INSTITUTE OF
COGNITIVE
SCIENCE

Psychological Processes in Discourse Production

Walter Kintsch
Department of Psychology
University of Colorado

Technical Report No. 99

Institute of Cognitive Science
University of Colorado
Boulder, Colorado 80309

October 1980

Paper presented at the workshop on "Psycholinguistic Models of Production", University of Kassel, West Germany, July 14-16, 1980. Preparation of this report was facilitated by Grant MH15872 from the National Institute of Mental Health.

Psychological processes in discourse production

Walter Kintsch

University of Colorado

Paper presented at the Workshop on "Psycholinguistic Models of Production", University of Kassel, West Germany, July 14-16, 1980. Preparation of this report was facilitated by Grant MH15872 from the National Institute of Mental Health.

Psychology has recently rediscovered its interest in reading comprehension, partly in response to a national malaise: the literacy problems in our schools. The reading problem has been joined by an equally disturbing writing problem. The development of a psychology of writing, of writing problems, and of teaching writing has acquired a certain urgency. Others, of course, have been concerned with these issues for a long time, and psychology is clearly a latecomer in this area. The question is, what does it have to offer. There are no concrete achievements to point to yet, which is hardly surprising given the lack of work on writing until very recently; but there is promise.

Since classical antiquity there have been treatises on good writing, on style, and on teaching writing. What distinguishes the psychological approach to writing from these continuing efforts? Not more than a small, but important shift in perspective: from a concern with text structure and the writer's abilities, to the actual process of discourse production. Psychologists attempt to model writing from what they know about the basic cognitive operations, explicitly taking into account human processing limitations. It is precisely through knowing what these limitations are and how they affect the outcome of the writing process that psychological theories attain predictive power, and, we hope, will open up new ways of looking at writing and writing problems.

It seems desirable to place some general constraints on the development of a psychological theory of writing. First, such a

theory should be closely related to a corresponding theory of comprehension. This is not to claim anything as primitive as that writing is merely comprehension in reverse; obviously, there are significant task differences, but the two behaviors are complementary, and it would be pointless to try to understand one without the other. Comprehension is a component of writing, because the writer in editing and monitoring his output must comprehend what he writes. In this essay I propose to investigate writing processes against the background of the theory of text comprehension of Kintsch & van Dijk (1978).

The second constraint is even more important. An enormous, ancient body of knowledge already exists on writing; a psychological theory will have to deal with that, somehow. Let us not reinvent the wheel. The task is to make use of rhetoric (journalism, copywriting, advertising) - not to neglect it, or to disguise well known facts as new discoveries. A good theory is one that assures that psychology and related disciplines can take advantage of each other. They should not be independent of each other, nor should one subsume the other. There is, moreover, another body of scientific knowledge that is relevant for a theory of writing: the laboratory work on memory and cognitive processes. Many psychologists, when they start working on complex processes, disregard the laboratory results and the theories based on them as too simpleminded, and insist on starting from scratch once again. Here, I want to propose the opposite strategy: to build a theory

of the writing process cumulatively, building both on the list-learning paradigms of the psychological laboratory and on the analytic and descriptive tradition of rhetorics.

The Flower and Hayes model. Indeed, there is something else to build upon: the recent work on writing within the cognitive science framework. There is not much as yet: an interesting program to write stories by computer in the Yale tradition (Meehan, 1976), a review by Black (1979), and a few attempts at psychological modeling: Bruce, Collins, Rubin, and Gentner (1978) and Flower & Hayes (1979), who provide a useful foundation for the present model; and experimental work by Voss, Spilich & Vesonder (1980), who are concerned more with one of the subprocesses of text production than with a general model of writing, as will be discussed below.

Flower & Hayes provide a task analysis of the writing process (as such this is not very different from Bruce et al.), and substantiate it via protocol analysis. They are able to show that the processing stages they discuss have distinct psychological characteristics, and they develop a methodology that can be used effectively in the exploration of the writing process (especially Atlas, 1979). I propose to use Flower & Hayes' basic analysis of writing into subprocesses, and conceive of my task as modelling (some of) these subprocesses in sufficient detail so that they become more than mere names for unanalyzed global processing units.

Flower & Hayes distinguish three major subprocesses of writing: planning, translating, and editing. The planning process is subdivided into three components: generating ideas, organizing, and

goal setting (the latter is a control process that decides when to schedule generating, organizing, translating, or reviewing). Two components of the editing process are also distinguished: an automatic editing that interrupts other ongoing activity, and a more systematic reviewing process. They show that these subprocesses can be distinguished in various ways (e.g. by means of the written products of each, notes from the generation process, outlines from organization, well-formed sentences from the translation process). But they do not go into sufficient detail about how these subprocesses work.

I am going to concentrate here on the subprocesses of generating and organizing. I cannot at this point present detailed formal models, but merely a program for how such models can be developed within the framework of the Kintsch and van Dijk theory of comprehension. I shall say very little about translating, indicating only how these processes would fit into the general framework.

