in Experiment 1 was again employed, but this time subjects were asked to rank the importance of the break using a 1-5 continuous scale, with 5 as an important break and 1 an unimportant break. This was done to give subjects more latitude in deciding on the importance of the syntactic break.

Data analysis. TEPs were calculated for each sentence and each subject in the manner they were calculated in Experiment 1. The TEPs were averaged across subjects and were then used to construct hierarchical sentence structures for each sentence according to the following iterative procedure, originally proposed by Grosjean and Lane (1977): "First, find the shortest value in the sentence. Second, cluster the two elements separated by that value by linking them to a common node and delete the value. (If three or more adjacent words are separated from each other by the same value, make one cluster of these words: trinary, quaternary, etc.) Finally, repeat the process until all values have been deleted."

Parsing values and grammatical CIs were obtained as in Experiment 1. Parsing values were then used to construct hierarchical structures following the iterative procedure used for TEPs. Finally, predicted performance structures were obtained for each sentence type from the model proposed by Grosjean et al. (1979).

Thus, for each of the four sentence structure types, we have four sets of 14 mean values (each sentence type is represented by two exemplars and each exemplar contains seven word boundaries). Two sets are predictor values: the CIs based on the surface structure trees of the sentences elaborated by linguists on the basis of a "classical" model of sentence descriptions and the predicted performance structure model; and two sets are actual performance data: TEPs and parsing values.

For each sentence type, the 14 TEP values were correlated with the corresponding parsing values and with the predictor values: CIs and PIs.

Results and Discussion

The invariance of performance structures. Grosjean et al. (1979) reported a high correlation between the pause durations produced in slow reading and the parsing values given by subjects for identical sentences ($r = .92$) and concluded that performance structures are quite invariant across experimental tasks. Grosjean et al. (1981) confirmed this finding in a different language modality, American Sign Language, with four different paradigms: signing at slow rate, parsing, relatedness judgments of pairs of signs taken from each sentence, and probed recall; the mean coefficient of correlation across all four tasks was .73. As can be seen in Table 2, the correlation between TEPs and parsing values is also very high: .87 for Type 1 sentences (balanced with few

Table 2
Pearson Product-Moment Correlations Between TEPs and Parsing and TEPs and Predictor Values (the Linguistic Complexity Indexes, CIs, and the Performance Structure Indexes, PIs)

	Type 1	Type 2	Type 3	Type 4
	TEP and Experimental Data			
TEP-Parsing	.87	.83	.36	.05
	TEP and Predictors			
TEP-CI	.89	.42	−.12	−.12
TEP-PI	.89	.82	−.12	−.26

Note–Type 1 = balanced sentences, few deletables; Type 2 = unbalanced sentences, few deletables; Type 3 = balanced sentences, many deletables; Type 4 = unbalanced sentences, many deletables. Each correlation is based on 14 data points (for $N = 14$, $p < .01$ when $r \geq .66$).

deletables) and .83 for Type 2 sentences (unbalanced with few deletables). However, and as expected, for sentences with many deletable elements, the correlations between the two measures are low: r = .36 for Type 3 sentences and r = .05 for Type 4 sentences.

The similarity between the performance structures obtained from recall and parsing is also illustrated in Figures 2 and 3 (bottom structures), in which we present the performance structures obtained for Type 1 and 2 sentences. We can see that very different tasks produce very similar performance structures, and we can conclude that the two paradigms are probably tapping the same subjective sentence organization imposed by the speaker-listener. This confirms the earlier findings of Grosjean and his colleagues.

The influence of deletable elements and of unbalanced sentences on TEPs. The results obtained in Experiment 1 led us to expect high correlations between TEP and CI when sentences were balanced and when they contained few deletables, but lower correlations when sentences were unbalanced and when they contained many deletable items. As can be seen in Table 2 (second row) these expectations were borne out. When sentences had constituents of unequal length (for example, Type 2 sentences), the correlation between the TEPs and CIs was much lower (r = .42) than when the constituents were of the same length (r = .89) for Type 1 sentences. This significant difference ($z = 2.3$, $p < .025$; Weinberg & Goldberg, 1979) is well illustrated in Figures 2 and 3, in which we present the surface structure and TEP structure of Type 1 and 2 sentences. In Figure 2, we find a very good match between the performance structure based on recall and the surface structure of the sentence, whereas in the second figure, in which the sentence has an NP of one word and a VP of seven words, the TEP structure is very different from the surface structure. Here, the main TEP break has been moved from the NP-VP boundary to the NP/prepositional phrase break within the VP. These results confirm those obtained in

