DOCUMENT RESUME

ED 197 326 CS 005 867

AUTHOR Lesgold, Alan M.: Perfetti, Charles A.
TITLE Interactive Processes in Reading: Where Do We Stand?
INSTITUTION Pittsburgh Univ., Pa. Learning Research and Development Center.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
REPORT NO LRDC-1980/18
PUB DATE 80
NOTE 39p.: Small print may be marginally legible.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Cognitive Processes: Epistemology: Language Processing: *Learning Theories: *Psycholinguistics: *Psychology: *Reading Processes: *Reading Research: Reading Skills

ABSTRACT

Much of the current research in reading processes has been aided by movements in experimental psychology known as information processing psychology, cognitive psychology, and cognitive science. The information processing movement has contributed three important ideas: (1) Logogen theory of a cognitive response unit that is sensitive to the set of auditory, visual, and semantic features associated with a given word: (2) the idea of a limit on the amount of conscious mental processing in which a person can engage at one time: and (3) the cascade theory, which holds that there are two or more levels of processing that have several properties, each operating continuously on outputs at the next lower level. Separately, a methodology has developed of trying to understand acquisition of a skill by studying differences between people of greater or lesser expertise. The newly emerging cognitive science movement has contributed ideas derived from the work on speech understanding and the distinction between event-driven and goal-driven (bottom up versus top down) processing. (Author/HTH)

1980/18

# INTERACTIVE PROCESSES IN READING: WHERE DO WE STAND?

Alan M. Lesgold and Charles A. Perfetti

Learning Research and Development Center

University of Pittsburgh

To appear in
A. M. Lesgold and C. A. Perfetti (Eds.),
Interactive processes in reading,
Hillsdale, NJ: Lawrence Erlbaum, Inc.,
in press.

The research reported herein was supported by the Learning Research and Development Center, supported in part as a research and development center by funds from the National Institute of Education (NIE), United States Departments of Health, Education, and Welfare. The opinions expressed do not necessarily reflect the position or policy of NIE, and no official endorsement should be inferred.

# INTERACTIVE PROCESSES IN READING: WHERE DO WE STAND?

Alan M. Lesgold and Charles A. Perfetti

Learning Research and Development Center

University of Pittsburgh

## Progress in Theorizing About the Reading Process

It has long been evident that reading is a very complex activity, but only recently has the necessary set of tools for directly understanding that complexity begun to appear. Until a few years ago, the best we could do was to attack each aspect of the reading process as a separate research problem, more or less as the proverbial set of blind men tried to understand the elephant. This often entailed a need for many different blind men, i.e., a large number of different (but still oversimplified) research approaches, in order to gain any useful knowledge. Unfortunately, each simplistic approach manufactured its own theories of reading dysfunction, and there proliferated a complex typology of reading disorders. Since the mechanism that does the reading and the experience base that results in learning to read are both complex, it was possible to isolate apparent examples of each of these disorders, and the blind-man approach has proven to be quite useful to the special education field, aiding in understanding the many ways in which the verbal processing apparatus can fail.

However, the majority of inadequate readers do not, we suspect, have rare or exotic problems that are well suited to analysis via a complex typology. To understand why these children and adults cannot read well, we need to understand the overall reading process well enough to be able to identify its points of vulnerability, those components that must work efficiently for effective reading to occur. The positive message of portions of this book and other recent work is that we are closing in on a major portion of that goal. In the pages that follow, we offer some suggestions about why we think progress in this area has accelerated and what we think needs to be done next.

## Recent Influences on Reading Research

We believe that much of the current work has been aided by a few seminal contributions of the past decade or two. These developments have come from overlapping movements in experimental psychology known as information processing psychology, cognitive psychology, and cognitive science. The information processing movement, born in vigilance and attention work that began during World War II, has contributed three important ideas: Morton's logogen theory (1969), the idea of a limit on the amount of conscious mental processing in which a person can engage at one time (Atkinson & Shiffrin, 1968; Kahneman, 1973; Newell & Simon, 1972), and the cascade theory of McClelland (1979), all of which have heavily influenced our recent work (Lesgold & Perfetti, in press). Somewhat separately, there has developed a methodology of trying to understand acquisition of a skill by studying differences between people of greater and lesser expertise

(e.g., Chase & Simon, 1973). This work complemented a long tradition in reading research (comparing "good" and "poor" readers) by suggesting specific knowledge sources as responsible for expertise. Finally, the newly emerging cognitive science movement, an integration of cognitive psychology with the artificial intelligence domain, has contributed ideas derived from the work on speech understanding (Erman & Lesser, 1978) and the distinction between event-driven and goal driven (or bottom-up vs. top-down) processing (Bobrow & Norman, 1975). We consider each of these areas in turn.

Logogen theory. The neuron has been an influential metaphor for understanding many higher level aspects of cognition (see, for example, McCulloch & Pitts, 1943). An important theoretical program that derives from this metaphor is the logogen theory of John Morton (1964/1968, 1969). Morton proposed that for each word one is able to recognize, there is a response unit, called a logogen, that is sensitive to the set of auditory, visual, and semantic features associated with that word. When the number of features that are currently active (i.e., being looked at or recently thought about) exceeds the logogen's threshold, that unit is automatically activated, and all the features are made available to the rest of the cognitive apparatus. Because logogen activation is automatic and does not require attention, the logogen theory is a theoretical forerunner of automaticity theories of reading. Indeed, Morton and Long (1976) have presented data that suggest that logogens are not subject to the capacity limitations that characterize higher levels of cognitive processing (see next section).

Morton's contributions go further than we can consider in this writing. For one thing, the logogen is not only an embodiment of an automated recognizer for a word; it also has a natural extension to accommodate contextual processes (cf. Morton, 1969). Context is simply the set of active or recently active semantic features. Thus, a top-down or contextual influence is nothing but the activation of semantic information patterns to which the logogen is sensitive. The problem with the logogen formulation is that it leaves unanswered the question of what a feature is. Although an answer to this question is important for any substantial theory of semantic processes, it seemed less critical in the formulation of how a word recognizer might use information in a general way without regard to semantic structure.

