JOURNAL OF MEMORY AND LANGUAGE 29, 133–159 (1990)

Sentence Memory: A Theoretical Analysis

WALTER KINTSCH AND DAVID WELSCH

University of Colorado, Boulder

FRANZ SCHMALHOFER

University of Freiburg, FRG

AND

SUSAN ZIMNY

Indiana University, Pittsburgh

How sentences from a discourse are recognized or verified can be explained by combining theories of item recognition derived from list-learning experiments with hypotheses about the representation of text in memory within the framework of the construction-integration model of discourse comprehension. The implications of such a theory of sentence recognition are worked out for two experimental situations. In the first experiment, subjects read brief texts and were then tested for recognition with verbatim old sentences, paraphrases, inferences, and contextually related and unrelated new distractor sentences after delays from 0 to 4 days. Differential decay rates for the wording and meaning of the text and for situational information were observed. The theory provides a good quantitative account of the data. In the second experiment, the speed-accuracy trade-off in sentence verification for two subject groups with different prior knowledge was studied for old verbatim sentences and inferences. Qualitative predictions derived from the theory with the parameter estimates from the first study were in agreement with the data. Readers without an adequate situational understanding (novices) were found to make quick judgments based on surface and textbase characteristics of the test sentences, while experts in addition utilized their situation model successfully, which required more processing time.

A large number of experiments on recognition memory exist for lists of words or pictures. Several models of recognition memory are available today which account very well for most of the phenomena observed in these experiments. Can these theories also account for experimental data when the materials used are not lists of items, but coherent discourse? By combining the essential features of current models of recognition memory developed for list-learning studies with a model of discourse comprehension and assumptions about the representation of discourse in memory, a model of sentence recognition and sentence verification can be obtained that accounts for major features of sentence-recognition data. Thus, we do not propose developing a new model for sentence memory. Instead, we shall combine existing models of list-learning and text-comprehension processes to derive a theoretical analysis of sentence recognition and verification.

We begin by comparing three current models of item recognition (Gillund & Shiffrin, 1984; Hintzman, 1988; Murdock, 1982) and determine their common essential features, to be used for modelling sentence memory. We will then review some notions

This research was supported by a Grant No. 15872 from the National Institute of Mental Health to Walter Kintsch and by Grant Schm 648/1 from the Deutsche Forschungsgemeinschaft to Franz Schmalhofer. Address reprint requests to Walter Kintsch, Department of Cognitive Science, University of Colorado, Campus Box 345, Boulder, CO 80309.

0749-596X/90 $3.00

about the representation of discourse in memory from van Dijk & Kintsch (1983) and briefly sketch the construction-integration model of discourse comprehension (Kintsch, 1988). Finally, we will show how these theoretical assumptions in combination provide an account of sentence recognition and verification data. We demonstrate that our model can be made to match a set of sentence-recognition data in which old verbatim sentences, paraphrases, inferences, and new sentences are used as test items for retention intervals varying between an immediate test and a four-day delay (Experiment 1). The model is further evaluated by testing qualitative implications for subject groups with different prior knowledge with respect to the speed-accuracy trade-off in sentence-verification judgments (Experiment 2).

Models of Item Recognition

Three models of recognition memory will be considered here, those of Hintzman (1988), Murdock (1982), and Gillund and Shiffrin (1984). All three models are formulated rigorously so that quantitative predictions are possible, and all appear to be empirically adequate in the domains to which they have been applied.

At first glance, the three models appear to be about as different as they could be in their basic makeup: Murdock's is a distributed memory model; Hintzman postulates multiple episodic traces; Gillund and Shiffrin conceive of memory as a network of interassociated nodes, while the other two models employ feature vectors. However, these models share some essential similarities when they are expressed formally, and it is these that we shall use as a basis for a model of sentence memory.

Hintzman (1988). This model is a multitrace model, in which each experience leaves its own memory trace. Memory traces, as well as test items, are represented as feature vectors, the values of the features being 1, −1, or 0. The similarity of a memory trace to some probe is the (weighted) dot product of their corresponding feature vectors. The total activation of a probe, its Intensity I, is given by the sum of the similarity values of the probe with all traces in memory. E(I) = 0 if the probe does not resemble any traces and increases as the quality of the match improves. For recognition judgments, the intensity I is fed into a decision mechanism.

Murdock (1982). Murdock also represents memory traces as well as test items as feature vectors. However, a single vector now represents the memory trace of a whole list of items with which the feature vectors of the test items are compared on a recognition test. Once again, corresponding features of the memory vector and the test vector are multiplied and the resulting values are summed to obtain a retrieval strength value, which is then used as input into a decision system. There are other versions of distributed memory models for item recognition which differ from Murdock in their mathematical formulation, but these differences are irrelevant at this general level of analysis.

Gillund and Shiffrin (1984). Unlike the previous two models, items in this model are represented as nodes related to each other by associate links in a retrieval structure. Suppose that there is a set of items [I], a test node T, and a context node C, with the similarity between a test node and an item I being S(T,I), and the similarity between the context node and item I being S(C,I). For recognition, the memory probe is assumed to consist of T and C, and the activation resulting from comparing the memory probe with item I is given by the product S(T,I)*S(C,I). The total activation of T is just the sum of the activations for each of the items in memory, and, as in the previous models, serves as a test statistic for a decision system.

Obviously, this brief description does not do justice to the three models considered here. Nevertheless, it suffices to make a few important points. The discrepancy in their verbal formulation notwithstanding,

they agree on three crucial mathematical properties. First, in all models the target is compared to *all* memory traces, and the sum of the comparison values provides the relevant test statistic. This sets these models apart from the previous generation of recognition models, where a recognition decision was thought to be dependent only upon the similarity of the target item to its corresponding memory trace. This is a crucial feature of item recognition. However, it does not appear to matter much exactly how this comparison between the set of memory traces and the target item is performed: whether the traces are summed first, and then the comparison is made (as in Murdock), or whether the comparisons are made first and their outcomes are then summed (as in Hintzman and Gillund & Shiffrin) makes no difference for present purposes.

Similarity between trace and target in the Hintzman and Murdock models is computed by the dot product of the corresponding feature vectors. In Gillund and Shiffrin, the links in the associative network represent familiarity values directly. The discourse comprehension theory as formulated in Kintsch (1988) lends itself more naturally to the latter approach, though a more molecular analysis would be possible in principle.

Finally, all three models use a decision mechanism to turn strength measures (Intensity, Familiarity, Similarity) into Yes–No decisions.

These three mathematical properties sufficiently specify the recognition mechanism for the model to be proposed here. The ideosyncratic features of the three models will be neglected in favor of these formal communalities. The fact that all three models fit recognition data about equally well implies that the features common to these models are responsible for the fit to the data. The other differences among the models represent either differences in theoretical metaphors and verbal interpretations of the common formal substance of the model, or require for their resolution a broader framework than just laboratory studies of item recognition.[1]

Levels of Representation

In experiments using simple stimulus materials it is common practice to represent the outcome of a match between a memory trace and a test probe by a single, unitary value. For sentence materials or discourse, on the other hand, this is no longer sufficient, and different types of information, which may play different roles in retrieval and decision making, must be distinguished (e.g., Ratcliff & McKoon, 1989). According to van Dijk & Kintsch (1983), three levels must be distinguished in the memory representation of discourse. At one level, a text is characterized by the exact words and phrases used. This is the surface level of representation. Syntactic theory provides the tools for the description and analysis of this level of representation. At another level, not the exact wording but the semantic content of the text must be represented. Both the local (microstructure) and global (macrostructure) characteristics of the text play a role here (Kintsch & van Dijk, 1978). Several representational schemes have been developed within linguistics, semantics, artificial intelligence, and psychology for this purpose. We shall use here the propositional representation first introduced in Kintsch (1974). The situation model is the third level of representation important for text comprehension (van Dijk & Kintsch, 1983). What is represented at this level is not the text itself, but the situation described by the text, detached from the text structure proper and embedded in pre-established fields of knowledge. The principle of organization at this level may not be the text's macrostructure, but the

[1] The authors of the models discussed here are concerned with general models of human memory. The formal similarity noted above does not hold outside the domain of item recognition.

knowledge schema (e.g., an appropriate script or frame) used to assimilate it.

In a number of experimental studies it has been shown that these three levels of representation can be distinguished in sentence recognition experiments (e.g., Fletcher & Chrysler, in press; Schmalhofer & Glavanov, 1986). Old verbatim sentences are represented at all three levels of representation: the surface structure, the textbase, and the situation model. Paraphrases of old sentences, on the other hand, differ in terms of the surface structure from what is stored in memory, but not at the textbase and situation model level. Inference statements that were not directly expressed in the text differ from the memory representation both in terms of their surface structure and propositional content, but they are part of the same situation model. Finally, contextually related, but not inferable test sentences differ from the memory representation at all three levels. Thus, by looking at the differences among these types of test sentences, estimates of the memory strength at each level of representation may be obtained in sentence recognition experiments.

The Construction–Integration Model

The construction–integration model of Kintsch (1988) describes how texts are represented in memory in the process of understanding and how they are integrated into the comprehender's knowledge base.

The crucial features of the model are as follows. Comprehension is simulated as a production system, the rules of which operate at various levels: some build propositions from the linguistic information provided by the text; some generate macropropositions; some retrieve knowledge from the comprehender's long-term memory that is related to the text, thus serving as mechanisms for elaboration and inference. All these rules share one general characteristic: they are weak, "dumb" rules that do not always achieve the desired results. In addition to what should have been constructed, these rules generate redundant, useless, and even contradictory material. In contrast, most other models of comprehension attempt to specify strong, "smart" rules, which, guided by schemata, arrive at just the right interpretations, activate just the right knowledge, and generate just the right inferences.

Smart rules necessarily must be quite complex, and it is very hard to make smart rules work right in ever-changing contexts. Weak rules, as they are used here, are obviously much more robust—but, left to themselves, they do not generate acceptable representations of the text. Irrelevant or contradictory items that have been generated by weak rules, however, can be eliminated, if we consider not just the items generated by the rules, but also the pattern of interrelationships among them. Generated items which are irrelevant to the text as a whole will be related only to one or a few other items. Contradictory items will be negatively connected to some of the other items in the network of items. Relevant items, on the other hand, will tend to be strongly interrelated—be it because they are derived from the same phrase in the text, or because they are close together in the textbase, or because they are related semantically or experientially in the comprehender's knowledge base. Thus, if activation is allowed to settle in the network, an integrated representation of the relevant items is obtained.