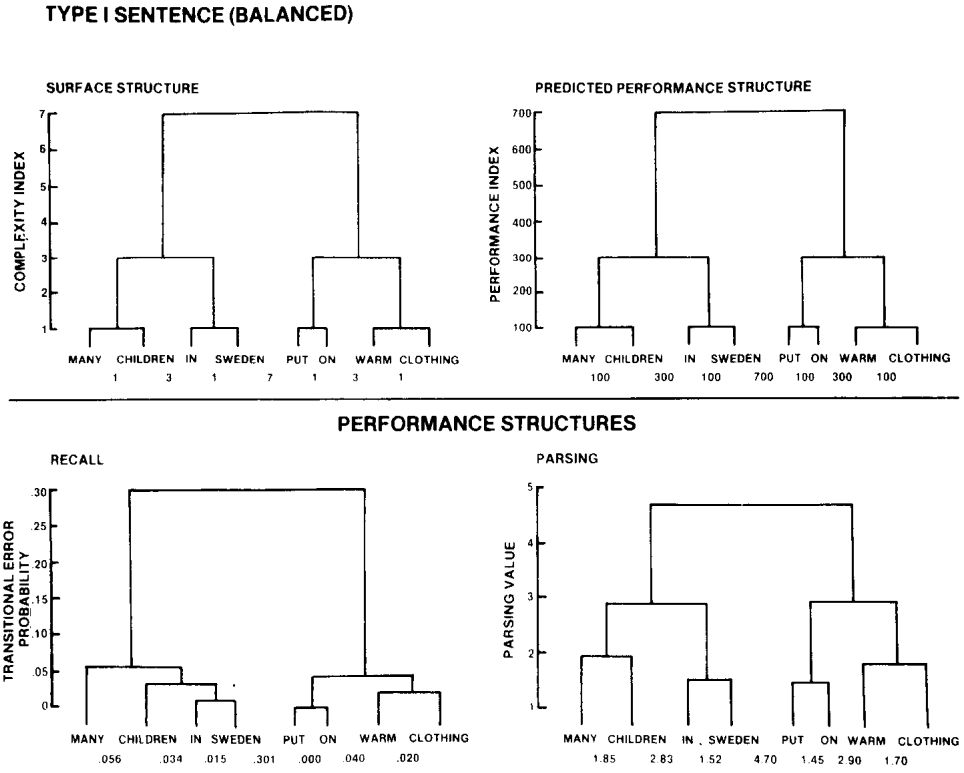

Figure 2. The surface structure of a Type 1 sentence (balanced), its predicted performance structure, and the performance structures obtained from rote recall (transitional error probabilities) and parsing. The values obtained from each paradigm and from the two models are used to give height to the nodes of the respective structures along a ratio scale.