Basically, the Morton logogen is similar to certain aspects of one of the recent speech understanding models discussed later, HARPY. By having an automatic recognition response whenever a threshold number of critical "features" is activated, a theory can account for such phenomena in reading as speed-accuracy tradeoffs, word frequency effects, context effects, etc. On the other hand, although logogen theory is an important precursor of more recent work, it (at least in its earlier published versions) does not tell us enough about the overall structure of the word recognition process as it relates to reading. Further, the logogen seems to be an appropriate model only for the automated level of performance in word processing. We still need to learn how more complex, inferential, semantically driven conscious processes become "compiled" into logogens.

A limitation on processing capacity. One way to describe the

inability of a person to perform some function is to say that he or she has limited capacity. Such a statement, by itself, is a nonexplanatory restatement of that person's inability. However, if it is possible to specify the nature of the capacity limitation in some detail, then a limited-capacity account becomes a more useful explanation. In the case of reading, less skilled readers have sometimes (e.g., LaBerge & Samuels, 1974; Perfetti & Lesgold, 1977) been characterized as having problems that involve a capacity limitation. The argument has been that less practiced components of word recognition require a substantial allocation of processing capacity that otherwise could be used for higher-level aspects of the reading process. It is necessary, however, to specify better what processing capacity is in order for this sort of argument to be a contribution.

Several approaches to characterizing this limited-supply commodity have been proposed. Newell (1980) has suggested an interpretation based upon production system models of cognition. A production is a conditional mental operation; it is performed only when its specified conditions are satisfied. Any computational model of cognition can be specified as a memory structure combined with a set of productions and a discipline (set of rules) that specifies the order of execution when the conditions of several productions are simultaneously satisfied. The conditions of productions consist of patterns to be matched against active portions of memory. Some patterns are very specific, whereas others are more flexible, containing free variables (essentially, "wild cards") as part of the

pattern to be matched. This flexibility means that parts of the pattern to be matched are not completely specified (e.g., "If someone has a sister, that someone is a brother or sister"; as opposed to "If John has a sister, then he is a brother or sister"). When such a pattern is matched to active memory, the free variables must be bound to the specific parts of the pattern for which they are to stand. Newell has proposed that there is a limitation on the instantiation, binding, and use of such variables. That is, there is a limit on the speed at which conditions of productions containing variables can be tested (i.e., it takes time to match someone to a specific person).

Within this approach to limited capacity, a more expert reader would presumably be modeled as one whose competence consists in having a very rich set of specific productions rather than only a smaller set of vague, nonspecific productions. The approach argues that one trades off generality for execution speed. A general production (e.g., "If the word starts with CA, its first syllable may have the sound /kae/) contains unbound variables in its condition, whereas a more specific one does not (e.g., cattle is pronounced /kaetel/. Since variable binding is a bottleneck in the system, it will be performed only if the number of productions that fit to the point of variable binding is not too large. A specific production that recognizes a situation exactly will not be impeded by this bottleneck.

In the early stages of reading, it is necessary to teach children some productions with unbound variables. These include the phonics principles that permit children to sound out words they haven't seen and the rules for recognizing familiar word stems with common affixes.

Some theories of cognitive learning, such as Anderson's ACT (1976), postulate a second source of productions with unbound variables; they assert that productions with variables in their conditions are generalized from more specific productions that are the result of specific experiences. For example, experience with the word CAT may lead to a temporary behavior of treating any word that starts with CA as CAT. Such a generalization mechanism is the basis for any adaptive performance. However, in the case of word recognition, it may be counterproductive. In ACT, simple trial and error will, with practice, tend to compensate for excessive generalization by strengthening successful productions and weakening those that are too general.

To summarize, one approach to theorizing about the limitations on thinking ability is to characterize the limit as an inability to match the conditions of productions very quickly when they contain unbound variables. Children acquire productions with unbound variables through instruction, such as phonics rules and stem+affix rules, and also through overgeneralization that is adjusted with practice. Thus, children should show limited capacity effects once they have learned the barest rudiments of reading, and these effects should persist until removed by extended practice. No one has addressed the issue of whether it is possible to design instruction to minimize the formation of, or need for, productions with unbound variables, but this would seem like a sensible issue for future work.

Another aspect of limited capacity is the limited duration of those memories that are "partial products" of cognition. Originally,

psychologists spoke of long-term memory and short-term memory, with the short-term memory being extremely limited (Miller, 1956). More recently, it has become apparent that cognition requires a considerable amount of what Hunt (1973) has called "working memory." Several forms of evidence suggest that the contents of one's recent experience are temporarily available for further cognitive processing. For example, when one reads a sentence in a discourse, one usually can retrieve enough of the prior sentence(s) to resolve anaphorical references, even when those references are quite vague or indirect. If there have been intervening context changes, this retrieval becomes harder or impossible (cf. Lesgold, Roth, & Curtis, 1979).

There are several possible mechanisms for working memory loss. The simplest is to assume that working memory decays after a certain amount of time. Such an assumption is compatible with most global models of reader disability. One assumes that poor readers do too much slow (attention-demanding) processing. Thus, their working memories will decay before they are needed, at least some of the time. Unfortunately, a model of this sort cannot explain why working memory availability is diminished by shifts in the topic or context of a discourse. Consequently, it may be worthwhile to consider a more elaborate theory of working memory--one in which there is the possibility for more specific differences in working memory between better and less skilled readers.

One such theory would be that all working memory is simply a manifestation of episodic (Tulving, 1972) memory. Episodic memory can be thought of as a content-addressable trace of ongoing cognitive

experience. When part of the content of an experience is used as a retrieval cue, the rest is returned, with a noise level that decreases as the input more closely matches the total cognitive activity under way during that experience. Some of the content of any such episodic memory is irrelevant, but variable, "system noise." Such noise will be more of a problem as the time between storage and retrieval increases (the components of the noise can be thought of as undergoing random walk; Landauer's model, 1975, is a specific variation on this theme). Thus, there will tend to be a trade off between the recency of storage for an episodic trace and the amount of partial content needed to retrieve the rest of the trace, because the number of irrelevant matching features will decrease as time passes between storage and attempted retrieval. Context shifts would also tend to decrease the match between the current ambiance of features and that of the prior context.