A simple example will illustrate these processes. Suppose we are concerned with the meaning of "bank" in "A large amount of money was lost when the bank was robbed by a masked gunman." A smart rule would assign "bank" the proper meaning on the basis of contextual information—we know money is more likely to be lost when a financial institution is robbed than a river bank. A dumb rule constructs interpretations for both of the meanings of "bank" that are known, "bank-1 was robbed" as well as "bank-2 was robbed." However,

we construct not just isolated propositions, but interrelate them in a network. This is possible because the propositions of a text are related in various ways—syntactically, semantically, via the discourse structure, and through general world knowledge. In consequence, a network of interrelated propositions can be obtained, in which "bank-1" is strongly connected with the rest of the text, the "money," the "masked gunman," etc., while the bank-2 propositions would not be connected with the rest of the network. Activation in this network will collect in those parts of the network that are tightly interrelated, and the isolated "bank-2" propositions will become deactivated. Thus, the network rejects the inappropriate interpretation that had been constructed. The construction–integration model in this way achieves with weak, robust construction rules followed by a spreading activation stage (integration) the same result that smart but complex rules would have achieved.

Kintsch (1988) not only describes the relevant details of this model, but also reports some results that (a) suggest that this kind of a model may capture some features of human comprehension processes better than "smart" comprehension models, and (b) demonstrate that the model is computationally adequate in some reasonably complex domains.

The construction–integration model provides a natural account of sentence recognition. First, comprehension of a paragraph is simulated in the way just outlined, resulting in a memory representation consisting of text propositions, plus whatever knowledge elaborations and inferences survived the integration process. These items have some sort of activation value—central, important propositions being more highly activated than peripheral ones—and they are related to each other in the ways specified by the model. Formally, this means we have an activation vector **A**, specifying for each constructed element a final activation value, and a coherence matrix **C**, specifying the relations among these elements. Together **A** and **C** characterize in the model the memory representation achieved as a result of comprehending this paragraph.

The model is then given the to-be-recognized test sentence to comprehend, for which it will construct the same kind of representation. In recognition, the representation of the test sentence is compared with the representation of the whole paragraph: we determine how much of the total activation flows into that part of the network that represents the test sentence. In the case of an old sentence, all parts of it are already part of the network and we merely have to add up the total activation of all its constituents. In the case of a new sentence, new elements have to be appended to the network. These are connected to the network in exactly the same way as the network was established originally. In the case of an inference or paraphrase, typically some of the elements correspond to already existing nodes in the network, while others have to be appended to the network. If a test sentence fits in well with the original text (e.g., it is actually a part of it), it will become strongly activated. If it has no connections at all to the original material, it will not be activated at all. The more similar it is to the original, the more connections there will be, and the more highly activated the test sentence will become. Thus, we can use the amount of activation that flows from the original paragraph to the test sentence as a measure of its familiarity or strength, and use a decision rule to derive a binary recognition response.

Consequently, the proposed model of sentence recognition is based on three components: a recognition mechanism from the list-learning literature, the notion that discourse is represented at different levels, and the processing mechanisms of the construction–integration model. The test item—the test sentence—is compared, at each level of representation, against all items in memory—the whole text. The

comparison yields an index of the similarity between what is remembered and the test item, as measured by the amount of activation that flows from the memory representation into the test item. This similarity index is then used in a decision mechanism. Thus, the recognition principles derived from the list learning literature have been embedded into the framework of the construction–integration model.

In the next section, an experiment on sentence recognition from discourse will be described. These data will provide the setting for the detailed and formal development of our model.

Sentence Recognition

Experiment 1

Zimny (1987) studied sentence recognition for verbatim old sentences, paraphrases, inferences, and two types of distractor sentences for retention intervals up to four days. She constructed 18 texts of about 150 to 200 words each, based on the scriptal norms of Galambos (1982). Each text described a sequence of scriptal events (e.g., "Nick goes to the movies") by stringing together high-frequency, familiar actions from the norms, interspersed with some nonscriptal material (e.g., his girlfriend wore a dress with pink polka dots). The reason for constructing these texts according to script norms was so that we knew what sort of situation model was likely to be constructed for each text, namely a script-based one. Linguistic analyses specify the structure of the surface representation for arbitrary texts, and propositional analyses are similarly general, yielding textbase hierarchies for a wide variety of texts. Unfortunately, this is not the case for the situation model: for most texts we have no clear idea what sort of a situation model would be generated. Consequently, we must work with special cases where enough research has been done to establish this kind of information. Research in this area has therefore focused on a few cases such as maps, as in Perrig and Kintsch (1985); mental models, as in Johnson-Laird (1983); or scripts, as in Bower, Black, and Turner (1979) as well as the present case.

For each text, Zimny constructed five test sentences which vary in terms of their level of discourse representation. Old sentences appeared at test as they had in the original text, and are represented at the surface, textbase, and situation model levels. Paraphrases involved minimal word order or single word changes; they are identical with sentences from the text at the levels of their textbase and situation model, but differ in some ways in their surface structure. Inferences were sentences that could be inferred by readers from the surrounding context with high reliability; these sentences fit into the same situation model as actual sentences from the text, but they differed both in terms of their textbase and surface representations. While an attempt was made to keep the test sentences similar in terms of their length and complexity, they obviously had to differ in numerous ways, with some being much more salient and recognizable than others. Therefore, Zimny wrote three different versions of her texts, so that each sentence could serve either as an old, paraphrase, or inference sentence. In addition, two entirely new test sentences were used with each text. One sentence was contextually appropriate, while the other was unrelated to the theme and context of the text and served as the baseline for the recognition analysis.

One group of subjects was asked to recognize the test sentences for each text right after reading the text. Subjects were instructed to answer "Yes" if they thought they had seen the sentence before, and "No" otherwise. Three other groups of subjects received the test sentences after delays of 40 min, 2 days, or 4 days.

The results most relevant for present purposes are shown in Figs. 1 and 2. Figure 1 shows the percent "Yes" responses subjects gave to old test sentences, paraphrases, inferences, as well as context appropriate and context inappropriate distractor items as a function of delay. The old

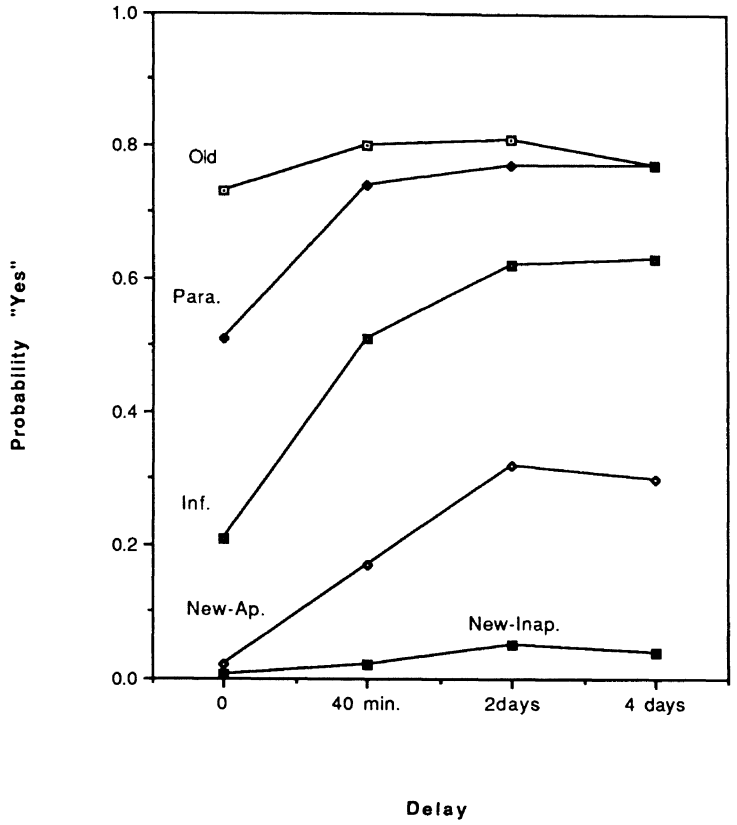

FIG. 1. Probability of Yes responses for old sentences, paraphrases, inferences, and context appropriate and inappropriate new sentences as a function of delay; after Zimny (1987).

sentences, paraphrases, and inferences as well as the main effect of delay were both significant statistically, but most importantly, there was a significant interaction between these factors, $F(6,280) = 38.7$, $p < .001$. Figure 2 provides estimates of the trace strengths at the three levels of representation over the delay intervals. The percent "Yes" data were first turned into d' measures by using the context inappropriate distractor items as a baseline. This transformation was necessary to remove strong, delay-dependent bias effects from the analysis: on the immediate test, subjects used a strict criterion for saying they had seen a sentence before, but after four days they were willing to assert this on the basis of much weaker evidence. (Note the increase in Yes responses as a function of delay in Fig. 1). Secondly, difference measures between the d's were computed. The difference between the memory strengths of old sentences and paraphrases provides a measure of the strength of the surface representation (how something was said). The difference between the strengths of the paraphrase sentences and inferences provides a measure of the strength of the textbase representation (whether something was actually said in the text or not). And

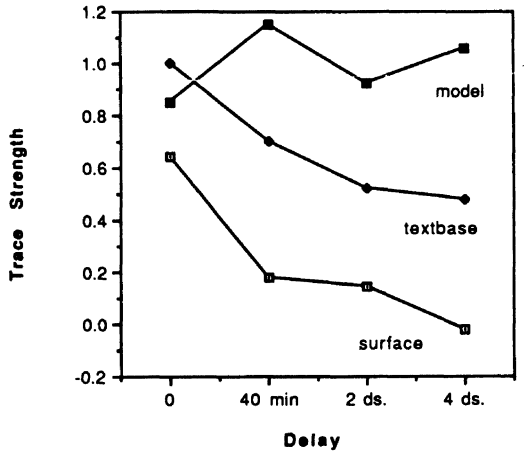

FIG. 2. Estimated strengths of the surface, textbase, and model traces; after Zimny (1987).

finally, the difference between the strength of the contextually appropriate distractor items and the inference sentences provides a measure of the strength of the situation model (whether something is true in the given situational context or not). These difference values are plotted in Fig. 2. A statistical analysis of these data revealed that, in addition to significant main effects, the interaction between delay and trace type was also significant statistically, $F(6,280) = 6.29, p < .001$.

