DOCUMENT RESUME

ED 154 341 CS 004 061

AUTHOR Hoe, Alden J.; Hopkins, Carol J.
TITLE Parsing Word Strings from Text with a Computer: Implications for Reading Instruction.
PUB DATE May 78
NOTE 14p.; Paper presented at the Annual Meeting of the International Reading Association (23rd, Houston, Texas, May 1-5, 1978); For related document see CS 004 062

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage.
DESCRIPTORS Beginning Reading; *Computers; *Content Analysis; Elementary Education; *Phrase Structure; *Reading Instruction; Reading Skills; Sight Vocabulary; Syntax; *Word Frequency; *Word Recognition
IDENTIFIERS *Word Strings

ABSTRACT

Compilation of a list of the most common phrases used in reading was begun with the rationale that the quick recognition of phrases would facilitate reading comprehension. These first efforts showed that categorizing phrases by parts of speech did not provide acceptable levels of accuracy. The system that was effective, however, used a computer program that recorded every consecutive two- and three-word sequence in the text sample and determined which of these word strings recurred most frequently. The computer program makes possible samplings of large amounts of text-50,000 words or more-thus eliminating the idiosyncrasies of text sampling. The researchers who developed this system believe that the common phrases it identifies should be taught in much the same manner as common words are now taught in beginning reading instruction. (RL)

Alden J. Moe
205 Education
Purdue University
West Lafayette, IN 47907



# PARSING WORD STRINGS FROM TEXT WITH A COMPUTER: IMPLICATIONS FOR READING INSTRUCTION

Alden J. Moe and Carol J. Hopkins[1]

Purdue University

A paper presented at the twenty-third annual convention of the International Reading Association, Houston, May 4, 1978

The ability to quickly associate meaning with units of written language is considered crucial to the comprehension of text (Smith, 1971). Among the units of written language the reader must process are individual words, phrases, clauses, sentences and discourse structures. Word lists have been compiled for reading instruction with the criterion that the words be common, and that these common words be taught early in reading instruction; several such word lists have

[1]The authors gratefully acknowledge the help of Lee Congdon who developed the computer program discussed in this report and Robert Hieb who made many trial runs in the process of the program development.

become popular and are widely used by classroom teachers (Thorndike and Lorge, 1944; Dale and Chall, 1948; Carroll, Davies, Richman, 1971; Harris and Jacobson, 1972). List of common phrases (or common word strings), however, are not found in the research literature or in instructional materials (basal reader, manuals, workbooks, etc.) even though it is believed that the quick recognition of phrases will facilitate comprehension. The only available list of common phrases is the one compiled by Dolch (1948) thirty years ago.

The purpose of this report is to provide a rationale--or a justification--for the need to identify common word strings in text. This justification will touch upon some theories of language and/or reading processing and present our implications for reading instruction. In addition, we will describe some of the stages that brought us to the point where we felt we could actually parse word strings from text with a computer and identify the most common.

Both "word strings" and "phrases" have been used at this point to indicate word groups where the words appear together in text. A more precise definition of intra-sentence word groups such as phrases, clauses, and strings is deferred until a following section.

## Significance of the Problem

The more automatic the recognition of the chunks of language being read and the less effort expended on decoding, the greater the likelihood of complete comprehension. LaBerge and Samuels (1974) and Samuels (1976) refer this as automatic decoding or automaticity. Samuels (1976) states that "in order to have both fluent reading and good comprehension, the student must go beyond accuracy to automaticity in decoding" (p. 323).

In other words, the reader has a limited amount of cognitive energy, or ability, or memory with which to accomplish the reading task; the more cognitive energy used for decoding, the less for comprehension. The development of automaticity probably begins at the word level; but LaBerge and Samuels state that if the reader

> begins to organize some of the words into short groups or phrases as he reads, then further repetitions can strengthen these units as well as word units. In this way he can break through word-by-word reading and apply the benefits of further repetitions to automatization of larger units. (p. 315).

The importance of "phrase reading" over "word reading" is demonstrated by noting differences in the fixation length of naive and fluent readers. For example, first-grade children may make two fixations per word whereas high-school seniors make one fixation for about every two words (Taylor, Frackenpohl, and Patter, 1960). And in a study of third- and sixth-grade readers, Rode (1974-75) found that the eye-voice span was longer for the older readers suggesting that the older readers attempted to decode the larger units of meaning.

The work of Wisher (1976, 1977) provides further evidence that the reader uses his understanding of syntax "to parse word strings into convenient processing units" (p. 601). It is likely that understanding or semantic integration occurs between phrases and clauses (more likely clauses, but that discussion is beyond the realm of this paper). Further support is provided by Fodor and Bever (1965) who found that listeners group words (for understanding) according to the syntax of the sentence.