One might add the assumption that memory nodes matched by conditions of an executing production are automatically stored as part of episodic traces if they are matched by bound variables. If matched by free variables, their storage into episodic memory is assumed to be not as complete. Such an episodic working memory would have the property of being "bigger" for people who have a rich array of specific productions than for people who have learned only very general productions (containing many unbound variables). That is, a bigger proportion of the content of relevant episodic traces will be task relevant for the expert than for the novice, since more of the content will have been generated by the expert's own specific

procedural knowledge and less by temporary variables. As a result, the information will be retrieved more reliably.

To summarize this section, we note that capacity limitations have been a popular way of talking about why some children don't read well. In recent years, both the empirical work on reading ability differences and the work on cognitive simulations of reading have allowed enough specification of detail for this approach to become valuable. In contrast to the earliest work on reading ability differences, it is currently more likely that a theory of ability differences will talk about the interaction of cognitive software of different types with the general bottlenecks in the human information processing system than about possible differences in system hardware, although some level of hardware differences may be present.

Cascade theory. The measurement of reaction time has been the dominant empirical technique of the information processing tradition in psychology. There are several reasons for this. First, time is a primitive unit of physical measurement. Consequently, the early psychophysics work that spawned psychology chose response time as a suitably rigorous dependent variable. Second, Saul Sternberg (1969) developed a class of experimental designs using reaction time to test theories of cognitive processing in which component processes execute in strict linear sequence. Finally, there has often been no other measure with fine enough grain to capture the level of theorizing in current research. The data of earlier psychological work, such as overall proportions of correct responses, etc., are quite overdetermined by today's theories of mental processing and thus are

inadequate for testing theoretical validity. Within current paradigms, one thing that is true of even the smallest mental process is that it takes time.

Even with this impressive history, the methodologies for reaction time analysis have generally been inadequate. This is because the basic approach was to assume that treatment manipulations could be found that would independently affect only one component of a process and that reaction time changes produced by such manipulations were completely due to changes in the function of the target component. When components are assumed to interact while they are operating or when the speed at which they operate depends upon the quality of data they receive from lower level components, the existing methodologies are not wholly adequate.

More recently, McClelland (1979) proposed a new type of relationship between components of a mental process to augment the prior model of purely sequential and noninteractive components. This new relationship is the basis for his cascade theory. McClelland developed in considerable detail a basis for reaction time designs that test cascade models, and it is likely that such models will be useful alternatives for theorizing about specific mental functions, including reading. The assumptions underlying cascade theory are a somewhat generalized version of the assumptions presented by Perfetti and Roth (in press), and we do not consider all of them here. However, we should briefly review what a cascade model is.

According to McClelland, a cascade model is one in which there are two or more levels of processing that have several properties.

First, each component at a given level operates continuously on outputs of components at the next lower level. Second, each component is continually outputting, with some time lag, the current state of its computations based upon the input levels it has been receiving. Finally, there is no direct transmission of data from higher to lower levels. The efficiency of each component is determined by the rate at which it responds to input. The output quality of each component is determined by the asymptotic activation level for the component (the clarity and completeness of the output it can generate given sufficient time).

The McClelland model is more restricted than some current theoretical approaches, such as the Rumelhart interactive model (1977), which does not define a strict directionality of relationships between different levels of processing. (Note also that the directionality assumption is not followed by Rumelhart and McClelland, in press.) On the other hand, there are only a few indications that the less restricted approaches require strong bidirectionality to account for available data. Perhaps this is because our theorizing sophistication is still ahead of our empirical capabilities. In any event, we can view the McClelland work as an important extension of our ability to closely identify theory with data, even if it turns out to be too restrictive as an overall model.

McClelland raised some important points in his article that deserve some discussion here. For example, he demonstrated that the same data can have a different interpretation and even multiple interpretations under the cascade theory assumptions. In the case of

multifactor RT experiments, for instance, a statistical interaction of two treatment factors no longer means that the two types of manipulations must be affecting the same process component. Under the assumptions of cascade theory, any one of the following three possibilities could produce an interaction: (a) the two manipulations affect the efficiency of the same process component; or (b) the two manipulations affect the asymptotic level of output from the same component (i.e., output quality); or (c) one manipulation affects the rate (efficiency) of a component whose rate is a limitation on overall system efficiency while the other manipulation affects the asymptotic activation level (output quality) of some other component. Further, the lack of an interaction effect does not rule out the possibility that two manipulations might affect the same component.

McClelland's contribution goes beyond pointing out a set of alternative models with which current data may be consistent. He presented examples of such alternative models that appear to have great potential. For example, he commented in his (1979) article on the interaction in word recognition of attentional variation and different levels of word frequency. A cascade model that he proposed for this relationship would have the rate at which word recognition components respond be determined by the level of attention allocated to recognition and the asymptotic level of activation for the demon that recognizes a word determined by that word's frequency. Thus, low frequency words would be recognized more slowly and less accurately than high frequency words, especially when attention was diverted to other components than word recognition. An obvious extension of the

model would state that practice using a particular word will improve its asymptotic level, perhaps with different improvement rates for different children.

Although such a model is quite appealing (especially to authors who have publicly stated hypotheses that are less precise variations on this theme), it is important to consider whether the greater precision can lead to greater possibilities for empirical validation of such hypotheses. The techniques McClelland cited, unfortunately, seem better suited to experimentation with competent (probably adult) readers than with children and seem especially unsuited to experimentation on children who read poorly. The problem is that the primitives of a cascade theory, from which more complex predictions are generated, are functions that show individual process component output as a function of processing time. Directly gathering the data for estimating those functions seems to require complex techniques, as most reading process components execute in a few hundred milliseconds or less.

The two methods that have been used thus far for such measurements are deadline tasks and response-signal tasks. In a deadline task, the subject must respond by a particular deadline, a fixed number of milliseconds after the stimulus is presented. By varying the deadline, it is possible to construct a speed-by-accuracy plot, which is the operating characteristic function that we need for directly testing a cascade model. An alternative approach requires that the subject respond as quickly as possible after a response signal. By varying the latency from stimulus onset to response

signal, one can generate the operating characteristic function.