Figure 2 shows some interesting trends. First of all, surface memory was found only on the immediate test. Memory for the textbase was quite strong initially, decreased with delay, but remained above zero even after four days. Situational memory, on the other hand, stayed at a high level, independent of delay.[2] These are the data that will be modelled here.

The Memory Representation of the Text

To derive theoretical predictions for the data from the Zimny experiment, somewhat different aspects of the construction–integration model will have to be emphasized than in Kintsch (1988). In Kintsch (1988) the memory representation of a text was developed only at the propositional level: surface traces, as well as situational representations were neglected. Obviously, these distinctions must be made explicit in a treatment of sentence recognition. On the other hand, the focus of Kintsch (1988) was on the performance of the model as an inference engine—something that we shall neglect in the present application of the model. The reason for omitting this aspect of the model is that it does little actual work in the present application, and that its inclusion would make an already complex story even more complicated. This simplification does introduce some distortions, however, which will have to be considered after the simplified case has been presented.

The Zimny data are averaged over subjects and sentences. Predictions will be derived for a single text which is much briefer than the original texts used by Zimny, and for only a few specific test sentences. While these materials are not atypical, it is certainly the case that for another text example and other test sentences somewhat different quantitative predictions and parameter values may have been obtained. Thus, predictions for a "typical" subject and material set are compared here with data averaged over subjects and materials.

The following two-sentence text will be used as the input text: *Nick decided to go to the movies. He looked at the newspaper to see what was playing.* (This is the beginning of a text based on a Going-to-the-Movies script used by Zimny (1987), which then continues through the whole event.) In Kintsch (1988), this text would have been broken down into propositional units (such as NICK, (GO-TO,NICK,MOVIES), etc.) which then would activate knowledge (perhaps *Nick wanted to see a film*) through their associative links in the reader's long-term memory store. This propositional structure would be consolidated through an integration process which eliminates the context-irrelevant knowledge that had been activated. For the sake of simplicity, we omit the knowledge activation process in this application, and only look at the actual text contents, as explained above. Instead, the role of surface properties of the text as well as the situation model in sentence recognition will be modelled: we make explicit in our analysis the linguistic relations as well as the scriptal relations among the input units in the text.

A simulation of the model constructs a network of text elements that specifies how strongly each element is related to every other element in the network. We are con-

[2] The task-dependent nature of these results should be emphasized: long-term memory for surface features is frequently observed in other contexts, as is forgetting of situational information. Forgetting rates are clearly material- and task-dependent (for a review, see Kintsch & Ericsson, in press).

cerned with three types of relationships, corresponding to the three levels of representation of text in memory. Within each level, we specify relation strengths in terms of distances among nodes in a coherence network. The pattern of interconnectedness among these nodes will determine the degree of activation each element will receive.

In Fig. 3, 10 word groups (linguistic elements, L) have been distinguished in the text. Most of these correspond to propositions (P) as well as elements of the situation model (M), except P7 and M7 do not have a corresponding linguistic element L7. The linguistic elements form syntactic chunks (S) according to the phrase structure of the sentences [e.g., L3 (*to-go-to*) and L4 (*the-movies*) combine to form the chunk S3]. Together, L and S constitute the elements of the surface representation of the text. (They are distinguished here merely for convenience, to allow a ready comparison between the actual words and phrases used in the text and the propositions or situation model elements corresponding to these words or phrases.) The graph shown in Fig. 3 allows one to calculate a distance matrix among the L- and S-elements: for instance, L1 is one step away from S1, three steps away from L2, and not connected to L10.

The propositions P1 to P9 are connected to each other in a somewhat different pattern. Following Kintsch (1974), one can approximate the structure of a propositional textbase by noting the pattern of argument overlap among the propositions. For example, P1 appears as an argument in P2, P3, P5, and P8, while P2 overlaps with P1 and P3. The textbase structure obtained via argument overlap is shown in Fig. 4. This network defines a distance matrix among the propositional elements: P2 is a single step away from P1, three steps away from P7, and four steps away from P9.

A similar distance matrix can be computed for the elements of the situation model. Since the text was explicitly con-

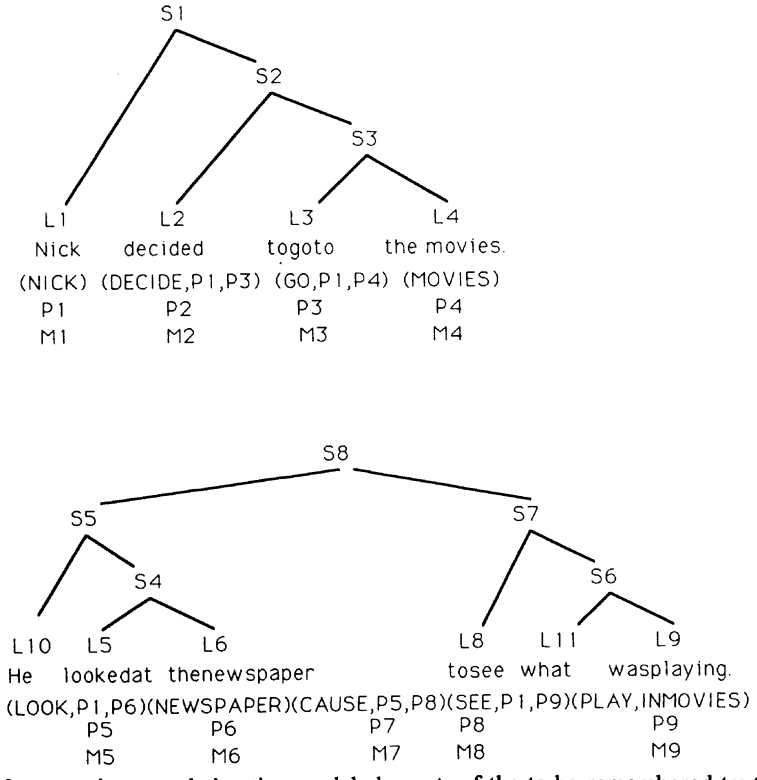

FIG. 3. Surface, textbase, and situation model elements of the to-be-remembered text.

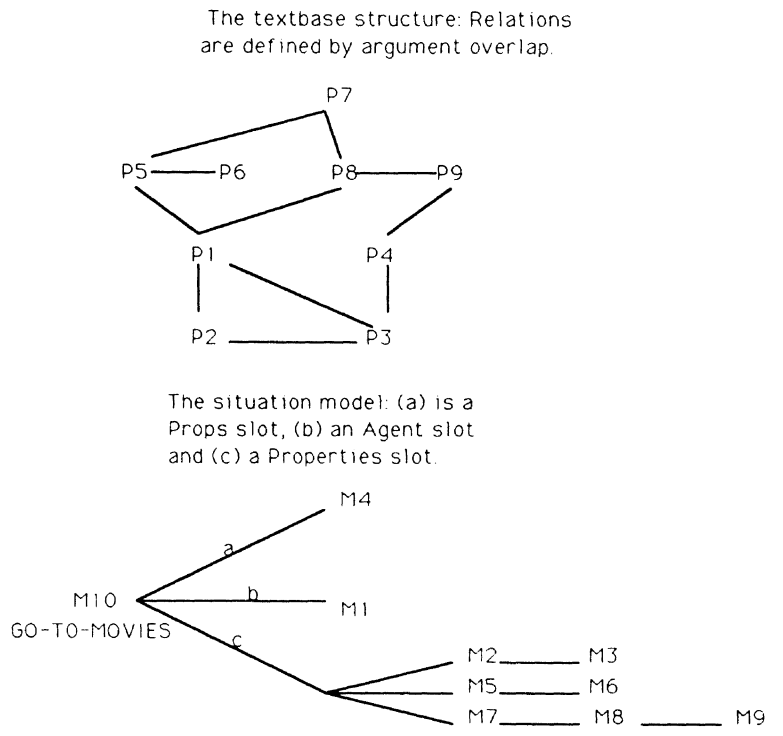

FIG. 4. The coherence nets formed by the textbase and the situation model.

structed from script norms, it is assumed that the situation model in this case is structured as a script (i.e. as a schema with slots for Properties, Agents, Preparatory Steps, etc.) (e.g., Schank & Abelson, 1977). The script header M10 must be added to the items directly derived from the text—an exception to the policy of neglecting all inferences in the present application of the model. The resulting structure is also shown in Fig. 4. This time, M2 is one step away from M3, two steps from M1, one step from M7, and three steps from M9.

It is not necessary to think of L1 (the exact word used in the text), P1 (the corresponding proposition), and M1 (an element of the situation model) as three distinct objects in the reader's memory representation. It is the same "Nick" in all three cases, but viewed once from a linguistic perspective where it enters into a certain set of relations with other linguistic elements, once considered as a proposition which plays a role in the textbase, and once considered in terms of its role in the "Go-to-the-Movies" script. For analytic purposes it is useful to distinguish L, P, and M units, but what matters conceptually is that text elements enter into different relationships with other elements, depending upon the level of analysis: surface, propositional, or situational.[3]

For the analyses in Figs. 3 and 4 it was necessary to work with a particular phrase structure grammar, textbases were constructed in a particular way, and the scripts were assumed to have particular forms. There are, of course, other phrase structure grammars, textbases need not be based on argument overlap, and different assumptions about the slots of a script could be made. However, the analyses used here are well-motivated and well-established, and most alternative analyses would in practice

[3] The reason we do not just have an element "1" instead of L1, P1, and M1, adding the three types of relationships together, is that on recognition tests we are usually dealing with only one of these elements, but not the others.

be highly correlated. Nevertheless, more sophisticated analyses (e.g., a textbase that explicitly takes into account causal connections) might lead to marginally better results.

The relationships shown in Figs. 3 and 4 define a network which can be represented by a coherence matrix, which will provide the basis for the integration process. The rows and columns of this matrix are given by the elements L1 to L11, S1 to S8, P1 to P9, and M1 to M10. The entries of the matrix designate the strength of the relationship between row and column elements. At this point numerical parameters must be estimated for the strength of relations among the elements which are shown in the graphs of Figs. 3 and 4. An unsystematic trial-and-error procedure was employed to obtain these estimates. Intuition suggests that local relations in the surface structure and textbase are quite strong but weaken rapidly as the distance between items increases. Hence, values of 5 and 3 were used in the coherence matrix for items 0 (the relationship of a node to itself) and 1 step apart in either in the surface structure or in the or textbase. All other connections were set to 0. On the other hand, scripts are more stable long-term memory structures, allowing for more long-distance relations, so that strength values of 4, 3, 2 and 1 were assigned to items 0, 1, 2 and 3 steps apart in the script structure, respectively. Finally, a value of 4 was used to tie together the same node at different levels of representation [e.g., L1 to P1, and P1 to M1 (we assumed there were no direct connections between L1 and M1)]. In consequence, the effective connections for the surface and textbase elements in the coherence matrix correspond to the links shown in Figs. 3 and 4, but the connections among the model elements are much richer, since not only neighboring nodes are directly connected, but also nodes two and three steps apart in Fig. 4.