The importance of being able to read phrases has been discussed

by a number of reading educators including Bond and Tinker (1975), Harris and Sipay (1975), Heilman (1972), Heilman and Holmes (1972), and Zintz (1975); they believe that good readers organize the text they read into meaningful units such as phrases. However, many poor readers do not do this and comprehension is poor even when they have been pre-taught each individual word in the selection (Oaken, Weiner, and Cromer, 1971), and it has been found that training in the reading of phrases has improved the reading of remedial students (Amble, 1967).

Phrases, Clauses, and Word Strings

In our earliest efforts we were interested in identifying common, reoccurring phrases such as prepositional phrases.[1] For reasons which will be explained later, those efforts were unsuccessful so we resorted to identifying common word strings. At this point a discussion of what is meant by phrases, clauses and word strings is appropriate. In the sentence below there is a noun phrase (Little children) followed by verb phrase (were playing) which is followed by a prepositional phrase (in the park).

Little children were playing in the park.

The noun phrase and the verb phrase (Little children were playing) form

---

[1] While transformational grammar theory does not provide for the categorization of phrases according to parts of speech, we found the traditional labels useful. In transformational grammar, a sentence may be divided into a noun phrase and a verb phrase. Additional information in this area may be found in DeStefano (1978) and Jacobs and Rosenbaum (1968).

a main or independent clause. Any group of consecutive words (Little children, Little children were, children were playing, were playing in, playing in the, in the park, and so on) constitutes a word string.

We were--and are--primarily interested in common word groups which we expected to be true phrases according to a traditional grammarian's definition. However, a more appropriate descriptor for the word groups we identified is the term "word string" (or word strings).

Early Efforts at Parsing Phrases

Because we wanted to be able to analyze large amounts of texts (initially we felt at least 15,000 words), the application of computer technology was a critical part of our work. The fact that certain kinds of analyses may be accomplished through the use of computers has been demonstrated (Kucera and Francis, 1967; Carroll, Davies and Richman, 1971; Harris and Jacobson, 1972, Moe, 1973; and Hopkins and Moe, 1975). However, this study required programming of a somewhat different nature.

We identified five types of phrases (prepositional, participal, gerund, infinitive and verb) which commonly appear in written materials. Since it was anticipated that prepositional phrases could be identified by the computer with a high degree of accuracy we worked with a computer programmer to develop such a program. By programming the computer to locate all prepositions (with a list of 52 prepositions stored in the computer's memory) and then parse out the preposition and the two-word string which followed it, we found that indeed it was possible for the computer to identify these three-word strings with 99% accuracy. That is, we only missed about 1% of the prepositional phrases. A

problem arose, however, in that even though these three-word strings began with a preposition, they did not all function as prepositional phrases. We then eliminated prepositions from the list which rarely seemed to function as the first word in prepositional phrases. After many program revisions and trial runs we were able to parse out almost all (97-99%) of the prepositional phrases out of the text. However, we were still parsing out many word strings which were not prepositional phrases. And when we examined the strings which had been parsed out, only about 62% were actual prepositional phrases; we found this level of accuracy to be unacceptable.

Later Efforts

We then decided to approach the problem of identifying common phrases from a completely new perspective. Rather than categorizing phrases by parts of speech, another computer program was developed which identified every consecutive two- and three-word sequence found in the written text, store it in memory, and, at the end of all text input, tabulate all possible two- and three-word strings.

Through much trial and error, the investigators were able to develop a program that parsed out common word strings which are, by traditional definitions, actual phrases or which are the first two or three words of an actual phrase. Some of the common word strings identified, however, cannot be categorized by traditional definitions (and are, therefore, simply referred to as common strings). Once the new program was operational a corpus of 16,000 words analyzed previously with the old program, was reanalyzed. This analysis led us to decide that if we were going to make claims that we had identified common word

strings in written text, that many, many more samples of written text needed to be analyzed. In order to eliminate the idiosyncrasies of text sampling we believe that large amounts of text--over 50,000 words--should be used in subsequent analyses.

Major Implications

There appears to be little disagreement that the more able readers process larger chunks of text more rapidly than the less able readers. And it is agreed, we think, that our instructional practices should be such that our students are led to the point where they may, with a single fixation, read whole phases of two or three or four words. As to how children should be brought to this point, however, may be a debatable issue among reading educators. We believe that common phrases should be taught in much the same manner in which common words are taught and a suggested procedure is presented here.

If we know that "in the" and "of the", for example, are common word strings in text then it seems reasonable that they be taught as a group with a noun found in the text the students are to read. Since "in", "the", and "of" become part of a reader's sight vocabulary very early they will already be familiar to the student. The task is to get the student to read the function word(s) and the content word, which may or may not be a part of the student's sight vocabulary, quickly.