Both methods impose an additional processing load on the subject. Further, it is unlikely that a small child can understand the response signal task in the context of a response requirement that always seems faster than normal. Children may not understand what it means to be as accurate as possible but to take no more than, say, half a second. Even adults require training on such methods (Wickelgren, 1977). Thus, the experimental procedures suggested by McClelland will not work in studies of children's reading problems. It remains to be seen whether techniques such as making more refined use of the density function for correct and error RTs from simpler tasks (for an example, see Grice, Nullmeyer, & Spiker, 1977) can get around this problem.

This does not mean that the cascade theory will not be important and useful. It has begun to deal with the problems of directly verifying interactive component theories of processing with reaction time measures. Further, the specific cascade proposal is one of a class of models that can account for process interaction data such as that reported by Perfetti and Roth (in press). Nonetheless, the increasing specificity and complexity of theories such as McClelland's highlight the problem we wish to address next.

Issues of methodology. For a variety of reasons, empirical work in the study of reading skill acquisition has lagged behind theoretical work in recent years. In large part, this is because the artificial intelligence discipline has recently become strong enough to foster work in other areas. This work has given us not only metaphors for our own theories but also simulation methodologies for

exploring the implications of our theoretical work. Unfortunately, methods for empirical verification of our richer and more detailed theories have not emerged as quickly, though there are hopeful signs that this is changing.

The basic problem is that children do not provide rich enough behaviors in a laboratory setting. Further, there is considerable "error variance" in their responses (some of which theories should account for and some of which is perhaps best characterized as attentional variability). Thus, even though children, and especially less-skilled children, produce less complex behaviors and have less tolerance for experimental tasks, they also require more experimental trials in order to produce stable data. We are left with many degrees of freedom in our models but little detail in our data. The problem becomes even more severe when models of learning to read are being tested.

Four general approaches have emerged that we wish to discuss. First, whereas responses must be kept simple, stimuli can be varied in complex ways. Second, a large battery of different tasks can be used. Third, several sophisticated forms of data analysis have been applied to this problem. Finally, techniques of developmental psychology are being adapted to the study of long-term learning. We briefly explore several examples of these four approaches.

The sophisticated manipulation of stimuli is perhaps best illustrated by some of the experiments reported by Rumelhart and McClelland (in press). They had simple tasks, such as identifying single letters within words (although they used adult subjects,

children can also do such tasks). What was varied was the asynchrony between the time that the target letter was presented and the time that other letters of the word were shown. These time differences were of millisecond magnitude, a very subtle manipulation, yet they produced data adequate to the testing of a rather broad and important principle.

A second example is found in the work of Frederiksen (1978a, in press). Frederiksen has combined the use of a large number of tasks with theoretically relevant stimulus variations within tasks. This permits very specific tests of complex hypotheses about the sources of reading inadequacy in his high school subjects. Frederiksen has also pioneered the use of structural analyses of correlational data to verify complex theories (1978b; personal communication, 1980). With appropriate care, it is possible to gather considerable detailed data about children's reading performances using both natural and laboratory tasks. What is difficult, if not impossible, is to gather rich enough data in true experiments (in which all relevant independent variables are manipulated directly by the experimenter).

One way around this is to use recent structural equations modeling techniques (e.g., Joreskog & Sorbom, 1978) to test complex hypotheses against correlational data. Such tests allow one to specify the hypothesized set of skills that are present in each of a battery of tests and also to verify hypotheses about the extent to which one component skill of reading enables improvement in another. A recent dissertation (Lomax, 1980) nicely demonstrates this technique and shows the verification of a "bootstrapping" model in which word

recognition skill enables improved reading speed, which in turn enables even better word recognition. It also provides a guide to the relatively complex details of the technique.

A final empirical approach has been the application to reading acquisition of the developmental techniques of cross-sectional (e.g., Curtis, in press; Doehring, 1976) and longitudinal (Calfee, 1980; Lesgold & Curtis, in press) comparison. In essence, these techniques expand the Frederiksen type of approach to include multiple testings over the course of learning to read (in either the same or different subjects). When combined with the causal modeling techniques just discussed, it should be possible to generate the learning trajectories of a variety of very specific subskills for both more and less successful readers and to verify hypotheses about the sources of overall reading skill. In particular, we expect that longitudinal data, when analyzed using the Joreskog structural equations approach, will permit both specification of the components of skill at successive levels of reading expertise and the understanding of the mechanisms whereby children of different aptitude levels improve their skills. This leads us to our next topic, comparisons of experts and less skilled people.

Expert-novice comparisons. One approach to studying the problems children have in learning to read has been the comparison of children of differing levels of skill. This approach has a long history in reading research but has recently been most prominent in studies of high vs. low achievers in the reading curriculum (e.g., Curtis, in press; Frederiksen, 1978, in press; Perfetti & Lesgold, 1977). Such

work has been largely empirical, and is, for the most part, discussed elsewhere in the literature (Lesgold & Perfetti, in press). Our purpose at this point is to suggest that another form of contrastive research be given more attention. This is the building of empirically verified models of children's reading performance at different levels of expertise as a means of better understanding how learning to read happens.

This general approach has been analyzed into three steps by Glaser (1976). First, one must construct a model of skilled performance. Second, procedures must be developed for specifying the status of the learner's skills at instructionally relevant points in the course of learning. Finally, procedures for producing transitions from one skill level to the next need to be specified. This is essentially a means-ends approach to the problem of instruction that does not by itself represent a major breakthrough. The important breakthrough comes from the realization that specific simulation models of the different stages of reading expertise may be possible, and that it may even be possible to test instructional hypotheses by seeing if they produce transitions of a less-expert model into a more expert one. At the time this chapter was written, there were a number of projects under way using variations on this approach to specify learning mechanisms for physics (Larkin, 1980), arithmetic (Brown & Van Lehn, 1980), geometry (Anderson, Greeno, Kline, & Neves, 1980), and computer programming (Polson, Atwood, Jeffries, & Turner, 1980). In the future, we expect to see similar efforts for reading. Hopefully, such modeling will be done in tandem with some of the more

sophisticated empirical procedures already discussed.