The parameters estimated here are unique up to a multiplicative constant. These estimates in part reflect general constraints, such as the farther away, the weaker the connection must be. Within these constraints, the exact numerical values obtained result from goodness of fit considerations. For instance, replacing 5-3-0 for surface and textbase connections with 4-2-0 gives somewhat less satisfactory fits, but replacing it with 4-3-2-1-0, the values used for the situation model connections, destroys the fit entirely, as does a 5-3-0 choice for the situation model. Thus, there seems to be a substantive interpretation for these estimates: surface and textbase connections are strong initially, but very local, while the important feature of the situation model connections is that they reach beyond their immediate neighbors.

In this way a 38 $\times$ 38 coherence matrix was obtained for the text under consideration. Each of the 38 items was assigned an initial weight of 1/38 in an activation vector A1. This activation vector was successively multiplied with the coherence matrix to allow the activation to spread from the initial elements through the connections specified by the coherence matrix to other parts of the network, and finally, to settle in those parts of the network where the greatest interconnectivity exists. After each multiplication, the resulting activation vector was renormalized so that the sum of all activation values was 1. After 7 such cycles the average change in activation was less than .0001, and the process of spreading activation was stopped at that point. Figure 5 shows the pattern of activation over the 38 elements in the activation vector. L and S elements wind up with relatively low activation values (because only a few linguistic connections contribute to the spread of activation, given the matrix structure and parameter values assumed above). P elements are more strongly activated, partly because they are embedded in a more strongly interconnected network than the linguistic elements, and partly because they are directly connected to the dominant M elements. The reason for the higher activation of the M elements is of course their much

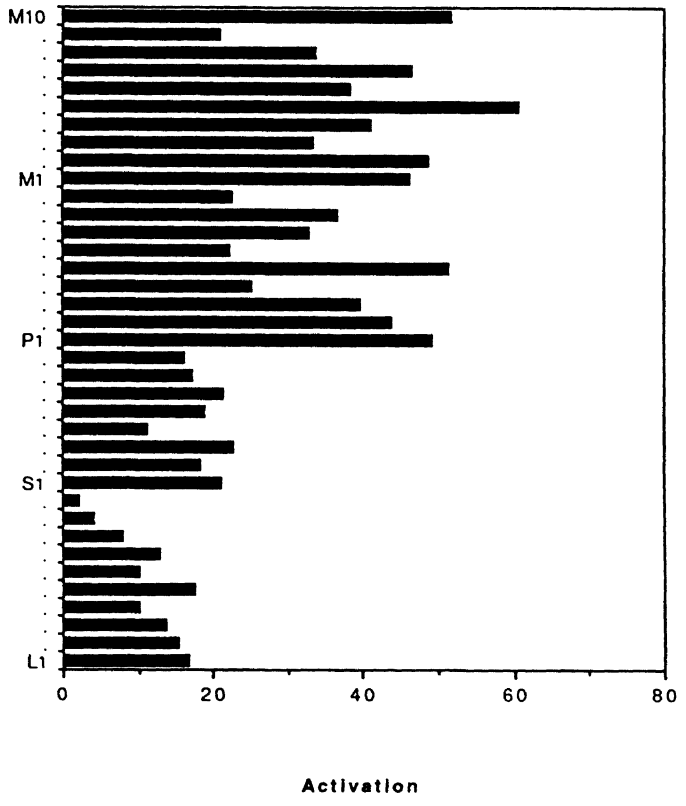

FIG. 5. Final activation values (multiplied by 1000) of the language units (L1 to L10), the surface chunks (S1 to S8), propositions (P1 to P9), and model elements (M1 to M10).

greater interconnectedness. Note that the only inference admitted here, the "Going-to-the-Movies" script header, has become one of the most highly activated items.

The memory trace after reading the text, then, consists of three components: the 38 elements that were constructed from the text (in the general case, these would be augmented by a substantial amount of activated knowledge—inferences and elaborations), their interconnections as represented by the coherence matrix C, and their activation values, given by the activation vector A.

Recognition of Test Sentences

We can now turn to the recognition test. First, consider an old test sentence that is taken verbatim from the original text (e.g., *He looked at the newspaper*). As in the memory models discussed above, the familiarity value of this sentence is based on the dot product T*A, where T is a vector with unit activation in all elements associated with the test sentence and A is the activation vector.[4] The results of this calculation are shown in Table 1. Paraphrases, inferences, and other new test sentences are treated in exactly the same way as old test sentences, except that the construction processes upon reading the test sentence now introduce elements into the network which were not present in the original representation of the to-be-remembered text.

Consider a paraphrase, such as *Nick*

[4] It is difficult to decide whether the sum or the average provides a better test statistic. Summing the activation values of the elements, as is done in taking a dot product, favors longer test sentences over shorter ones. It is an empirical question whether subjects are more likely to respond "yes" to longer sentences, if the average activation value of the elements is held constant. We are not aware of experimental evidence that could decide this issue.

TABLE 1
TEST SENTENCES AND THEIR FAMILIARITY VALUES

OLD "He looked at the newspaper"		PARAPHRASE "Nick studied the newspaper"	
L10	18	L1	19
L5	17	studied	0
L6	10	L6	10
S4	11	S	2
S5	24	S	4
P1	53	P1	53
P5	52	P5	51
P6	22	P6	22
M1	46	M1	45
M5	58	M5	58
M6	36	M6	36
Total	347	Total	300

INFERENCES "Nick wanted to see a film"		"Nick bought the newspaper"	
L1	11	L1	17
wanted	0	bought	10
to-see	0	L6	1
a-film	0	S	3
S	0	S	4
S	0	P1	52
S	2	[BUY,P1,P6]	24
P1	39	P6	14
[WANT,P1,P]	8	M1	44
[SEE,P1,P]	8	[BUY,M1,M6]	40
[FILM]	1	M6	19
M1	56	Total	228
[WANT,M1,M]	52		
[SEE,M1,M]	30		
[FILM]	49		
Total	256		

NEW "Nick went swimming"	
L1	19
went	0
swimming	0
S	1
S	4
P1	58
[GO,P1,P]	13
[SWIM,P1]	13
M1	45
Total	153

Note. The activation values of each element of a test sentences are shown (multiplied by 1000). Old elements are labelled as in Fig. 3; new elements are written out or, in the case of linguistic chunks, indicated by an S.

studied the newspaper. This time, the test sentence is only in part contained in the existing memory representation of the original text, so that we have to add several new elements to the coherence matrix in order to represent both the existing memory trace and the given test sentence. An inspection of Fig. 3 shows that there are three elements to be added: the word *studied* (but not the proposition P5, which re-

mains unaffected by the substitution of a synonym), as well as two new S elements (in place of S4 and S5). These three new elements are added to the coherence matrix and connected with the existing memory structure in the same way as the original elements themselves were interconnected (so that *"studied"* is two steps away from L6, *"the newspaper"*, but three, via the two new S-units, from L10, *"he"*, etc.). Thus, an expanded coherence matrix C_p is obtained. Activation is now spread through this new structure until the activation vector A_p stabilizes, which occurs after just 2 cycles. Table 1 shows the resulting pattern of activation for this test sentence. Its familiarity is slightly below that of the old, verbatim sentence, in qualitative agreement with Zimny's data.

The computation of familiarity values is shown for two inference sentences in Table 1. The first test sentence, *"Nick wanted to see a film"* is composed almost entirely of new elements, requiring the addition of 12 items to the original coherence matrix. It is a plausible inference (though not a logically necessary one), and its familiarity value comes out quite high, though well below that of the paraphrase sentence. The second inference sentence *"Nick bought the newspaper"* shares more elements with the original memory structure, but does not fit into the script structure as tightly as the first (*wanting to see a film* is itself a preparatory step in the Movies script, while *buying the newspaper* is just something appended to the newspaper introduced earlier). As a result, the second inference receives slightly less activation than the first. Finally, the familiarity value of a distractor sentence *"Nick went swimming"* is computed in Table 1; its only connection with the original paragraph is the name "Nick," and it receives the lowest activation value, as it should.

With additional assumptions about forgetting, further predictions can be derived. Suppose we simulate memory for two delay intervals, a short delay, corresponding to Zimny's 40-min interval, and a long delay, corresponding to the 4-day delay. We want to derive predictions for the time of recognition testing (i.e., after the paragraph has been read and after forgetting has taken place). We are assuming that the effect of forgetting is a weakening of the connections between the items in memory, with the connections among surface traces decaying most rapidly, textbase connections less so, while the situation model remains intact, as in the Zimny study (Fig. 2). Numerically, this means that we set surface and textbase connections to 4 and 2 for 0- and 1-step distances (instead 5 and 3) to simulate the short-delay test. For the long-delay test, all surface connections are set to 0, and textbase connections to 3 and 1, for 0- and 1-step distances, respectively. (Note that we are in effect collapsing acquisition and retention into a single matrix here.) Then, the same calculations are performed as in Table 1. However, the resulting activation values are not directly comparable across the three delay intervals, because of the way activation vectors have been renormalized after each multiplication. By keeping the total activation always at 1, the activation vectors indicate only relative values among the items in each vector, but not absolute values across different matrices. Such a model would incorrectly imply that the overall response strength does not decrease with delay, although the individual connection strengths are assumed to decrease. In order to avoid this consequence of the normalization procedure, each activation vector must be weighted by the total sum of all entries in the corresponding coherence matrix (alternatively, we could have incorporated this weighting in the normalization procedure). If there are many and numerically stronger connections in a matrix (immediately after reading), activation will reach a higher level than if there are fewer and weaker connections (after 4 days). These absolute strength values for the three

delay intervals are shown for old sentences, paraphrases, inferences, and new sentences in Fig. 6.