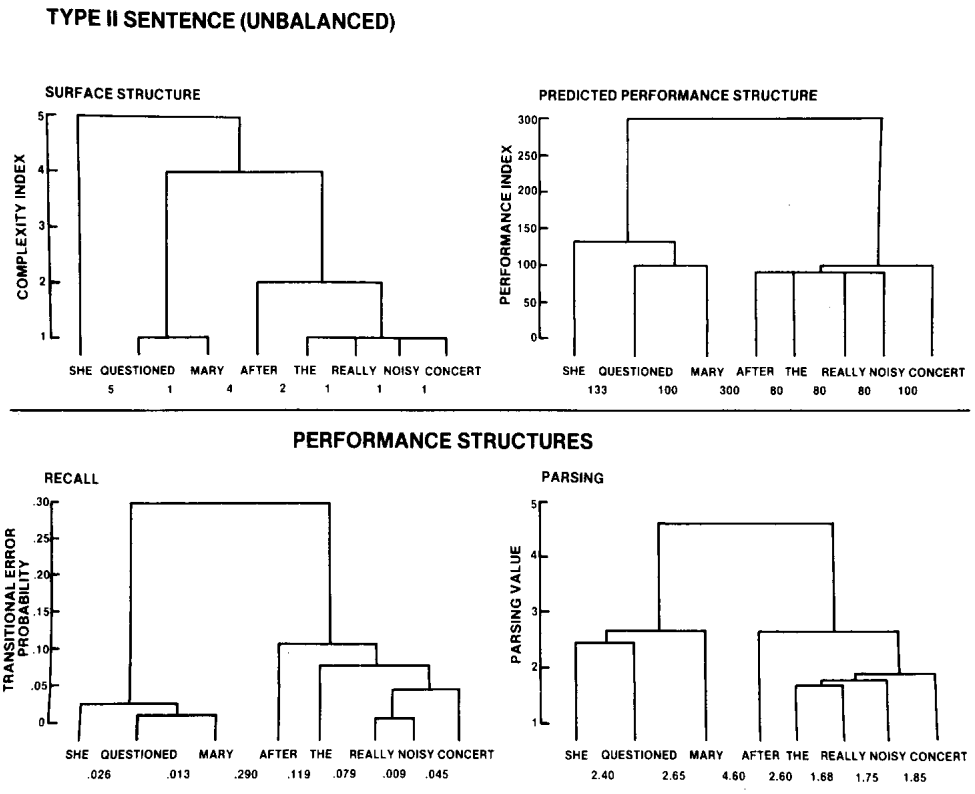

Figure 3. The surface structure of a Type 2 sentence (unbalanced), its predicted performance structure, and the performance structures obtained from rote recall (transitional error probabilities) and parsing. The values obtained from each paradigm and from the two models are used to give height to the nodes of the respective structures along a ratio scale.

Experiment 1 and explain in part the fairly good correlation obtained by Johnson (1965) between his TEPs and the linguistic structure of his sentences (r = .64): Johnson used fairly balanced sentences.

We also expected the results to confirm the role played by deletable elements (adjectives, adverbs, and prepositional phrases) in the recall of sentences and, thereby, on the computation of TEPs. As can be seen in Table 2, the correlations between TEP and CI and between TEP and parsing values for sentences with many deletables (Types 3 and 4) are close to zero. This is explained by the fact that when deletable elements are left out by subjects in their recall, the TEPs increase considerably, even though there is no major break at that particular location. Neither parsing values nor surface structure indexes can then predict the TEPs produced by subjects.

The prediction of TEPs. In Experiment 1, we found that TEPs were better correlated with other performance data (pausing and parsing) and with the Grosjean et al. (1979) PIs than with the CIs of the sentence, and we concluded tentatively that TEPs would probably be better predicted by a performance model that takes into account not only the syntactic structure of the sentence, but also the length of the constituents than by the sole surface structure model. As can be seen in Table 2, the performance model in question is indeed a better overall predictor of TEP for sentences containing few deletables. Both the CIs and the PIs are good predictors of the TEPs of balanced sentences (they both account for 79% of the TEP variance), but the performance model is a better predictor of the TEPs in unbalanced sentences. The PIs account for 67% of the TEP variance in these sentences, whereas the CIs account for only 18% of the variance (the TEP-CI and TEP-PI correlations for Type 2 sentences, .42 and .82, were significantly different at the .001 level; Weinberg & Goldberg, 1979).

The prediction strength of both the surface structure and the performance model is illustrated in Figures 2 and 3. In Figure 2, the sentence is balanced and both models are good predictors of the TEP data (as well as of the parsing values). For example, the main break in both predictor structures is after "Sweden" and the highest TEP and parsing values are also at that break. However, in Figure 3, in which we have a right-branching sentence with a short NP (one word) and a long VP (seven words), the surface structure model predicts a main break after the NP ("She"), whereas the performance model predicts it after the object NP ("Mary"). This is precisely the place at which we find the highest TEP and the largest parsing value, thereby confirming the strength of the performance structure model. (We should note here that neither model predicts the performance structure of sentences containing many deletable elements; as can be seen in Table 2, all correlations between either CIs or PIs and TEPs obtained for Type 3

and 4 sentences are close to zero and nonsignificant.)

We can conclude from this that TEPs (in sentences without deletables) can best be predicted by a model that takes into account not only the surface structure of the sentence, but also the speaker-listener's need to output constituents of equal length. Grosjean et al. (1981) explain the superiority of the performance model over the sentence structure model by the fact that the speaker-listener may initiate any sentence processing task (reading at slow rate, parsing, rote recall, etc.) with a baseline structural expectation that probably corresponds to the most economical hierarchical code, that is, an unmarked, binary tree. Any departure from this symmetrical structure would require an increase in the structural information stored. The hierarchical structure in any particular sentence, then, would entail substantial, little, or no change in the baseline expectation, the unmarked tree.

If the surface structure tree (or phrase marker) is close to the unmarked tree (which is the case of most published data), it is tempting to conclude that processing proceeds in terms of the surface structure tree: Balance does not need to be accounted for. When it comes to the unbalanced sentences, however, the unmarked tree and the phrase marker assign different hierarchical structures to the sentence. In this case, sentence processing seems to strike a compromise: Subjects weight the phrase marker by the unmarked tree in performing the task at hand. It should be noted that such an operation does not necessarily contradict Johnson's (1969) decoding operation model or Frazier and Fodor's (1978) "sausage machine" model. These two models put the stress on the derivation of the phrase marker, whereas the present model does not address this question. In fact, it takes the assignment of a phrase marker by the speaker-listener for granted and attempts to explain only how such a structure is combined with a balanced tree to produce a performance structure.