A note is in order as to the domain of writing processes to which the present model will be applied. The model is general, but for a start I would like to neglect the truly creative aspects of writing: writing something has never been said or thought of in quite that way before by the writer, or even by anyone else. I want to come back to that, but first consider the kind of writing that is for the most part reformulation of familiar ideas - writing a textbook, an essay or an editorial, the college student writing a theme on "Religion" or a story on "A baseball game".

A. On generating ideas

Flower & Hayes have very little to say: the generating process derives its first memory probe from information about the topic and the intended audience; as items are retrieved, associative chains are formed, which are broken whenever something not useful to the writing task is retrieved.

What does a writer have to go on, beyond the empty page staring into his face, when he starts writing? First of all, a topic that he wants to write on. This topic may be very vaguely defined, e.g. to write on "Religion", or it may be quite specific, e.g. "Discuss Marx' statement that Religion is the opium of the people". These two topics differ in the amount of constraint they impose upon the idea generation process. Often, a more constrained generation episode may follow a less constrained one, as in the case where the general topic "religion" is successively refined in writing an essay, but transitions from very specific to quite general topics also occur. The topic only partly controls the idea generation process. There are other important constraints: the type of text being generated (a story for the New Yorker, a textbook chapter, or an editorial?), and closely connected with that, the readers for whom the text is being written. Note that the writer is trying to communicate with his audience, however indirectly, and that this process is therefore regulated by the Gricean conventions that govern all human communication (Grice, 1975): to say what is relevant, not to say too much, or too little,

to say it so that it can be understood. Indeed, because of the remote relation between reader and writer, a writer must observe special, more stringent conventions than a speaker who has the advantage of personal contact (e.g. Hirsch, 1977; Rubin, 1979). Thus, we have two kinds of constraints to consider: the topic, and conventions about text type and audience. What they are constraining is a search for ideas in the writer's knowledge base.

We have to make some assumptions, therefore, about the nature of that knowledge base. Knowledge is part of a person's long-term memory. Various models of LTM have been proposed, ranging from almost random, minimally organized (Landauer, 1975) to intricately structured (in terms of conceptual hierarchies, scripts, frames, schemata). However, there may be disadvantages to conceiving of scripts and schemata as LTM structures; instead, one may think of them as being constructed when needed for some purpose, in a particular context. This construction process makes use of information in LTM, of course, but the information is not in itself tightly organized and is structured only on demand, for the purposes of a particular task. The relevant relations among units of information must be stored in LTM, but there are always multiple relations and hence many potential structures; when we construct a frame from this rich set of possibilities, we pick one particular pattern of stored LTM relations, suitable for our task. Imagine a huge road map of the U.S. as a two-dimensional analogue of LTM: all paths are potentially there, but if we want to go from Denver

to San Francisco via Atlanta, a particular path lights up, as in a subway map. In that sense, frames are constructed from LTM, if and when they are needed. There is not only one fixed frame, but in complex situations, a number of alternatives; flexibility is gained from mere rote elements.

In comprehending discourse, for instance, one does not pull a frame out of LTM, ready made, and then plug text propositions into the slots of that frame, but one constructs a frame suitable for organizing a particular text, using relevant knowledge. Script application, in the sense of Schank & Abelson (1977), is a special case where the relevant knowledge (about a procedure) is fairly simple and leaves few options (there is only so much one can do in a restaurant under normal conditions); but consider the reader who reads an essay on "Religion": there is no frame to be pulled out to guide his writing, but there is much loosely organized knowledge about religion that can be used to construct suitable frames ("facts" in the Kintsch & van Dijk, 1978, terminology) in interaction with the text.

We conceive of LTM as a huge propositional network that is organized only in the sense that certain relationships exist among the propositional nodes (e.g. they may have common arguments), and in the sense that related propositions are somehow close to each other. A spatial metaphor is useful here. The memory network is like the universe. In a multi-dimensional space, there are galaxies of knowledge on this or that, with subclusters like star

systems. There is a lot of local structure there, as well as some very specific relations within a knowledge cluster, but if observed from enough distance, all we have is a locally uniform space, dense in some regions, full of holes in others. Holes may be filled when new knowledge is acquired through learning, and where someone may have nothing but holes in his knowledge system, others have grown a thicket of nodes. But in some unspecified way these nodes are grown so that related nodes are deposited in neighboring regions of the space. (Landauer, 1975, has described a mechanism that assures that contemporaneous experiences are stored more or less together). Thus, the knowledge space contains an area where information can be found that has to do with "religion" (including personal, spatio-temporally tagged episodic nodes), and, at some distance from it, a space for "Quantum Theory" a vast void for most people. How can information regarding some more or less well specified topic be retrieved from such a knowledge base?

Obviously, people do not have a procedure to search their knowledge spaces systematically. Otherwise, given enough search, they would always find any piece of information that is actually stored. What the exact processing limitations are that make such a procedure impossible is unknown.