We want to compare Fig. 6 with the data shown in Figs. 1 and 2 to assess the goodness of fit of the model for sentence recognition proposed here. It is not a straight comparison, however, for the data are based upon averages over many subjects and a substantial number of texts and test sentences. Figure 6 is based upon the few examples shown in Table 1. Would another set of sample sentences yield identical, or even similar results? Yes, in the sense that we can find many sentences of each type that would be essentially interchangeable with the ones we have chosen; no, in the sense that we could find many sentences that would not.

That is not necessarily a weakness of our approach—it is more a reflection of the reality of sentence recognition. One old sentence is not necessarily like another one, and paraphrases and inferences are even less so. Given a text, one can define a set of "old" sentences or phrases (e.g., for the text in Fig. 3, there is the first sentence. Old 1; the first phrase of the second sentence-,which we have actually used as our example and which we shall call Old 2; and the second phrase of that sentence, Old 3). One can compute the activation values (summed over delay) for all three of these test sentences (or phrases), and find an average value. The variation among these sentences is substantial, however: Old 1 (*"Nick decided to go to the movies"*) has an activation value which is 124% of the average; Old 3 (*"to see a what was playing"*) is only 72% of the average; and only our choice, Old 2 (*"He looked at the newspaper"*) is close to the average, 103%. Old 3, in fact, is

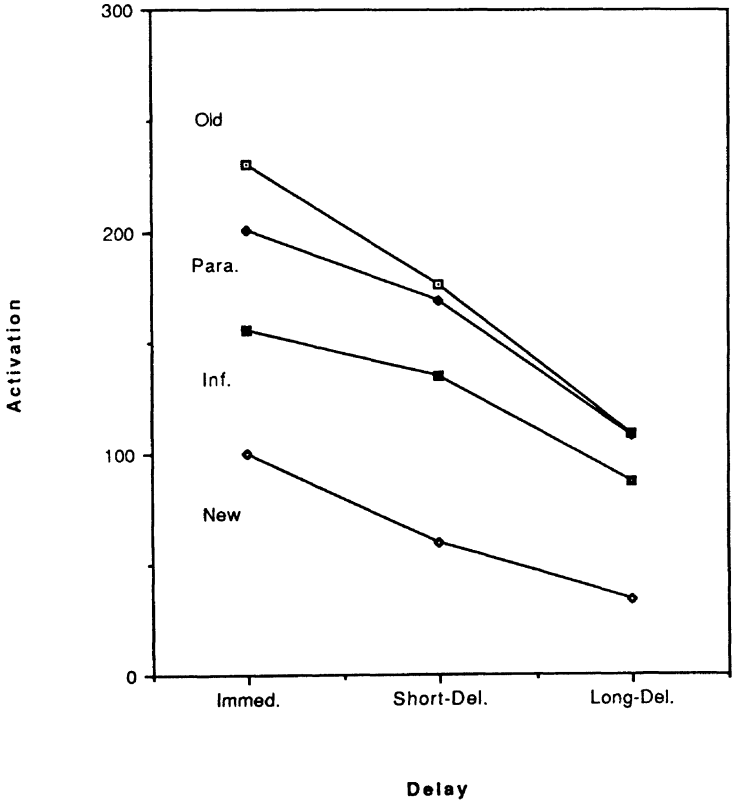

FIG. 6. Absolute activation values for the old test sentence, paraphrase, inference, and new test sentence as a function of delay.

closer to the paraphrase in Fig. 6 than to Old 2. Intuitively, that is not at all surprising: Old 1 seems just so much more memorable than Old 3. And it is surely no rarity empirically to find that a subtle paraphrase of a salient sentence is recognized better than many old sentences.

The situation is even more complex for inferences. We have used two inferences in Table 1 which differ considerably in their activation values. Indeed, *"Nick wanted to see a film"* is slightly closer in activation to the paraphrase in Table 1 than to the other inference, *"Nick bought the newspaper"*, and it takes little imagination to come up with inferences even more discrepant. Consider *"Nick wore pants"*, which we can infer from our little story probably with greater certainty than *"Nick bought the newspaper"*. In our model, it would receive an activation value identical to *"Nick went swimming"*, our false test sentence. But that is just as it should be, because that is what we would expect to find empirically! We talk about (pragmatic) "inferences" as if they constituted a class of sentences with respect to a given text that has certain well-defined common properties. That is just not the case. There are all kinds of inferences, and they will behave very differently in experiments, as well as in our model. The experimenter carefully constructed her stimulus materials by selecting typical, well-behaved sentences, and avoiding unusual ones as best she could. The theoretician must similarly perform his analyses on materials that are reasonably prototypical. There is no pool of "inference sentences" (or, for that matter, paraphrases or distractors) from which items could be sampled randomly, and to which, therefore, the results could be generalized. The domain of generalization is an informal one: texts and sentences that are like the ones used here. One could, in principle, derive predictions for all the materials used in an experiment. However, the labor involved would be prohibitive and not much would be gained, because we would still want to generalize our results not just to the sentences actually used in the experiment, but to all sentences like those used—an ill-defined domain. Our primary concern is not to predict what will happen with a set of particular texts and sentence materials, but with exploring whether the processes postulated by the theory can provide a good account of sentence recognition.

Obviously, Fig. 6 gives a fair qualitative account of the data in Figs. 1 and 2. The differences in response strengths between old items and paraphrases disappear as delay increases, and old items, inferences, and new items converge, but not completely. In order to go from the strength values shown in Fig. 6 to Yes–No responses, further assumptions need to be made about how strength values are transformed into Yes–No decisions. A simple response-strength model was assumed employing a ratio rule. The probability of a Yes response was computed by subtracting from each strength value a delay-specific threshold value and dividing the result by the total response strength, mapping the strength values into the [0,1] interval. Thus, four parameters need to be estimated for this purpose: a threshold for a Yes response for each of three delay intervals, and a value for the total response strength. The reason for introducing delay-specific thresholds at this point lies in the bias effects observed over the four-day delay in the Yes responses in Zimny's experiment (Fig. 1): we removed these biases by focusing on d' statistics and activation strengths, but now that we want to account for Yes responses, these bias effects have to be reintroduced. The fourth parameter, on the other hand, is simply a scaling factor needed to map strength values into probabilities. These four parameters were estimated by the method of least squares. The resulting fit to the data from Fig. 1 is shown in Fig. 7.

It would be hard to improve the fit of the

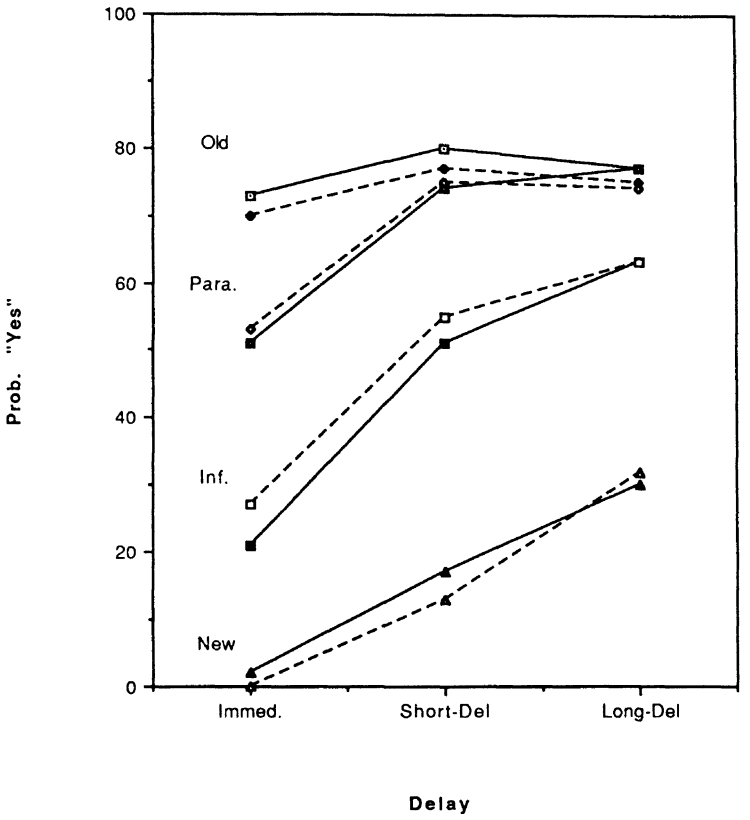

FIG. 7. Observed (——) and predicted (------) percent yes responses as a function of sentence type and delay.

predictions in Fig. 7 through more sophisticated methods of parameter estimation for the coherence matrices, or a more elaborate decision model. Clearly, the present model does very well, in that it gives a good qualitative account of the data (Table 1; Fig. 6), as well as a good quantitative fit (Fig. 7).

In evaluating the fit of the model it must be remembered that we have not constructed an ad hoc model for sentence recognition, but have put together this model from various existing components: a recognition mechanism from the list-learning literature, ideas about the memory representation, and a model of comprehension processes from recent work on discourse processing. Neither is there anything new about the way memory representations are constructed here: phrase structure chunks, textbases, and scripts are all familiar and widely used. Even the parameters in the model are constrained, both a priori (connection strengths can decrease with delay, but not increase), and empirically (surface traces must decay rapidly, textbase traces more slowly and incompletely, and model traces not at all). A theory of sentence recognition has been constructed largely from old parts, and it appears to be empirically adequate.

Nevertheless, a more skeptical view is also possible. There are a large number of parameters in the theory, and although it is not known how many are really free to vary (nor how this relates to the degrees of freedom in the data), their precise values are probably underconstrained. Furthermore, illustrative predictions for particular test sentences are used as a basis for predicting data averaged over many texts and sentences as well as subjects. In short, it is not

entirely obvious what is responsible for the good fit that was obtained—the theoretical principles emphasized here, or the design decisions made in putting this theory together.

To some extent this dilemma reflects the fact that it is hardly ever possible to evaluate a complex theory with respect to a single set of data. Fortunately, the theory makes some additional predictions that do not depend on any further parameter estimation. If the model presented here is more or less correct, then other predictions about sentence recognition follow which can be evaluated at least qualitatively without further parameter estimation.