Assume, for example, that the student has a sight vocabulary of approximately 100 words and the the words "street" and "pond" are to be introduced as new words in a lesson. The concepts or the meaning of "street" and "pond" will be discussed with the student by the teacher. Then the teacher will present the printed form of the word

(either in isolation or in content). If the student is to become a rapid reader--and a rapid comprehender of text--then the reader should be able to read the phrases "in the street" and "in the pond" quickly since the meaning of the phrase is not in the word "in" or in the word "the" but primarily in the word "street" and more completely in the phrases itself. A similar case may be made for the presentation of larger chunks such as clauses and the procedures would be much the same.

Our purpose was to develop a system to identify common word strings. Since students must go beyond the word level in beginning reading, we believe that the use of common word strings found in text will facilitate the reader's ability to handle larger and larger units of text.

## REFERENCES


Amble, B. R. "Reading by Phrases," California Journal of Educational Research, 1967, 18: 116-124.

Bond, G. L. and Tinker, M. A. Reading Difficulties, Their Diagnosis and Correction, third edition, New York: Appleton-Century-Crofts, 1975.

Carroll, J. B., Davies, P., and Richman B. American Heritage Word Frequency Book. Boston: Houghton Mifflin, 1971.

Dale, E. and Chall, J. S. "A Formula for Predicting Readability," Educational Research Bulletin, Ohio State University, 1948, 27: 11-20; 28: 37-54.

De Stefano, J. S. Language, The Learner and the School. New York: John Wiley and Sons, Inc., 1978.

Dolch, E. W. Sight Phrase Cards. Champaign, Illinois: Garrard Publishing Company, 1948.

Fodor, J. A. and Bever, T. G. "The Psychological Reality of Linguistic Segments," Journal of Verbal Learning and Verbal Behavior, 1965, 4: 414-420.

Harris, A. J. and Jacobson, M. S. Basic Elementary Reading Vocabularies. New York: Macmillan, 1972.

Harris, A. J. and Sipay E. R. How to Increase Reading Ability, sixth edition. New York: David McKay, 1975.

Heilman, A. W. Principles and Practices of Teaching Reading, third edition. Columbus: Charles E. Merrill, 1972.

Heilman, A. W. and Holmes, A. H. Smuggling Language Into The Teaching of Reading. Columbus: Charles E. Merrill, 1972.

Hopkins, C. J. and Moe, A. J. "The Validation of a Synthetic Syllable Count Appropriate for Computer-Determined Readability Estimates." Paper presented at the meeting of the International Reading Association, New York City, May, 1975.

Jacobs, R. A. and Rosenbaum, P. S. English Transformational Grammar. Waltham, Massachusetts: Blaisdell Publishing Co., 1968.

Kucera, H. and Francis, W. N. Computational Analysis of Present-Day American English. Providence: Brown University Press, 1967.

LaBerge, D. and Samuels, S. J. "Toward a Theory of Automatic Information Processing in Reading," Cognitive Psychology, 1974, 6: 293-323.

Moe, A. J. "Word Lists for Beginning Readers," Reading Improvement, Fall, 1973, 10: 11-15.

Oaken, R., Weiner, M. and Cromer, W. "Identification, Organization and Reading Comprehension for Good and Poor Readers," Journal of Educational Psychology, 1971, 62: 71-78.

Rode, S. S. "Development of Phrase and Clause Boundary Reading in Children," Reading Research Quarterly, 1974-75, 10: 124-142.

Samuels, S. J. "Automatic Decoding and Reading Comprehension," Language Arts, March, 1976, 53: 323-325.

Smith, F. Understanding Reading: A Psycholinguistic Analysis of Reading and Learning to Read. New York: Holt, Rinehart and Winston, 1971.

Taylor, S. E., Frackenpohl, H. and Pattee, J. L. Grade Level Norms for the Components of the Fundamental Reading Skill. Bulletin #3, Huntington, New York: Educational Development Laboratories, 1960.

Thorndike, E. L. and Lorge, I. The Teacher's Word Book of 30,000 Words. New York: Teachers College Press, Columbia University, 1944.

Wisher, R. A. "The Effects of Syntactic Expectation During Reading," Journal of Educational Psychology, 1976, 68: 597-602.

Wisher, R. A. "Linguistic Expectations and Memory Limitations in Reading." Paper presented at the twenty-second annual convention of the International Reading Association, Miami Beach, May, 1977.

Zintz, M. V. The Reading Process, second edition. Dubuque: Wm. C. Brown, 1975.