How people actually go about retrieving information from memory has been studied quite extensively in the context of list learning tasks. Most memory specialists today subscribe to some

version of a generate-edit theory. The nature of the retrieval cue is crucial for the generating process (Tulving, 1979); extensive use is made of inter-item associations in generating (Anderson & Bower, 1972); through a process of pattern completion the retrieval cue recovers the to-be-recalled item (Kintsch, 1974): the editing process which weeds out inappropriately generated items can be highly complex and task specific (Mandler, 1979). A model of recall in a list learning task which embodies all these features has recently been proposed by Raaijmakers (1979). It is the best and most complete model of recall available today, both because it has been worked out formally and in detail, and because it embodies the important principles mentioned above which have emerged from 20 years of study of this experimental paradigm. I shall investigate the applicability of the RS model to a new task: the task of retrieving information from memory within certain constraints, i.e. the process of generating ideas in writing.[1]

There are two problems. One must decide where and how to look for an idea in the vast knowledge space, and then actually produce it. For the second problem - the actual memory search - the RS theory provides a ready-made solution, as I shall show below, if we can handle the first one.

<u>On the construction of a retrieval cue and the establishment of a search set</u>. The question is, where to look for information in memory, or, less metaphorically, what questions to ask and in what order to explore them.

LTM is to be probed with a retrieval cue. This cue has several components. First, there are the contextual features, paramount in list-learning tasks. They are less important now, but can not be neglected (the familiar writing environment, versus the idea that finally opos up in the shower; today's headache). Next, there are the constraints imposed by the intended test type and audience. Certain searches can be excluded on the basis of such considerations, without actually having to examine their outcomes. Finally, and most importantly, there are the content specifications of the to-be-generated ideas. These may be extremely general, as in the "religion" example, in which case all sorts of ideas will be generated that will be difficult to organize, or they may be very specific, as in the "opium of the people" example, which will generate fewer ideas, but in such a way that they will be easier to organize.

Thus, we have a retrieval cue and could go on searching for ideas with that cue. But a whole new set of problems emerge at this point, for no text is written on the basis of one generation episode. Many memory searches are needed, and the question arises how these are controlled. What are the search strategies writers use? Are some better than others, and why?

How do people construct an outline for what they are going to write on (a search plan, not an outline for the final product, which would be a macrostructure). The principle is clear: consider some proposition A that is currently in the writer's short-term

store (it would be the topic at the beginning of the process). It has certain relations to other propositions in the knowledge structure, say V, W, X, Y, and Z. A search plan is obtained by deciding to explore X next, or X, Y and Z in that order. What governs such decisions?

There are many questions and almost no answers. The additional constraints mentioned above concerning text type and audience certainly play a role; e.g. they might exclude alternatives V and W right away as inappropriate. But beyond that, we know very little. Each note in a search tree represents a decision to continue the search in a certain way in the face of many alternatives. When do people explore a search tree in depth, when do they make a breadth-first search? Some pathways seem obligatory (e.g. the pro and contra in the delights of opium). How does the level of constraint interact with the choice to follow up the search in a particular way, (e.g. if X is a well-constrained path, and Y is vague). Most importantly, what effects do human processing limitations have on their search strategies? Are short-term memory limitations a factor, and if so, how are they overcome through the use of external memory devices?

There seem to be at least two promising ways to investigate such questions. First, through the observation of actual human search processes in writing, using protocol analysis or experimental procedures. For instance, Polson (1980) has described search behavior in a problem solving task as depth-first to a

certain level, followed by an exploration of other branches to the same level, with repeated reiterations. Polson speculates that this search pattern reflects an adaptation to short-term memory limitations: a depth-first strategy makes the least demands on short-term memory capacity, but elaborating a path too far has its dangers, because possible constraints from other branches are neglected in the process. To what extent is such a search strategy specific to the highly constrained problem-solving task used by Polson? As a second example, consider a study by Atlas (1979), who had subjects generate ideas to answer a letter. Expert writers exhibited considerable planning in this task, in contrast to novices. Clearly, such methods could be used to investigate empirically the kind of search strategies writers employ in the idea generation process.

But the empirical investigation must be supplemented theoretically. Theories of information retrieval and search processes provide a starting point (e.g. Winston, 1977). They are not psychological theories, however, because they fail to consider human capacity limitations and the nature of the human knowledge base. Nevertheless, an exploration of the formal properties of search processes may be instructive, and eventually lead to a psychological theory of memory search strategies. How is the search to be guided, by means-ends analysis or by productions? If so, how are the productions themselves to be controlled (by recency, importance?) Do people make use (or should they?) of

the heuristics that computer scientists employ to make searches more efficient (disaster cut-off, futility cut-off, feed-over)? Given that the knowledge base is infinite for all practical purposes, how is it possible for depth-first searches to be used successfully? How do short-term memory limitations affect breadth-first strategies? At present, there is no model that would permit us to answer such questions.