Speed-Accuracy Trade-Off Functions

In deriving the predictions for the Zimny (1987) data shown in Figs. 6 and 7, two different inference statements were used as examples. Both were pragmatic inferences that people were likely to make in this context, but they differed in interesting ways. The first inference, *"Nick wanted to see a film"* is strongly related to the text at the level of the situation model: it is a common (though not a necessary) prerequisite for going to the movies. On the other hand, at the textbase and surface levels, the connection is made only by a single term, *"Nick"*. In contrast, the second inference, *"Nick bought the newspaper"*, shares both *"Nick"* and *"newspaper"* with the original text at the surface and textbase levels, but is not directly related to the going-to-the-movies script; it is merely an addendum to *"newspaper"*. This makes an interesting difference in the way the present model handles these statements.

As was shown in Table 1, the *wanting-to-see-a-film* inference accrues more activation (256 units) than the *buying-the-newspaper* inference (228 units). However, there is a significant difference in the speed with which this accrual occurs. In the first case, the amount of activation attracted by the inference statement in the first cycle is low (173 units, or 68% of the eventual total), and rises rather slowly over 13 cycles to its asymptotic value. The second inference, on the other hand, gets most of its activation right away (198 units, or 87%, so it is initially the stronger one) and reaches asymptote in 9 cycles. These examples suggest that model-based inferences are weak initially but increase in strength to a high value with enough processing, while inferences that are based more on surface similarity acquire activation quickly, but do not change much with further processing. In the model, this is obviously a consequence of the fact that surface and textbase relations are very local, while the situation model network is more extended. The way to test this hypothesis would be to collect speed-accuracy trade-off data for inference statements differing as outlined above. Such data are not available, but the arguments presented here suggest that it would be interesting to look at other speed-accuracy trade-off experiments in which differences at the level of the situation model are likely to play a role. For example, we may compare subjects who have developed an adequate situation model (experts) to less knowledgeable subjects (novices) with a weak situation model but appropriate surface and textbase strategies. We would then predict that inference sentences rise faster to their asymptote for novices than for expert subjects. The performance asymptote itself, however, should be higher for experts than for novices.

One such experiment has recently been performed by Schmalhofer, Boschert, and Kühn (1990). Schmalhofer et al. collected data from novices and experts verifying sentences from a highly technical text (an introduction to some features of the programming language LISP). They found rather striking differences in the speed-accuracy functions for these two groups of subjects, and we shall try to account for these differences by means of the hypotheses suggested above. In the Zimny data we

are dealing with different types of inferences (surface- vs. model-based similarity), while Schmalhofer et al. deal with different types of subjects (experts with a good situation model and novices with an incomplete or faulty situation model). In both cases, the present model predicts quite different speed-accuracy trade-off functions for inferences because of the role the situation model plays in these decisions.

Experiment 2

Schmalhofer et al. (1990) had 39 subjects study brief texts introducing them to the programming language LISP. Half of the subjects had no programming experience, while the other half were proficient in the programming language Pascal (but had no experience with LISP). Therefore, the subjects with programming experience (the expert group) presumably knew about functions in general, and when studying the LISP text, could employ their function schemata to understand what they were reading (i.e., construct an appropriate situation model). Novices (the novice group), on the other hand, were presumably unable to do so within the relatively short time they were allowed to study these texts. However, they certainly could understand the words and phrases they read and form a coherent text base.

Subjects were tested on four texts. An example of a text used in the experiment is shown in Table 2, together with two types of test sentences: an old verbatim sentence and a correct inference. Paraphrases of old sentences and incorrect distractor items were also used in the experiment, but since we do not derive theoretical predictions for these items, they will be omitted here. Subjects were asked to verify whether or not the test sentences were true, and to provide confidence judgments.[5] When a test sentence was presented, a subject made six responses in a sequence, at 1.5-s intervals when signal tones were presented. The first response signal occurred 750 ms before the sentence appeared on the screen. Obviously, subjects could only guess at that time, but as the other response signals were presented they had increasingly more time to fully process each test sentence. The last response signal differed from the previous ones, indicating that there was no time pressure for the final response. Each sentence could thus be fully processed.

TABLE 2
A PARAGRAPH FROM THE TEXT USED IN EXPERIMENT 2 AND SAMPLE TEST SENTENCES

Original text
The function FIRST is used to extract the first S-term from a combined S-term. The function FIRST has exactly one argument. The argument of the function must be a combined S-term. The value of the function is the first S-term of the argument.
Test sentences
Old
The function FIRST is used to extract the first S-term.
Inference
A single S-term is produced by the function FIRST.

Figure 8 shows the probability of Yes responses to old verbatim sentences for the two subject groups as a function of time. Subjects start at a value near 50%, but improve rapidly and reach an asymptotic accuracy of 93% and 94%, respectively, in about 9.5 s. The curves for expert and novice subjects are clearly identical. A 2 (Groups) × 5 (Response Signals) ANOVA on arcsine transformed data, omitting the first pure guessing point, yielded a highly significant main effect for Response Signals, $F(4,144) = 8.83$, $p < .0001$, but no group effect nor interaction, both $F < 1$.

The data for the inference statements in Fig. 9 are quite different. Neither main effect is significant, $F(1,36) = 2.76$ for Groups and $F < 1$ for Response Signals, but there is a significant interaction, $F(4,144) = 2.72$, $p < .05$. Novices tend to accept infer-

[5] Schmalhofer (1986) has found the same pattern of responses for verification as for sentence recognition.

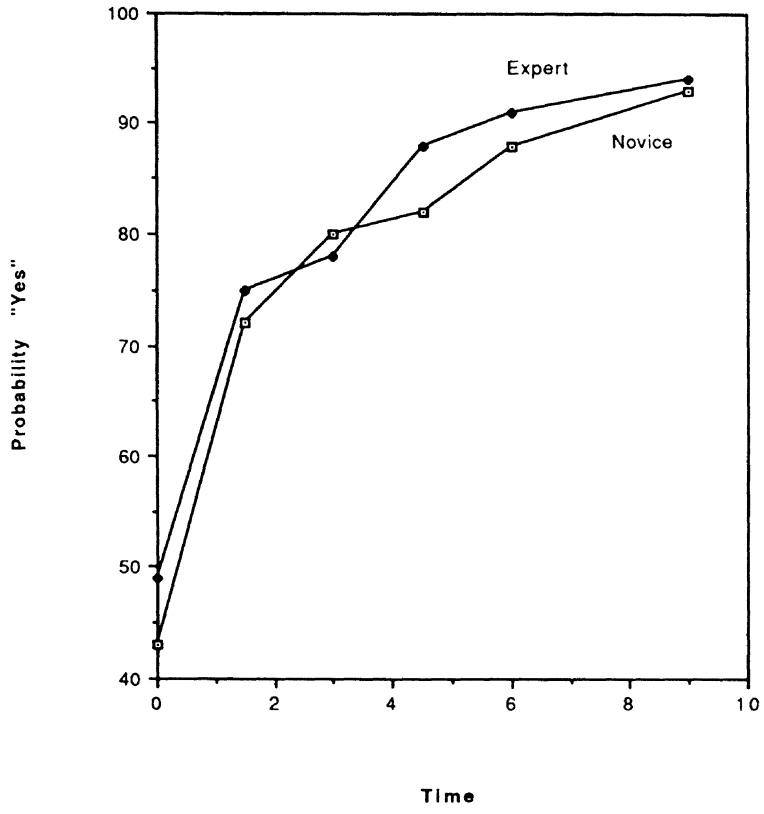

FIG. 8. Judged correctness of old, verbatim test sentences as a function of processing time for experts and novices; after Schmalhofer et al. (1990).

ence statements as true initially, but further processing only confuses them. They treat inferences much like old sentences at first, but then the frequency of their Yes responses actually decline instead of rising continually as in Fig. 8. Experts, on the other hand, show a slow but steady increase in Yes responses throughout, ending up at a higher level than novice subjects. Both groups of subjects take more time on the final, unconstrained response for inferences than for old sentences. These findings can be readily interpreted within the construction–integration model.

On-Line Integration

In previous work with the construction–integration model, the sentence was assumed to be the processing unit, purely for reasons of convenience: as long as one is not much concerned with what happens within a sentence, this is a useful simplification. However, if one is interested in how activation develops during the reading of a test sentence, the fiction of the sentence as a processing unit must be abandoned. Instead, it will be assumed here that words are the processing units. As each word is read, all elements that can be constructed at this point are constructed and added to the existing net, which is then re-integrated. Thus, each sentence contains as many processing units as it has words (or, rather, word groups, the L-units in Fig. 3).

In order to illustrate how this model works, we first simulate the construction of the original text representation. Since we are not interested in the on-line properties of this process, this is done in exactly the same way as with the Zimny data: all the appropriate L, S, P, and M units are constructed and connected according to the same principles as in Figs. 3 and 4. A function schema, with slots for "Name," "Use," "Input," and "Output," provides

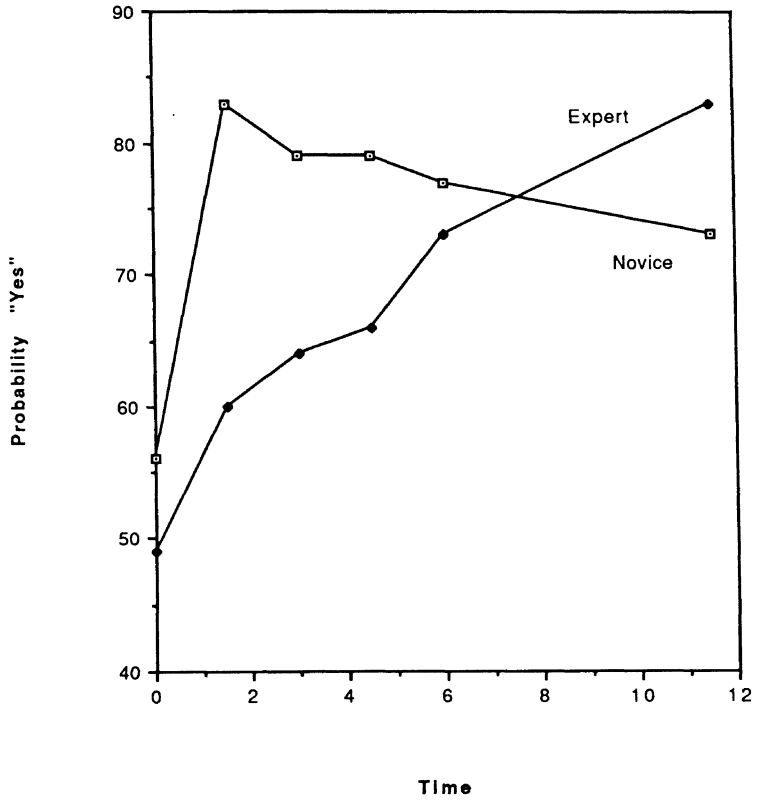

FIG. 9. Judged correctness of inferences as a function of processing time for experts and novices; after Schmalhofer et al. (1990).

the basis for the situation model. The resulting network is then integrated, and a pattern of activation is obtained which, together with the net of interrelationships itself, characterizes the memory representation formed for the to-be-remembered text.