Contributions from the speech understanding work. Another major source of guidance for interactive models of reading is the work stimulated by a major Defense Advanced Research Projects Agency (DARPA) effort in the early 1970's to develop speech understanding systems (Department of Computer Science, Carnegie-Mellon University, 1977). DARPA conducted a competition among several institutions to produce a speech understanding system with a certain level of skill and efficiency by 1976. The goals were set very high, and it appeared that none of the projects would meet them. Two very different programs developed at Carnegie-Mellon ended up coming very close to meeting the requirements. One of them, HEARSAY-II, differed from the other efforts primarily in having a looser control structure and many different levels of relatively independent decision processes. The other, HARPY, had a more tightly structured control flow and was compiled, or optimized, in ways that precluded easy modification.

It is becoming increasingly clear that there are a number of rather elegant principles embedded within the HEARSAY effort that may be quite useful to our task of modeling another difficult, multiprocess, understanding activity, namely reading. In this section, we explore some of these principles and also consider the thesis that HEARSAY is a good step toward modeling relatively novice performances, while other approaches to speech understanding, such as HARPY, are better but less complete characterizations of expert performances.

One interesting comment can be made about the expert-novice

difference as characterized by intelligent systems such as HEARSAY-II and HARPY. In contrast to the suggestions of some reading researchers, the expert models are more "bottom-up" than the novice models. That is, models such as HARPY do not have a central high-level strategy mechanism controlling which components are allocated attention, at least not to the extent that models such as HEARSAY-II do. Hence, the progression is from top-down novices to bottom-up experts (just as in chess; Chase & Simon, 1973). This suggests that we will want to be extremely careful in theorizing about the top-down aspects of reading. Mature readers most likely accomplish the recognition of words in a relatively bottom-up manner, as some authors (Lesgold & Perfetti, in press) have suggested. Presumably, they behave in a more top-down manner in making sense of the sentences they are reading, especially if they are reading in a domain for which they have little expertise.

We begin by reviewing some of the properties of the HEARSAY system, relying upon the Carnegie-Mellon summary reports (Department of Computer Science, 1977). All of the candidate speech understanding systems are multilevel systems; that is, they contain interacting knowledge structures operating at several different levels of analysis. Although a multilevel structure is important any time complex recognition is required, it is almost an absolute necessity in speech processing because of the ambiguity of the speech signal. The very same sound sequence can have different meanings in different sound contexts. Much of this ambiguity escapes our everyday experience because we have developed multiple levels of processing.

In reading, there is less ambiguity of input, but the complexity of recognition and comprehension mandates a multilevel model nonetheless.

At each of the levels of processing in HEARSAY, there are relatively independent knowledge structures that are activated when specific conditions are satisfied in the course of processing and that act by making certain computed results available for examination by other knowledge structures. The structures look a lot like the logogens of Morton (1964/1968), but they exist at levels lower and higher than the word level. Knowledge structures communicate via a message center or blackboard, a sort of unrestrained short-term memory. The basic idea is that the speech signal triggers certain low level knowledge structures. Low-level output, combined with the original signal information, triggers more knowledge structures at higher levels, and this process continues until a high-level structure generates an overall interpretation in which it has great confidence in.

Such a system, if totally unconstrained, will suffer from combinatorial explosion of the set of triggered knowledge structures. That is, each knowledge structure will execute when it can and can trigger additional knowledge structures with its actions. If there is considerable ambiguity in the signal, this will produce a mushrooming effect with more and more knowledge structures ready to execute. As a result, processing resources are overtaxed, and a correct interpretation is likely to be obscured by the chaos. To avoid these problems, there must be a discipline imposed on the system that permits only some of the triggered knowledge structures to execute.

The specific discipline chosen will substantially determine the nature and effectiveness of the system.

The HEARSAY-II discipline is important for two reasons. First, it provides a lesson about how allocation of processing capacity might take place in a multicomponent system. Second, the experience of the HEARSAY project in trying out different levels at which to concentrate decisions about the allocation of resources may be instructive. Within HEARSAY, there are a number of levels of knowledge structures. Any scheme to decide which of the potentially applicable structures should execute must look at the current blackboard contents and decide how the probable effects of a particular knowledge structure will contribute toward selection of the best overall interpretation of the utterance being processed.

The problems faced by a speech understanding system and by a text understanding system are rather similar. The system can look at a hypothesis for the entire utterance and see which word and subword hypotheses would confirm it further, for example. Alternatively, it can select the strongest phoneme hypotheses and activate word and subword hypotheses based upon them. A number of other schemes are also possible, but any optimization scheme has the property that it must take into account the results of processing done thus far and must be able to predict, at least in part, what any given knowledge structure is likely to accomplish if attention is directed toward it.

In HEARSAY, there is a component of every knowledge structure called a response frame, which provides this prediction. Even though it seems a bit difficult to propose that we need to know what we are

going to do before we do it, one might argue that it is exactly this property that characterizes successful performance of any complex activity, including human thinking activity. For example, if we have a medical problem, we decide on a specialist without knowing exactly what diagnosis he or she will make. Within psychology, there is a long tradition, going back to James and Pillsbury, among others, of positing two levels of awareness (or allocation of attention). More recently, MacKay (1973) has demonstrated that unattended information, though not being consciously noticed, can sometimes be shown to have had some influence on understanding. Thus, it is not unreasonable to think of a psychological model that functions by having relatively independent knowledge structures that are able to do a little bit of processing automatically but that require conscious attention in order to complete their work. The lesson from HEARSAY is that such models can be very effective as understanders. We expect them to become more prevalent in the future.

The issue of the level at which most attentional allocation decisions should occur is raised by the HEARSAY work, but perhaps not resolved for tasks other than the processing of sentence-level spoken utterances. HEARSAY seemed to work best when it attempted to allocate attention to knowledge structures directed at confirming word and subword hypotheses that might extend hypothesized multiword sequences. That is, hypotheses, at all levels, that would have the effect of expanding highly weighted hypotheses of two or three consecutive words in a sentence by incorporating an additional word or two were selectively favored. It remains to be seen whether the word and

phrase level is critical in the meeting of top-down and bottom-up aspects of processing in reading, also; but it is a fact that for speech understanding, certain levels worked better as control levels for HEARSAY-II than did others.