Another problem concerning search strategies is suggested by an observation made about comprehension strategies (Miller & Kintsch, 1980): when people have a choice, they tend to select the largest possible unit to organize a text. That is, if there is a suitable frame, that is used; but if no frame fits the text, they still organize the text, but in terms of smaller, propositional units. Is there a corresponding phenomenon in the idea generation process? Is it the case that, at each stage in the process, people favor the most general retrieval probe?

But let me return from the unexplored work of psychological search strategies, setting this problem aside for future research. Suppose that we have our retrieval cue with which to proble the memory system and that it is constituted as it was hypothesized at the beginning of this section: some contextual cue, information about text type and audience, and some specification of the content of the to-be-retrieved memory node. For the most part, that is, except for the context cues, the constitution of the retrieval cue is under the writer's control. From now on the idea generation

process is automatic, however. Two things happen: the retrieval cue defines a search set, and ideas are then generated from that set.

The following assumptions are made about establishing a search set. The retrieval cue is directly related to one or more propositional nodes in LTM. These nodes are automatically contacted. Their number depends on the specificity of the information in the retrieval cue: there may be only vague and general constraints (once more, the "religion" example provides an illustration in which case many nodes with the argument RELIGION will be activated), or there may be a highly specific constraints (the "opium of the people" topic would activate only a small subset of the nodes from the previous example). Each activated node has a neighborhood of some fixed size. The union of the neighborhoods of all activated nodes constitutes the search set.

The number of nodes in the search set thus depends on two factors: the number of nodes contacted by the retrieval cue, which produces a set of possibly overlapping semantic neighborhoods; and secondly the density of the knowledge space which determines how many nodes are actually found in these semantic neighborhoods. Thus, the level of constraint imposed by the retrieval cue and the density of the knowledge space in the area of the search jointly determine how many nodes will be included in the search set. The tighter the constraints, the fewer nodes will have to be searched; the denser the nodes in the knowledge space, the more there are to be searched. The important point is that the size of the

search set is outside the subject's direct control: given a retrieval cue and a knowledge base the search set is constituted automatically.

How the search proceeds and what sort of ideas are recovered. The search itself is automatic, too, and follows the principles of the episodic memory retrieval model of Raaijmaker (1979). The only difference is that we are now looking for information that is defined in a different way. In a list-learning task, the goal is to recover words that are associated with a particular kind of context tag; here, the search is for information that is related to the retrieval probe in terms of content and other criteria (text type and audience).

In Raaijmaker, the items of the study list constitute the search set. They are associated among each other and to a context cue. Here, the search set is defined by the semantic neighborhoods contacted by the retrieval cue; again the items in the search set are interassociated, but they are not normally associated with the present context cue. If the search set contains n-1 memory nodes, plus the retrieval cue, we obtain an inter-association matrix $\{s_{ij}\}$ of size n x n. The search, according to Raaijmaker, has two components. First an item is selected, which occurs with a probability proportional to its relative strength of association to the retrieval cue. A sampled item is not necessarily recoverable, however, because not enough information may be stored at a particular memory node to make recovery possible. In other words, the

pattern completion process may fail. The probability that it succeeds depends in the model on the absolute strength of the association between the item and the retrieval cue.

If an item is not successfully recovered, the search continues until a certain number of failures to recover anything have occurred (This number is probably quite small - see Flower & Hayes). At that point, the generation episode is terminated. If a recovery attempt is successful, the search continues, but two changes are made: first of all, the recovered item is now added to the retrieval cue, so that the new search is doubly constrained, by the original probe as well as the already generated information; secondly, the association between the sampled item and the retrieval cue is incremented by some amount $\underline{\theta}$. Thus, as more memory nodes are recovered, the nature of the retrieval cue keeps changing, in that it is always redirected somewhat by the last item. At the same time, however, the whole associative structure is dynamic, too: the items already once sampled grow stronger and are resampled with an increased likelihood, which is equivalent to decreasing the sampling probabilities for the ideas that the writer has not yet thought of.

The Raaijmaker model is mathematically formulated, in detail. There is no point in reproducing here this formulation. We can simply accept it wholesale with only minor modifications. One of them concerns the editing process which recovered ideas undergo before they are actually produced. A decision model based on signal-detection theory seems appropriate here: the suitability

of each recovered idea is evaluated against some criterion derived from the content as well as the type- and audience-specifications of the original memory probe.

But so what? What can this model do for us? Can it be tested experimentally, as the Raaijmaker original can be tested with list-learning data? That depends on what kinds of data we can obtain. But even now, the model explains qualitatively a number of interesting phenomena about writing. Start with interference: one tries to think of something, but the mind keeps going down the same well-travelled track, and it is on the beach or in the shower (significant individual differences here!) that the crucial idea finally comes. Increasing the associative strength between sampled items and the retrieval cue accounts for this output interference, and changes in the contextual components of the probe for the shower-effect. Retroactive and proactive interference occur because certain associations are formed that are maladaptive for a task - they lead the search process astray.