An old, verbatim test sentence is recognized by computing the amount of activation of its elements at each input stage. Thus, the test sentence *"The function FIRST is used to extract the first S-term,"* is processed in seven input stages, as shown in Fig. 10. First, *"The function"* is processed, yielding the elements L2, P2, and M2. The second input unit comprises *"FIRST,"* that is the elements L3, S1, P3, and M3. The remaining input units and constructed linguistic, propositional and situational memory representations are also shown in Fig. 10.

Figure 11 illustrates how the model works for the inference statement *"A single*

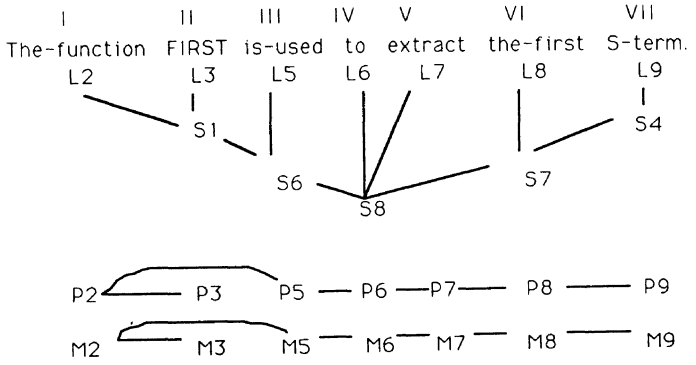

FIG. 10. An old, verbatim test sentence, processed sequentially in seven input stages.

S-term is produced by the function FIRST." Only one element is constructed in the first processing unit: the unit L20 *"a-single"* (the numbering takes account of what was already constructed in the processing of the original text). More happens next: *"S-term"* corresponds to L12, P9, and M9 of the original text. Furthermore, at this point the new S-element S18 is constructed, as well as the proposition P21, (SINGLE, S-TERM). Note that no new model element is constructed corresponding to P21, for there is no way to know where in the function-schema such an element should be placed. In the third input unit, not only the new surface element L21 is generated, but also the sentence unit S22 and the corresponding proposition P22 (PRODUCE, $, (SINGLE,S-TERM)). Both of these are at this point incomplete: we don't know as yet what produces (SINGLE, S-TERM)—the $-sign is used as a placeholder in the proposition—and we do not know all of the constituents of S20. S- and P-units are constructed as soon as possible, before all of the relevant information is available. This assumption in the present model is supported by results in the psycholinguistic literature, where it has been shown repeatedly that people assign words and phrases to plausible syntactic structures on-line, and do not wait until a complete analysis becomes possible (e.g., Frazier & Rayner, 1982).

The immediate processing strategy at the linguistic and textbase levels contrasts with a wait-and-see strategy at the situation model level. In the former case, there are powerful heuristics available that make immediate processing feasible [e.g., the Minimal Attachment strategy of Frazier & Rayner (1982), or the Referential Coherence strategy for forming a coherent textbase (Kintsch & van Dijk, 1978)]. The results may not be optimal (e.g., causal links are more useful in stories than mere referential links), or they may have to be revised eventually (as in garden-path sentences), but they yield useful approximations for on-line processing that can later be modified if necessary. Immediate processing is also used when situation model elements are encountered in a test sentence that are already available in the original memory representation of the text. In that case, it is assumed that they retain their original position in the situation model. (As all heuristics, this will sometimes be wrong, e.g. in the case of false test sentences.) Newly formed elements of the situation model, on the other hand, cannot be assigned on-line to a slot in the schema: where an element fits into a schema, or whether it doesn't fit at all or contradicts it, usually can be determined only after the whole sentence has been processed. Thus, the processing of new situation model elements is delayed until the sentence wrap-up. In Fig. 11, the elements M21 and M22, (PRODUCE, M2, (SINGLE,M9)), are therefore constructed in Input Stage 6 and assigned to the "Output" slot of the Function schema.

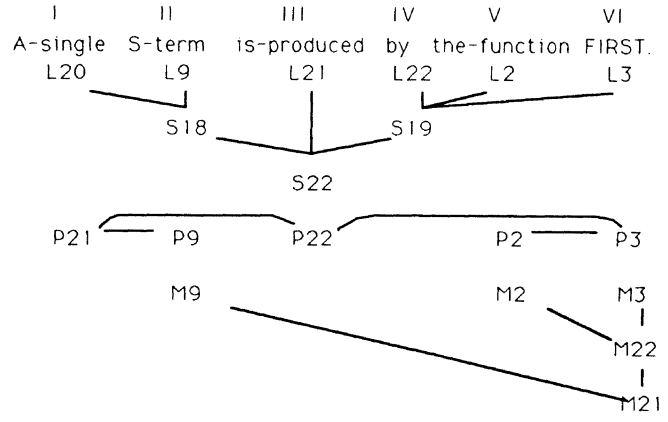

FIG. 11. A correct inference, processed in six input stages.

Fit of the Model to the Data

How well can this model account for the Schmalhofer et al. (1990) data? There are two striking features of these data: the fact that for old verbatim test sentences, the speed-accuracy trade-off functions increase steadily throughout the processing period and are essentially the same for novice and expert subjects; and the fact that experts have slowly rising, high-asymptote functions for correct inferences, while novices are characterized by fast-rising, low-asymptote functions. Statistically, we have observed an interaction between processing time and the two subject groups for inferences, but not for old sentences. The model implies both of these observations.

At this point, there are two ways to proceed. We could try to explore appropriate link values for the coherence matrix, estimate thresholds, and so on, as was done for the Zimny data, and attempt to fit the speed-accuracy data quantitatively. On the other hand, if we are satisfied with a qualitative fit only, computations could be based on the same parameters that were used in the Zimny data. This approach has some advantages in that it avoids the possibility that good fits are obtained merely because we happened to select just the right parameter combinations. There are no reasons at all why the same parameters should fit both sets of data, and good reasons why they should not (different subject groups, vastly different texts, different task demands—for superficial processing of many simple texts in one case and careful processing of much less material in the other). Nevertheless, if the model really has something to say about sentence recognition independent of the numerical values of the parameters in the Zimny simulation, one might expect that the qualitative pattern of the predictions would correspond to the main features of the new set of data. We have therefore chosen the second way to proceed. Our goal, then, cannot be to reproduce the actual pattern of the results presented in Figs. 8 and 9, but merely their essential feature: the absence of an interaction between processing time and background knowledge for old sentences, and a particular type of interaction—a more rapid initial rise in activation for low-knowledge subjects than for high-knowledge subjects—for inferences.

The difference between novice and expert subjects in the present model is that the former have only a fragmentary, partly correct situation model. Since we are only interested in qualitative predictions, the more radical assumption was made that novices have no situation model at all. Specifically, the speed-accuracy functions were simulated with the same parameter values that were used for the Immediate Group above, except that all link strengths were set to 0 in the situation model of the novices. The results are shown in Figs. 12 and 13 for old sentences and inferences. These calculations are based on the old sentence analyzed in Fig. 10 and the inference analyzed in Fig. 11.

Schmalhofer's speed-accuracy functions (Figs. 8 and 9) plot the probability of a Yes response against time. The model predictions are in terms of total activation against input stage.[6] Nevertheless, Fig. 12 manages to capture the relevant features of Fig. 8. Old, verbatim test items increase rapidly in strength and to a high level, the same for experts and novices. Inferences, on the other hand, rise faster for novices, but to a lower level, while the inference function for the experts rises more slowly initially but to a higher level (Fig. 13). Whereas for expert subjects the inferences were model-based, they were surface- and text-based for novices. This pattern of results is thus in close agreement with the Zimny data.

As in Experiment 1, the question arises how general the results shown in Figs. 12 and 13 are, since once again, we are comparing average data with predictions for single sentences. Typical as well as atypical

[6] Very similar predictions are obtained if the length of each input unit is made proportional to the number of cycles needed for the integration process to settle.

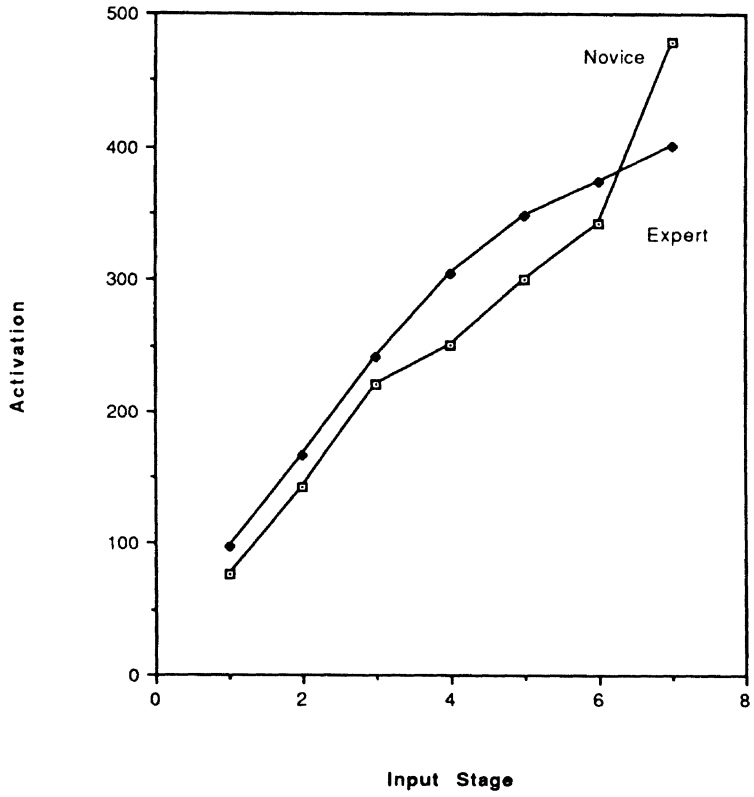

FIG. 12. Activation of an old test sentence as a function of processing time for experts and novices.

examples can be selected, just as in Experiment 1. For instance, the old sentence *"The argument of the function must be a combined S-term"* yields results much like the old sentence used in Fig. 12: steady increases for both groups throughout the whole time period, with most of the increase in the first half, though now the expert group has more activation than the novice group at all points. Similarly, the inference statement *"Combined S-terms used as arguments of a function are legal"* corresponds nicely to the results shown in Fig. 13: there is the slower initial increase in activation for the expert subjects (18% of their total in the first half, vs. 50% for the novice subjects), and the final cross-over on the last point. On the other hand, it is not hard to find atypical examples. Consider the inference *"The argument of the function may consist of five Lisptoms"*. For the first three cycles, this inference is identical with the old sentence just discussed. Hence we have an initial increase in activation for both the expert and novice subject groups. However, from the fourth cycle on, this inference behaves much like the other two examples that have been discussed here. In the first three cycles when the sentence can be differentiated from an old sentence, only 5% of the remaining activation accrues for expert subjects, vs. 57% for novice subjects. Thus, although the details of this example are much more complicated than for the other inference sentences discussed here, it shares some of the same qualitative features that characterize the recognition of inferences in the model.