One final comment might be made about the speech understanding models in particular and intelligent systems in general. This is that some models seem to be better theories of expert processing while others seem to be better theories of novice levels of skill. This does not mean that the expert-like models are more intelligent or more successful--many are very inadequate attempts at simulating expert behavior. Rather, it means that the style of the expert-like models is similar to the style of human experts as they have appeared in psychological studies of expertise.

To understand what an expert model is like, it may be useful to review what HEARSAY, which we consider a successful novice model, is like. HEARSAY has a very fresh mind. There are no constraints on short-term memory structure; the results of any mental process are available on the blackboard. Decisions are made in a conscious, hypothesis-testing mode which is optimized by attending first to more promising leads. The execution discipline, which decides how processing capacity (attention) is to be allocated, is extremely important to the success of HEARSAY for this reason. Finally, it is very flexible. New knowledge can be incorporated by simply adding additional knowledge structures.

In contrast, another Carnegie-Mellon model, HARPY, is more expert-like. It has automatic, clearly differentiated, short-term

knowledge pathways rather than an amorphous blackboard. The flow of control is managed by the components currently executing, with each component passing off control to the appropriate successor without the (conscious) intervention of a central strategy. Knowledge structures are larger and have more extensive output. Also, HARPY tends to prune from further consideration all but the most highly weighted of hypotheses currently being considered. Finally, because HARPY is finely tuned (compiled and optimized, in computer terms), it is less easily changed than programs such as HEARSAY-II.

The contrast between HARPY and HEARSAY-II shows both the strengths and weaknesses of the two as models of expert and novice behavior. HARPY is more efficient in large part because it quickly and accurately classifies the input and brings just the right knowledge structures to bear on it. On the other hand, it is less able to handle unexpected mutations of the input and less able to learn, yet we continue to feel that experts, at least expert readers, have the flexibilities that HARPY lacks. Nonetheless, we have learned a lot from the two models and expect that their influence on improved theories of the reading process has been and will be substantial.

## The Elusiveness of Phonological Processes

One of the process interactions of major theoretical and practical importance involves speech-based processes. In a recent work we included four chapters that have something direct to say about speech processes in reading (see Baddeley & Lewis; Levy; Katz & Feldman; Barron, all in Lesgold & Perfetti, in press).

A striking fact is that evidence for speech-based processes in skilled adult reading is fairly elusive. In a conference paper presented in 1976 but only recently published (Perfetti & Lesgold, 1979), we reviewed some issues concerning speech processes in reading, including experiments with lexical access and/or comprehension, and concluded that speech processes played an important role in supporting comprehension. Certainly, this was not an idiosyncratic conclusion, supported as it was by the research of Kleiman (1975) and Levy (1975) that appeared to demonstrate an immediate memory role for speech-based processes. More contentious was our conclusion that then available experiments could not "... be used to build a strong case against phonological coding" (p. 73) as a necessary aspect of word recognition in reading-like situations. That conclusion seems not to stand well in the face of more careful research since then, especially that of Coltheart (Coltheart, Besner, Jonasson & Davelaar, 1979; Davelaar, Coltheart, Besner, & Coltheart, 1978) on lexical access. There seems to be little reason to doubt that access to a word <u>can</u> occur without phonemic recoding. If so, the question becomes whether lexical access normally, rather than necessarily, involves some speech process. The focus shifts from whether access requires recoding to the conditions of reading that promote phonetic processes and to what function, if any, is served by such processes.

One reason for maintaining an interest in these questions is that children seem to rely heavily on speech processes while learning to read. There is indirect evidence for this in the fact that young readers who are relatively skilled show their most marked advantage

over unskilled readers in tasks involving production (naming) of words. Lesgold and Curtis (in press) make this point for children just learning to read and note that this difference persists at least through the elementary grades. Also, Hogaboam and Perfetti (1978) report bigger differences between skilled and less-skilled readers in vocalization latency than in word matching, in agreement with Lesgold and Curtis. More direct evidence relating early reading skill to speech processes comes from Liberman, Shankweiler, Liberman, Fowler and Fischer (1977) who report greater phonemic interference effects for skilled readers in a short term memory task. Barron (in press) suggests not only a phonemic memory factor but a phonemic lexical access factor that may favor skilled readers.

There is thus a mild paradox. Speech processes appear to be unnecessary for skilled reading, yet they are characteristic of beginning reading, especially for those who learn quickly. A reasonable way out of this paradox is to suggest a skill acquisition model based on differences between expert adult readers and novice children. Speech processes are important for beginning reading because the child must learn to map print to speech sounds. However, achieving an expert level of skill in reading involves learning to bypass the print-to-speech connection by acquiring a print-to-meaning connection (Wernicke, 1874, 1966, would have been happy with this sort of model). With extended practice at lexical access, attention to phonemic correspondences of letters drops out and perception of letter patterns automatically activates word concepts. The transition from novice to expert probably requires extensive practice, just as in

other areas of intellectual skill, such as chess. The result of this practice can be described as the replacement of generalized phonemic-based production sequences with specific, unified, word recognition productions.

However, there remain a few questions with this solution. The major question is whether phonemic codes continue to serve some role subsequent to lexical access. It seemed reasonable to conclude that even if phonemic codes are unnecessary for lexical access, they are still useful for later memory and comprehension. If so, it is reasonable to suppose that phonemic codes are activated during lexical access. A system would be rather inefficient if it postponed phonemic code access until required by comprehension blockage. What would be used to reaccess the code? Since the visual input would be gone, the only alternative other than re-examining the word(s) in question would be to reaccess the phonemic code via the semantic code. That could be a problem. At a minimum, it would be inefficient insofar as information from a semantic code is connected to a phonemic code more strongly in the name-to-meaning than in the meaning-to-name direction. Hence, retrieving a name given meaning would be more time consuming than the converse.

Perhaps more critical is that semantic information may underdetermine phonemic information. If so, accuracy as well as efficiency becomes a problem. For example, suppose in reading an American history text, the reader encounters the sentence, "Fillmore appeared to have enough influence to forge a compromise in the Senate." If the reader's code for the "meaning" of Fillmore is

something like [+Name, U.S. President, 19th Century] he or she does not have the information sufficient for reaccess to the name. There's nothing to keep the reader from accessing Jackson, Pierce, Harrison, or Tyler, instead.