A large search set (not enough constrained by the probe) will decrease the probability of sampling an item in the set, while a small, highly constrained set will improve it, for the same reasons that allow Raaijmaker to predict the list length effect in recall. There are many non-intuitive questions that can be explored with this model, e.g. concerning the interaction between level of constraint and density of the knowledge space (and hence learning effects). Furthermore, the strength and distribution

of inter-item associations is another factor that interacts with the previous ones in ways that simply can not be computed without a precise model.

It may also be useful to explore the temporal characteristics of the idea generation process, taking up the Bousfield & Sedgwick (1944) tradition in a new context. We can make, for instance, predictions about the rate of idea generation in relation to the number of ideas generated.

Are we going to find an analogue in writing to the part-list cueing effect in recall?

Finally, and most importantly, how does the search process interact with search planning? Can one identify which strategies are good under which conditions (knowledge density, proble specificity)?

I would claim, therefore, that the model seems fairly rich and promising, worth working out in detail. Note that it is a very global model. It abstracts from the content detail of semantic memory models ("A robin is a bird") in favor of global, crude statistics like the density of the knowledge space. It is looking at retrieval from afar!

B. Organization

Again, Flower and Hayes' characterization of this stage is inadequate: they describe five elementary operators - identify first or last topic; order with respect to previously noted topic;

find subordinate topic; find superordinate topic; identify category. Organizing, it is claimed, is done by these operators. Clearly, there is much more to organizing a text than that.

The question is, how ideas once generated after a search plan, are being organized. The search plan, used to generate ideas, need not be the same as the final organization of the text. In general, there will be reorganizations, as new relationships among ideas are discovered.

The top levels of the final organization eventually becomes the macrostructure of the text, while the lower levels form the microstructure. The output of the organization process, therefore, is the textstructure, ready to be put into words by the translation process. Of course, no strictly sequential stages are implied by the model: idea generation, organization, and translation may be scheduled in many different ways. At one extreme, all ideas might be generated first, then organized, and then translated into words. At the other extreme, a single idea may be generated at a time, related to whatever is already written, and expressed in words. What sort of process scheduling writers use under what conditions is pretty much an open question at this point.

What are the principles by means of which a set of ideas can be organized? There are two sources that can provide the necessary constraints to organize a text. They are not independent, but it is quite useful to distinguish them. Both content and form may

provide a basis for organizing a text. The very nature of the ideas that are to be organized may determine the organization. Ideas are always interrelated in some way, directly or indirectly, in the long-term memory structure. Sometimes these interrelationships are strong, direct, and quite unique. This is, for instance, the case in the memory systems called scripts. If a script applies, it is usually quite clear what is to follow what; indeed there are not many choices (disregarding literary techniques). There is only one straightforward organization for a given content. An example of this type of organizational strategy has been described by Voss et al. (1980). They asked subjects to generate an account of a half-inning of a fictitious baseball game. While different subjects generate quite different ideas in this task, the organization of the ideas is pretty much determined. Given the constraints of the task, subjects must start out with a description of the setting of the game (which teams are playing, who is at bat, what inning, what score, plus some optional material like names of the pitcher and first batter). Thereafter, the description of the game consists of a sequence of game states, rigidly ordered, and of actions which provide the transitions between these states. These actions can be described in considerable detail (see Spilich et al., 1979), and as to their level of specificity and importance. But the main point here is that, given the states and the actions, this content fully determines a unique organization.

Only in rather extreme cases is the content so all-powerful as in the example above, however. More frequently, content admits to many alternative organizations. Fortunately, the writer has available some additional, formal -rhetorical strategies to organize his material. Consider as a second example the essay on "Religion is the opium of the people". Suppose a writer has generated a number of ideas on this topic by the processes described in the previous section and now sets out to organize them. (Again, not all generation has to be done before organization, and it is quite possible that organizational strategies may also influence the search strategies in the generation process). The organizing task here is much more difficult than in the case of the essay on "Baseball". In the baseball case, the total number of different ideas that can be generated is not huge. After all, only so many things can happen in a game. Furthermore, their interrelationships are highly constrained, either by causal or temporal relations. In the 'Religion' essay, on the other hand, the number of different ideas that can be generated is much larger, and the relationships among the ideas are diverse and complex. Rhetorical strategies are needed to impose a structure on this material. Suppose, for example, that a writer decides on the over-all structure of an argument. We assume that he has internalized a series of strategies for organizing an argument. What these strategies are, we can take from rhetoric books. Thus, for instance, following Aristotle, a writer would immediately organize

his material into three major subsections: statement of the issues, assembly of evidence, and conclusion. The first question would be, what is the proposition to be argued? In the present case, this is given. Is the proposition clear, and if not what are the issues involved? What is the common ground? It does not matter that different writers, depending on the nature of their knowledge base and their attitudes, might identify very different issues. The only important consideration here is that in every case, this structure provides them with a way or organize their ideas. The writer can then go on, asking what are the facts in the case, how are they related to the issues, how are they to be evaluated. In this way, an over-all frame for the to be generated text is constructed: the proposition is elaborated into issues, facts are related in orderly ways to the issues. Finally, the question becomes, what does the evidence mean in relation to the argument? Again, various ideas can be assigned a function in the over-all structure, or, if that is not possible, they may be rejected at this stage of the process. The writer has some more choices at this point: to employ persuasion as part of his argument, or to check his argument for fallacies. In either case, he can take recourse to further strategies to achieve these subgoals. Thus, although the content relations were too complex to provide a ready-made structure for this essay, the rhetorical relations can do the job.