Obviously, Figs. 11 and 12 are only caricatures of the corresponding data in Figs. 8 and 9: our goal was merely to show that the model can predict the interaction for the inference sentences, and the absence of such an interaction for the old sentences without any further parameter estimation. The inherent features of the model, not any particular lucky parameter estimates, are responsible for this prediction. To actually fit the model to the data, however, we would need to estimate new parameters and

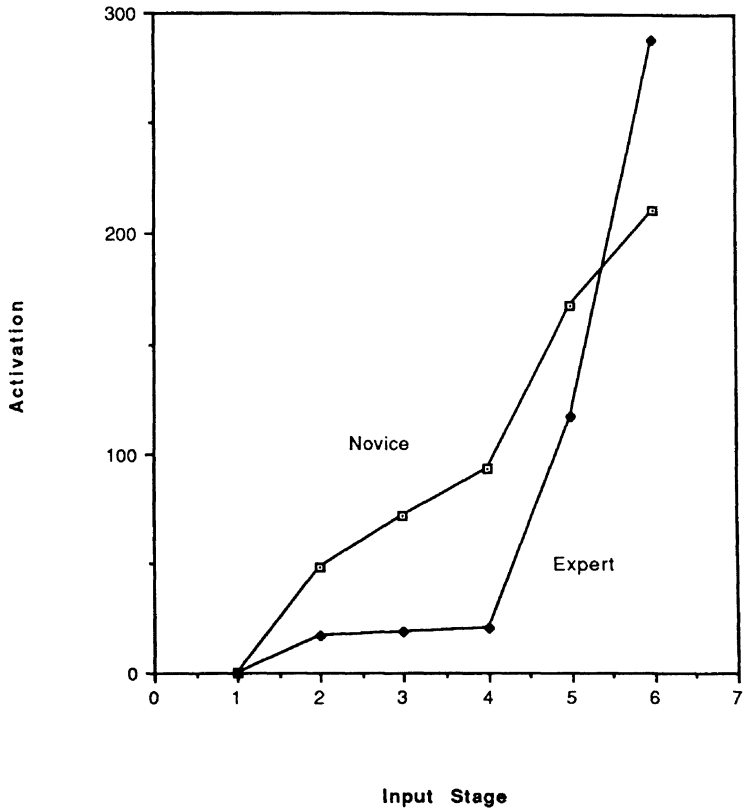

FIG. 13. Activation of a correct inference as a function of processing time for experts and novices.

to transform activation strengths into response probabilities, just as was done in Experiment 1.

CONCLUSION

A model of sentence memory from discourse has been developed and tested here which builds upon previous work on item recognition and discourse comprehension. The recognition mechanism used in this model has been derived from previous models of recognition developed to account for list learning data. Two elaborations from the domain of discourse comprehension were needed to enable this mechanism to deal with sentences from a coherent discourse, rather than with list items. First, sentences must be represented in memory at several levels of representation, each of which can contribute to a recognition or verification judgment. Second, the very processes of comprehension as formulated in the construction–integration model of Kintsch (1988) were shown also to be involved in judging whether a sentence had been experienced before as part of a discourse. Thus, familiar theoretical notions could be combined to provide an explanation for sentence recognition.

This explanation fared quite well when tested against the results of empirical investigations of sentence judgment. In Experiment 1, a good quantitative account of recognition for old sentences, paraphrases, and inferences was obtained for delays ranging from immediate tests to four days. However, due to the complexity of the comprehension model, a large number of parameters had to be estimated to match these data. Hence, we changed our strategy in Experiment 2 from one of fitting empirical results quantitatively to one of testing qualitative implications of the model which did not involved further parameter estimation. The data in question concerned the time course of sentence verification. It was shown that the model predicted major qualitative features of speed-accuracy trade-off

functions, without estimating new parameters. Thus, the model has been tested successfully against two large, complex sets of sentence memory data.

The model of sentence memory developed here is quite general and can be applied to many different texts and test sentences, with one serious restriction: in order to apply the model, one needs to know what the situation model would look like for the text and the subjects in question. Linguistic analyses as well as propositional textbases (the latter if necessary based on the default rule of argument overlap, as in the present case) can be constructed for any kind of text, but situation models are much less well understood. In particular, it is not clear how nonpropositional situation models (e.g., mental maps, as in Perrig & Kintsch, 1985) could be integrated into the present framework.

Earlier models of sentence recognition share some characteristics with the model proposed here, but differ in other respects. Two such model are the schema-pointer-plus-tag model of Graesser (1981) and the plausibility judgment model of Reder (1982). Both models, in common with an earlier generation of recognition models, conceptualize recognition as a match between the memory representation of an item and the item presented at test, thus violating a basic feature of current recognition models as discussed here. Furthermore, they are much less specific than the computational model presented here. In other respects, however, there are some communalities between these models and the present approach. Graesser distinguishes two stages of sentence recognition, one corresponding to the question "Is the item in the memory trace?," and the other to "Must the item have been in the passage?" (Graesser, 1981, p. 92). Reder similarly distinguishes between a plausibility judgment and a direct retrieval (Reder, 1982). Clearly, there are some parallels here between matches based on the surface and textbase representation on the one hand and matches based on the situation model on the other. One could, in fact, claim that what has been done here is to provide an explanation and computational mechanism for the phrase "plausibility judgment." Significant differences should not be overlooked, however. Reder, for instance, emphasizes the stage character of the process with plausibility judgments normally coming first, preempting direct matches. In the present model, matches at all three levels of the representation occur as soon as possible and in parallel, with the contribution of the situational match necessarily coming in rather late in the processing of a sentence, as the analyses of the speed-accuracy trade-off data in Experiment 2 show quite clearly. A recent report by Ratcliff and McKoon (1989) further supports these conclusions. They have shown in a sentence matching experiment that different kinds of information are available at different points during retrieval, with surface similarity playing the strongest role early in the time course of retrieval, in agreement with our conclusions.

One does not need a separate model for sentence recognition. If we put together what we know about the item-recognition process per se with the construction–integration model of discourse comprehension, we have a ready-made explanation for many of the phenomena of sentence recognition. Thus, the construction–integration model comes one step closer toward becoming a general theory of discourse comprehension and memory.

References

Bower, G. H., Black, J. B., & Turner, T. J. (1979). Scripts in memory for text. *Cognitive Psychology*, 11, 177–220.

Fletcher, C. R., & Chrysler, S. T. (in press). Surface forms, textbases, and situation models: Recognition memory for three types of textual information. *Discourse Processes*.

Frazier, L., & Rayner, K. (1982). Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences. *Cognitive Psychology*, **14**, 178–210.

GALAMBOS, J. A. (1982). *Normative studies of six characteristics of our knowledge of common activities*. Cognitive Science Technical Report No. 14. New Haven, CT: Yale University.

GILLUND, G., & SHIFFRIN, R. M. (1984). A retrieval model for both recognition and recall. *Psychological Review*, **91**, 1–67.

GRAESSER, A. C. (1981). *Prose comprehension beyond the word*. New York: Springer.

HINTZMAN, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, **95**, 528–551.

JOHNSON-LAIRD, P. N. (1983). *Mental models*. Cambridge, MA: Harvard University Press.

KINTSCH, W. (1974). *The representation of meaning in memory*. Hillsdale, NJ: Erlbaum.

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

KINTSCH, W., & ERICSSON, A. (in press). Die kognitive Funktion des Gedächtnisses. In K. H. Stapf & D. Albert (Eds.), *Psychologische Modelle der menschlichen Gedächtnistätigkeit*.

KINTSCH, W., & VAN DIJK, T. A. (1978). Towards a model of text comprehension and production. *Psychological Review*, **85**, 363–394.

MURDOCK, B. B., JR. (1982). A theory for the storage and retrieval of items and associative information. *Psychological Review*, **89**, 609–626.

PERRIG, W., & KINTSCH, W. (1985). Propositional and situational representations of text. *Journal of Memory and Language*, **24**, 503–518.

RATCLIFF, R., & MCKOON, G. (1989). Similarity information versus relational information: Differences in the time course of retrieval. *Cognitive Psychology*, **21**, 139–155.

REDER, L. M. (1982). Plausibility judgments versus fact retrieval: Alternative strategies for sentence verification. *Psychological Review*, **89**, 250–280.

SCHANK, R. C., & ABELSON, R. P. (1977). *Scripts, plans, goals, and understanding*. Hillsdale, NJ: Erlbaum.

SCHMALHOFER, F. (1986). Verlaufscharakteristiken des Informationsabruf beim Wiedererkennen und Verifizieren von Sätzen. *Zeitschrift für Experimentelle und Angewandte Psychologie*, **33**, 133–149.

SCHMALHOFER, F., BOSCHERT, & KÜHN. (1990). *Text- and situation-based learning*. Manuscript in preparation.

SCHMALHOFER, F., & GLAVANOV, D. (1986). Three components of understanding a programmer's manual: Verbatim, propositional, and situational representations. *Journal of Memory and Language*, **25**, 279–294.

VAN DIJK, T. A., & KINTSCH, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

WELSCH, D. (1989). *A connectionist approach to recognition memory for sentences*. Unpublished master's thesis, University of Colorado, Boulder, CO.

ZIMNY, S. T. (1987). *Recognition memory for sentences from a discourse*. Unpublished doctoral dissertation, University of Colorado, Boulder, CO.

(Received: May 27, 1989)
(Revision received: July 6, 1989)