There are two possible solutions to this problem. One is to assume that a reference-securing process uniquely determines the name. For example, the above example might be supplemented with a reference-securing code such as the one who was president 1850-1853 or the one nobody remembers, or the one whose name is the same as a linguist. The reference securing codes would uniquely determine the name needed. The advantage of this is that it relieves the reader from having to hold onto a name code. It allows an "abstract" meaning-reference code that reaccesses the name when necessary.

The problem with the reference-securing code lies in accounting for words without securable references. In the sentence example, one can imagine secure references for Fillmore easily enough. However, appeared, to, have, enough, and influence seem to resist reference securing. It is possible, in context, to secure reference for the entire phrase enough influence to forge a compromise, something like [the X sufficient to cause Y to agree to Z]. In general, phrases are more reference-secured than words. Thus, the reference securing hypothesis seems to suggest that while "lexical access" may describe an early stage of reading comprehension, the semantic processes necessary for securing reference, and hence necessary for keeping retrieval probability high, will operate over multiword phrases. There seems to be no reason to to disallow such processes.

The second possibility that allows for post-lexical name access is that phonetic (or phonemic) fragments are available. Consider the American history example again. Suppose the reader's code included [+Name, U.S. President, 19th Century, +/f- ]. The difference is that the code includes information concerning the initial phoneme. The probability of reaccessing the name is obviously greatly increased by this assumption. Name accessibility is increased even more if additional phonemic information such as other phonemes or even number of syllables is available. This alternative is actually a different form of the phonemic recoding hypothesis, with a built-in functional assumption. It assumes that some phonemic information is accessed with other lexical information and that at least some of it is kept available for consultation.

This version of the phonemic recoding assumption reveals a possibility of phonemic recoding that is often ignored. The code need not contain all the phonemic information needed to produce the word. It can be abbreviated or partial. This possibility is reflected neither in experiments on lexical access nor in those on sentence processing. The assumptions of existing studies seem to be that the complete sound of the word, the whole acoustic pattern, is what is involved. For example, experiments involving rhyme judgments in sentence processing (Kleiman, 1975) and pseudohomophone effects in lexical decision (e.g., Coltheart et al., 1979; Davelaar et al., 1978) have to assume that phonemic codes are similar to acoustic patterns of some sort. It's not clear that evidence from such research rules out phonemic recoding that is less complete.

In any case, the main point is that phonemic codes are useful for name access because name codes are needed to secure reference. However, if only those words with reference-securing potential (e.g., content words) need to be available, then a generalized phonemic coding procedure for all words would not be necessary. Instead, only reference-securable words would be name accessed. Syntactic words, for example, could be reconstructed.

The reason for assuming a name-accessible memory code is the usefulness of such information in comprehension and memory. Whether name codes are activated during ordinary comprehension or only during verbatim memory situations remains an issue. Baddeley and Lewis (in press) and Levy (in press) conclude that memory demands recoding but comprehension does not. If so, then we might conclude that lexical access will activate name codes just in case the reader's strategy is to have it so.

This would be a comfortable conclusion. Expert readers are good decoders and flexible word recognizers. When they need them, they generate and use name codes along with meaning codes. It might be necessary to expand this flexibility so that name code access could precede semantic access (difficult and rare words) or could follow it (high memory demands and comprehension obstacles). The problem with this is that it suggests a complex strategic component to reading when a simpler nonstrategic process would serve as well.

A less awkward model would assume that lexical access always activates phonemic codes. The only relevant strategic factor is whether a reader recodes in subword units and then uses that code to

consult meaning and to place it in the text representation. The attractive feature of this proposal is that the activated phonemic code is available for later memory scanning. A name code is thus available for securing reference. By this proposal, reading skill includes the rapid activation of all lexical information, including phonemic information. In any given situation, activation of phonemic information may precede or follow activation of semantic information depending upon the depth of semantic analysis required and the familiarity of the word. The issue then turns from whether speech recoding occurs to consideration of factors that control the activation time course of lexical properties.

## References


Anderson, J. R. Language, cognition and thought. Hillsdale, NJ: Lawrence Erlbaum, Inc., 1976.

Anderson, J. R., Greeno, J. G., Kline, P. J., & Neves, D. M. Learning to plan in geometry. Paper presented at 16th Annual Carnegie Symposium on Cognition, Pittsburgh, Pennsylvannia, 1980.

Atkinson, R. C., & Shiffrin, R. M. Human memory: A proposed system and its control processes. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 2) New York: Academic Press, 1968.

Bobrow, D. G., & Norman, D. A. Some principles of memory schemata. In D. G. Bobrow & A. Collins (Eds.), Representation and understanding: Studies in Cognitive science. New York: Academic Press, 1975.

Brown, J. S., & Van Lehn, K. Learning a procedural skill through exception conditions. Paper presented at 16th Annual Carnegie Symposium on Cognition, Pittsburgh, Pennsylvania, 1980.

Calfee, R. C. A Design for Examining Pattern Differences in Reading Abilities. Paper presented at American Educational Research Association annual meeting, Boston, Massachusetts, 1980.

Chase, W. G. & Simon, H. A. Perception in chess. Cognitive Psychology, 1973, 4, 55-81.

Coltheart, M., Besner, D., Jonasson, J. T., & Davelaar, E. Phonological encodings in the lexical decision task. Quarterly Journal of Experimental Psychology, 1979, 31, 489-507.

Curtis, M. E. Development of components of reading skill. Journal of Educational Psychology, in press.

Davelaar, E., Coltheart, M., Besner, D., & Jonasson, J. T. Phonological recoding and lexical access. Memory and Cognition, 1978, 6, 391-402.

Department of Computer Science, Carnegie-Mellon University. Speech Understanding systems: Summary of results of the five-year research effort at Carnegie-Mellon University. August, 1977.

Doehring, D. G. Acquisition of rapid reading responses. Monographs of the Society for Research in Child Development, 1976, 41 (2, Serial No. 165).