How seriously such a model deserves to be taken depends mostly on how well it is possible to specify the rhetorical

strategies that writers use to organize a text. This ought to be a feasible task, since we have a long tradition in rhetorics to go on. However, there are many questions. Have rhetoriticians really considered everything that is relevant to a psychology or writing? What is an appropriate formalism for expressing these strategies? How is their use controlled? I can do no more here than provide a few suggestions.

I am concerned mostly with expository texts and arguments, less with non-technical description and narratives, though similar principles apply there. That is, what sort of rhetorical strategies are available to help organizing an essay, an editorial, a theme, or a chapter in a textbook?

Strategies may be applied recursively. A writer may decide on one over-all form of the text, e.g. that of an argument. But as this form is elaborated, other structures may be embedded: terms are defined or illustrated, classifications are made, contrasts or similarities are established, the components or the functioning of some complex system are analyzed; definitions may occur within definitions, and contrasts within contrasts. Thus, the eventual complexity of the essay is in no way predetermined, and its structure is flexible and changeable. Strategies must be highly redundant in order to be applied successfully, but they need not be unambiguous, logical, or mutually exclusive. The formalism that is probably best suited for expressing such strategies is a production system. I shall indicate here what

some of the rhetorical productions might be for organizing expository text, without concern for precise formulations, nor for the control structure that would be necessary to make such a system function. Basically, what I am doing is to formulate as productions some rhetorical rules taken from Brooks & Warren (1949) and Pitkin (1977).

Let X,Y,.. be concepts, propositions, or strings of propositions. On the left side of the productions are stated some conditions; when they apply, the actions on the right side are taken. The actions always consist of organizing the text propositions, that is, of constructing a frame within which these propositions can be embedded:

1. Y identifies (i.e. specifies in space/time) X --→ term X + identification Y
2. Y illustrates (i.e. is prototype of) X --→ term X + illustration Y
3. Y defines X --→ term X + genus V + differentiae W
   compare X,Y
   Contrast X,Y
   ...
4. $X_1, X_2, \ldots \varepsilon X, \bigcup X_i = X, \bigcap X_i = \emptyset$ --→ classify X into X_is
5. X is similar to Y in terms of Z and Z is significant --→ compare X,Y
6. X is dissimilar to Y in terms of Z and Z is significant --→ contrast X,Y
7. X and/or Y --→ coordinate X,Y

8. X is physical whole with parts Y --→ list parts and relations among parts
9. X is process or operation with stages Y --→ list stages and relations among stages
10. X is physical cause or logical consequence of Y --→ cause + consequence
11. X is a complex with components Y --→ list components and interrelations
12. X intensifies Y --→ X + Y
13. X qualifies Y --→ X + Y
14. X specifies time of Y --→ X + Y
15. X specifies place of Y --→ X + Y
16. X specifies manner of Y --→ X + Y
17. X specifies significance of Y --→ X + Y
18. X evaluates Y --→ X + Y
19. X is a problem to be solved by solution Y --→ (to be expanded analogously to the definition example)
20. X is a question answered by Y --→ (to be expanded analogously).

Strategies 19 and 20 are not worked out here in detail, because just as in the case of definitions, any of the other strategies may be used to solve a problem or answer questions.

Consider how these strategies might be used to organize a text. Let us try a minitext with only two ideas: "it is springtime - the transients gather in Central Park". Strategy (14) applies, and we obtain the organization

{ it is springtime
{ the transients gather in Central Park

The case is of course trivial: with only two ideas, there is surely no need for an organizational strategy. But suppose that instead of a single time proposition, a long string of propositions had been generated describing spring in Boulder, the snow in the mountains, the warm sun below, and then equally many propositions about the gathering of the transients in the park. Now Strategy (14) comes in handy. It provides a framework with two slots to organize the text. Within each slot, propositions may be further organized by means of other strategies that might apply, but if the text is short, this is not necessary: we can get by with relatively crude means here, e.g. relating the propositions in each slot only via argument repetition. Thus, we obtain a text representation structured in terms of a frame (with possible embeddings) and with propositions related by argument repetition in the slots of the frame. And there is something more that Strategy (14) does for the writer in our example: it tells him how to express in words the relation between the time slot and the assertion slot in his frame - a topic to be discussed below under "Translation".

The list of 20 strategies presented here is no more than a suggestion for how this problem can be approached. It is inadequately formalized, probably incomplete, and certainly not detailed enough. For instance, at least two kinds of contrast need to be distinguished:

positive-negative (not x, but y) and concession - assertion (x, yet y). These distinctions need to be made because they are relevant for the translation process: different verbal expressions must be used in the two cases.