In short, the performance model is a better predictor in such tasks as parsing and rote recall than is the surface structure tree alone because the performance model weights the latter by an unmarked tree that reflects the enduring frame of reference. Martin (1972) thinks that unmarked trees of this kind are also the point of departure for assigning rhythmic structure to a sentence in speaking and comprehending.

CONCLUSION

We have shown in this study that TEPs obtained from a recall task are influenced by a number of factors that can have a powerful effect on the correlation between TEP and the sentence structure. The first is the presence of deletable elements in the sentence: If such items as adjectives, adverbs, and prepositional phrases are left out in the rote recall of the sentence, then the TEPs increase considerably, even though there is no major break at those particular locations. The second factor is the length of constituents: When constituents are of unequal lengths, TEPs will no longer reflect the surface structure of the sentence, and the main surface break in the sentence (e.g., the NP-VP break) will not have the highest TEP. It becomes clear from this that in order to obtain a high correlation between TEP and the linguistic structure, one needs to use balanced sentences that contain few deletables. As we have shown, correlation coefficients can be quite high when we suppress all, or almost all, deletable elements and use balanced sentences, and they can be very close to zero when the sentences contain many deletables and the surface structure is not balanced.

A second main finding is that the performance variables revealed by Grosjean et al. (1979) are also at work in a task such as rote recall: The performance structure model Grosjean et al. propose is a better predictor of hierarchical structures obtained from TEPs (in sentences with no deletables) than is the sole surface structure of the sentence. Thus TEPs join a number of other performance data (parsing values, pause durations, indexes of relatedness, and probe reacting times) that reflect the speaker-listener's subjective organization of sentence structure.

REFERENCES

Bever, T., Lackner, L., & Kirk, R. The underlying structures of sentences are the primary units of immediate speech processing. *Perception & Psychophysics*, 1969, 5, 225-234.

Fodor, J., & Bever, T. The psychological reality of linguistic segments. *Journal of Verbal Learning and Verbal Behavior*, 1965, 4, 414-420.

Frazier, L., & Fodor, J. The sausage machine: A new two-stage parsing model. *Cognition*, 1978, 6, 291-325.

Grosjean, F., Grosjean, L., & Lane, H. The patterns of silence: Performance structures in sentence production. *Cognitive Psychology*, 1979, 11, 58-81.

Grosjean, F., & Lane, H. Pauses and syntax in American Sign Language. *Cognition*, 1977, 2, 101-117.

Grosjean, F., Lane, H., Battison, R., & Teuber, H. The invariance of performance structures across language modality. *Journal of Experimental Psychology: Human Perception and Performance*, 1981, 7, 216-230.

Johnson, N. The psychological reality of phrase structure rules. *Journal of Verbal Learning and Verbal Behavior*, 1965, 4, 469-475.

Johnson, N. The effect of a difficult word on the transitional error probabilities within a sentence. *Journal of Verbal Learning and Verbal Behavior*, 1969, 8, 518-523.

Levelt W. V. M. Hierarchical chunking in sentence processing. *Perception & Psychophysics*, 1970, 8, 99-103.

Marmurek, H., & Johnson, N. Hierarchical organization as a determinant of sequential learning. *Memory & Cognition*, 1978, 6, 240-245.

Martin, E. Toward an analysis of subjective phrase structure. *Psychological Bulletin*, 1970, 74, 153-166.

Martin, J. G. Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review*, 1972, 79, 487-509.

Miller, G., & Isard, S. Some perceptual consequences of linguistic rules. *Journal of Verbal Learning and Verbal Behavior*, 1963, 2, 217-228.

Weinberg, S., & Goldberg, K. *Basic statistics for education and the behavioral sciences.* Boston: Houghton Mifflin, 1979.

NOTE

1. At least two routes are open in comparing two sets of data when a number of sentences are involved. The first, and the one we used here, is to compute a global correlation between the two, that is, a correlation based on the total number of word boundaries across all sentences (N = 154). The second is to compute individual correlations between the two variables for each sentence type and then to calculate the mean correlation across sentence types. We will not follow this second approach, as it gives results comparable to the first: We found, for instance, that the global correlation between TEP and pausing was .41 and the mean of the 14 individual correlations was .42, for TEP and parsing the correlations were .42 and .38, respectively, for TEP and CI we obtained .39 and .40, and for TEP and PI, .44 and .45, respectively.

(Received for publication August 12, 1980;
revision accepted February 4, 1981.)