Erman, L. D., & Lesser, V. R. HEARSAY-II: Tutorial introduction and retrospective view. Technical Report No. CMU-CS-78-117. Carnegie-Mellon University, Department of Computer Science, 1978.

Frederiksen, J. R. A chronometric study of component skills in reading. Technical Report 2. Cambridge, MA: Bolt, Beranek, and Newman, 1978(a)

Frederiksen, J. R. Assessment of perceptual, decoding, and lexical and their relation to reading proficiency. In A. M. Lesgold, J. W. Pellegrino, S. D. Fokkema, & R. Glaser (Eds.), Cognitive psychology and instruction. New York: Plenum, 1978(b).

Glaser, R. Components of a psychology of instruction. Review of Educational Research, 1976, 46, 1-23.

Grice, G. R., Nullmeyer, R., & Spiker, V. A. Application of variable criterion theory to choice reaction time. Perception & Psychophysics, 1977, 22, 431-449.

Hogaboam, T., & Perfetti, C. A. Reading skill and the role of verbal experience in decoding. Journal of Educational Psychology, 1978, 70, 5, 717-729.

Hunt, E. The memory we must have. In Shank, R. C. & Colby, K. M. (Eds.), Computer Models of thought and language. San Francisco: W.H. Freeman and Company, 1973.

Joreskog, K.G., & Sorbom, D. LISREL: Analysis of linear structural relationships by the method of maximum likelihood. (Users Guide. Version IV, Release 2) Chicago: International Education Services, 1978.

Kahneman, D. Attention and effort. Englewood Cliffs, NJ: Prentice Hall, 1973.

Kleiman, G. M. Speech recoding in reading. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 323-339.

LaBerge, D., & Samuels, S. J. Toward a theory of automatic information processing in reading. Cognitive Psychology, 1974, 6, 293-323.

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

Larkin, J. H. Enriching Formal Knowledge: A model for learning to solve problems in physics. Paper presented at 16th Annual Carnegie Symposium on Cognition, Pittsburgh, Pennsylvania, 1980.

Lesgold, A. M., & Perfetti, C. A. Interactive processes in reading comprehension. Discourse Processes, 1979, 1, 323-336.

Lesgold, A. M., & Perfetti, C. A. (Eds.). Interactive processes in reading. Hillsdale, NJ: Lawrence Erlbaum Associates, in press.

Lesgold, A. M., Roth, S. F., & Curtis, M. E. Foregrounding effects in discourse comprehension. Journal of Verbal Learning and Verbal Behavior, 1979, 18, 291-308.

Levy, B. A. Vocalization and suppression effects in sentence memory. Journal of Verbal Learning and Verbal Behavior, 1975, 14, 304-316.

Liberman, I. Y., Shankweiler, D., Liberman, A. M., Fowler, C., & Fischer, F. W. Phonetic segmentation and recoding in the beginning reader. In A. S. Reber & D. L. Scarborough (Eds.), Towards a psychology of reading. The proceedings of the CUNY conference. New York: John Wiley, 1977.

Lomax, R. G. Testing a component processes model of reading comprehension development through linear structural equation modeling. Unpublished dissertation, University of Pittsburgh, 1980.

MacKay, D. G. Aspects of the theory of comprehension, memory and attention. Quarterly Journal of Experimental Psychology, 1973, 25, 22-40.

McClelland, James L. On the time relations of mental processes: An examination of systems of processes in cascade. Psychological Review, 1979, 86, 287-330.

McCulloch, W. S., & Pitts, W. H. A logical calculus of the ideas immanent in nervous activity. Bulletin of Mathematical Biophysics, 1943, 5, 115-133.

Miller, G. A. The magical number seven plus or minus two: Some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97

Morton, J. Interaction of information in word recognition. Psychological Review, 1969, 76, 165-178.

Morton, J. A preliminary model for language behavior. In R.C. Oldfield & J.C. Marshall (Eds.), Language. Baltimore: Penguin, 1968 (reprinted from International Audiology, 1964, 3, 216-225).

Morton, J., & Long, J. Effect of word transitional probability on phoneme identification. Journal of Verbal Learning and Verbal Behavior, 1976, 15, 43-52.

Newell, A. HARPY, production systems, and human cognition. In R. Cole (Ed.), Perception and production of fluent speech, Hillsdale,NJ: Erlbaum, 1980. by the method of maximum likelihood. Users Guide. Version IV, Release 2. Chicago: International Education Services,

1978.

Newell, A., & Simon, H. Human problem solving. Englewood Cliffs, NJ: Prentice Hall, 1972.

Perfetti, C. A., & Lesgold, A. M. Coding and comprehension in skilled reading and implications for reading instruction. In L. B. Resnick & P. Weaver (Eds.), Theory and practice of early reading. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1979.

Perfetti, C. A., & Lesgold, A. M. Discourse comprehension and sources of individual differences. In M. Just and P. Carpenter (Eds.), Cognitive processes in comprehension. Hillsdale, New Jersey: Lawrence Erlbaum Associates, 1977, 141-183.

Polson, P., Atwood, M. E., Jeffries, R., & Turner, A. The processes involved in designing software. Paper presented at 16th Annual Carnegie Symposium on Cognition, Pittsburgh, Pennsylvannia, 1980.

Rumelhart, D. E. Toward an Interactive Model of Reading. In Dornic, S. (Ed.), Attention and Performance VI: Proceedings of the Sixth International Symposium on Attention and Performance Stockholm, Sweeden, July 28-August 1, 1975. Hillsdale:NJ, Lawrence Erlbaum Associates, 1977.

Sternberg, S. Memory-Scanning: Mental processes revealed by reaction-time experiments. American Scientist, 1969, 57, 421-455.

Tulving, E., & Donaldson, W. Organization of Memory. New York: Academic Press, 1972.

Wernicke, C. Der aphasische Symptomenkomplex. Breslau: Cohn und Weigart, 1874.

Wernicke, C. The symptom complex of aphasia. In R. S. Cohen & M. Wartofsky (Eds.), Boston studies in the philosophy of science. Volume 4. Dordrecht: Reidel, 1966.

Wickelgren, W. Speed-accuracy tradeoff and information processing dynamics. Acta Psychologica, 1977, 41, 67-85.