At the beginning of this section, two principles for organizing ideas were described: exploiting the constraints inherent in the ideas, and the use of rhetorical strategies. Although this distinction was useful, it is, strictly speaking, superfluous, for it is obvious now that the content-dependent organization is just a special case of rhetorical structure. Consider Voss' "Baseball" example. Strategies 10 (causal analysis) and 14 (temporal sequence) are sufficient to organize the ideas generated here, for actions in the game cause certain game states, which are followed by other actions, and so on, until the final goal state, the end of the half-inning, is reached. What is so special here is merely the relational impoverishment of the idea set - causal and temporal relations are just about the only ones possible, if one complies with the experimental instructions to describe the game action. Thus, although it is certainly convenient to talk about a baseball script, the notion is really superfluous: the script can be generated when we need it from a flexible, loosely organized knowledge base.

C. Translation

Translation, the third major component of the writing process, could and deserves to be treated in at least as much detail as

Generation and Organization. Again, we have a large body of knowledge to draw upon: while we relied on memory models and rhetoric before, we now can make use of linguistics and psycholinguistics (e.g. MacWhinney, 1980). I shall not do that in this paper, however, and restrict myself to some rather general observations.

We assume that at this point the writer has available both the macro- and microstructure of the text, that is, its complete semantic representation. The microstructure consists of propositions, interrelated by various semantic relations which we approximate by argument repetition, and further organized into frame units. These filled-in frames are called 'facts' in Kintsch & van Dijk. The macrostructure consists of the most relevant and significant facts, hierarchically organized. This is what the writer needs to put into words now.

Several problems must be distinguished. There is first of all the task of signalling to the reader which portions of the text are macrorelevant, and what sort of a macrostructure coincides best with the writer's intentions. The strategies by means of which the comprehender tries to reconstitute the writer's organization, or constructs his own, are discussed in van Dijk & Kintsch. Presumably, they correspond very closely to the production strategies of the writer.

Next, there is a whole set of problems concerned with translating the microstructure of the text into words. Translation strategies at the paragraph, sentence, and word level need to be

distinguished. A great deal of linguistic and psycholinguistic work is relevant here (e.g. considerations of style). The rhetorical organization strategies described above also have translation components associated with them. For instance, Strategy 12, which distinguishes between an assertion and its intensification, has the following conventional expressions associated with it:

x - indeed y, x - in fact y, x - even y, not merely x - but y, x - if not y, and in negative expressions not x - let alone y, x - to say nothing of y, x - not to mention y (Pitkin, 1977). Similar lists of connectives can be compiled for the other strategies. Thus, organization constrains translation.

Recent work by Kozminski (1980) illustrates how specific models of the translation process can be developed. Kozminski's problem is to organize and connect six sentences, each dealing with a significant aspect of a stockmarket report (sales, dividends, etc). Thus, the ideas are given here and already translated into words, but they need to be put into a reasonable order and linked by sentence connectives. Kozminski solves the problem by specifying the relevant knowledge base and organizational strategies. The knowledge base reflects the stockmarket analyst's assumptions about the causal relations and correlations among the six market indeces. This intercorrelation matrix defines several possible paths through the six-dimensional feature space. If a particular path is selected, the connectives between adjacent sentences can

be generated by a few strategies: what is relevant in reading these stock reports is the evaluation of the stock implied by each sentence, especially whether it is positive or negative. In terms of this evaluation, each sentence pair is either a contrast, comparison, coordination, intensification, or regression. Therefore, the strategies listed in the previous section apply, and since they carry with them a choice of possible sentence connectives, one of the possible connectives can be chosen. Thus, from a list of six sentences, Kozminski obtains a properly ordered, cohesive text.

Kozminski's work (and also Atlas (1979) mentioned before in another context) illustrates a promising research strategy: to isolate from the huge problem of text production manageable sub-components and to study these in detail, both experimentally and theoretically. Attempts to test the model as a whole would be futile.

C. Editing and Review, and D. Process Control

These are the two last components of a process model of writing, neither of which can be considered here. Flower & Hayes have done some important descriptive work on these problems, but of course most of the questions remain for further research: what are the conditions that trigger editing and review? How do they interact with the various search and organization strategies? How should they be used? How can they be used, given the flexibility and versatility of recent computer editing systems? What are the

conditions that transfer control from Generation to Organization to Translation, and vice versa? When are some control strategies to be preferred to others? Most of all, a more systematic approach is needed to investigate these problems than the random sampling of questions I have given here.

E. Creativity

It is important to say a few words here about the creative aspects of writing, if for no other reason, than to show that this important problem falls at least in principle within the domain of the model. So far, I have restricted the model to noncreative writing: within certain constraints, the writer generates ideas from his LTM and connects them along established lines (a kind of regurgitation).

In creative writing (as always, I am more concerned with the creation of ideas, rather than the way they are expressed), there is an additional processing component. The ideas generated from LTM are connected along new lines that do not simply retrace relations already established in LTM. Instead, new relations among ideas are inferred, and, indeed, new ideas are formed. (Formally, this amounts to the same thing, since both relations among ideas and ideas are represented as propositions).

One needs to describe a system of transformation operators, inference rules, or analogical processes that govern the generation of new propositions from old ones. This has been done for limited domains, such as set inclusion hierarchies, but there is room for

more research. In terms of experimental paradigms to study these processes, problem solving and analogical reasoning suggest themselves, for there it is not sufficient that some old idea be retrieved from LTM, instead a new one must be formed.

If we thus distinguish between the reproduction and the production of ideas, what properties of LTM provide a friendly environment for production? Materska (1976) has shown that reproduction and production (experimentally defined) are not always correlated positively. When reproduction reaches a very high level, production may in fact be inhibited. In terms of the present model, what is responsible for this inhibition effect - very strong, dominant associations? Since the same memory system is used in both tasks, what are the characteristics that permit a more efficient use of that system for one task than for the other?

Conclusion

There are too many questions without answers. The purpose of this essay was to show that the theoretical framework suggested here might eventually yield answers to these questions. I have here not a theory but a program for building and exploring one.

Footnotes

1. The term 'memory' is used here in a very broad sense, in accordance with its use in experimental psychology: it comprises all kinds of information stored in the brain, from personal experiences, to general knowledge, to procedures and strategies. Non-psychologists often find this use of the term puzzling, and object to such statements as 'ideas in writing are generated from memory'. But no harm is done, as long as one understands in which sense 'memory' is used in that phrase.

REFERENCES

Anderson, J. R. & Bower, G. H. Recognition and retrieval processes in free recall, Psychological Review, 79, 97-123.

Atlas, M. A. Addressing an audience: a study of expert-novice differences in writing. Technical Report No. 3, Document Design Project, Washington, D.C., 1979.

Black, J. Psycholinguistic processes in writing. In S. Rosenberg (Ed) Handbook of applied psycholinguistics. Hillsdale, N.J. Erlbaum, in press.

Brooks, C. & Warren, R. P. Modern rhetoric. New York: Harcourt, Brace, and Co., 1958.

Bruce, B., Collins, A., Rubin, A. D., & Gentner, D. A cognitive science approach to writing. Technical Report No. 89, Center for the Study of Reading, Cambridge, Mass., 1978.

van Dijk, T. A. & Kintsch, W. Strategies for discourse comprehension, in preparation.

Flower, L. S. & Hayes, J. R. A process model of composition. Technical Report No. 1, Document Design Project, Washington, D.C., 1979.

Grice, H. P. William James Lectures, Harvard University, 1967.

Hirsch, E. D., Jr. The philosophy of composition. Chicago: University of Chicago Press, 1977.

Kintsch, W. The Representation of Meaning in Memory. Hillsdale, N.J.: Erlbaum, 1974.

Kintsch, W. & van Dijk, T. A. Toward a model of text comprehension and production. Psychological Review, 1978, 85, 363-394.

Kozminski, E. In preparation.

Landauer, T. K. Memory without organization: Properties of a model with random storage and undirected retrieval. Cognitive Psychology, 1975, 7, 495-531.

MacWhinney, B. A framework for sharing points. In R. Schiefelbusch, (Ed.) Communicative competence: Acquisition and intervention. Baltimore, Md: University Press, 1980.

Mandler, G. Organization and repetition. In L. G. Nilsson (Ed) Perspectives on memory research. Hillsdale, N.J.: Erlbaum, 1979.

Meehan, J. R. The metanovel: Writing stories by computer. Research Report No. 74, Computer Science Department, Yale, 1976.

Materska, M. Relations between productive and reproductive utilization of information at different stages of learning. Polish Psychological Bulletin, 1976, 7, 115-123.

Pitkin, W. J., Jr. X/Y: Some basic strategies of discourse. College English, 1977, 38, 660-672.

Polson, P. G. Personal communication, 1980.

Raaijmakers, J. G. Retrieval from long-term store. Nijmegen: Stichting Studentenpers, 1979.

Rubin, A. D. A theoretical taxonomy of the differences between oral and written language. Technical Report No. 35, Center for the Study of Reading, Champaign, Illinois, 1978.

Schank, R. & Abelson, R. Scripts, plans, goals, and understanding. Hillsdale, N.J.: Erlbaum, 1977.

Spilich, G. J., Vesonder, G. T., Chiesi, H. L., & Voss, J. F. Text processing of domain related information by individuals with high and low domain knowledge. Journal of Verbal Learning and Verbal Behavior, 1979, 18, 275-290.

Tulving, E. Relations between encoding specificity and levels of processing. In L. S. Cermak & F. I. M. Craik (Eds.) Levels of processing in human memory. Hillsdale, N.J.: Erlbaum, 1979.

Voss, J. F., Vesonder, G. & Spilich, G. Text generation and recall by high and low knowledge individuals. Journal of Verbal Learning and Verbal Behavior, 1980, in press.

Winston, P. H. Artificial intelligence. Reading, Mass.: Addison-Wesley, 1977.