ED 062 091 RE 004 082

AUTHOR Harris, Albert J.

TITLE Rationale and Description of "Basic Elementary Reading Vocabularies."

PUB DATE May 72

NOTE 12p.; Paper presented at the meeting of the International Reading Association, Detroit, Michigan, May, 1972

EDRS PRICE MF-$0.65 HC-$3.29

DESCRIPTORS *Basic Vocabulary; Comparative Analysis; *Computer Programs; Content Reading; *Elementary Grades; Reading Materials; *Textbooks; *Word Lists

ABSTRACT

The Harris-Jacobson Basic Elementary Reading vocabularies contain 7,613 words found to be basic in 14 series of textbooks for grades 1 through 6. They include a core list, an additional list, a technical vocabulary, and a total alphabetical list. Comparisons with the Dale list and the Botel list show very high degrees of overlapping with the appropriate part of the Harris-Jacobson list. Comparisons have also been made with the Taylor list, the Kucera-Francis list, and the American Heritage list. Despite marked differences in the techniques used in compiling some of these lists, there is substantial agreement about the words that are basic for reading in the elementray school. References are included. (Author/AW)

Albert J. Harris
52 Salisbury
Century Village
West Palm Beach, Florida 33401

International Reading Association Convention, Detroit, Michigan

Session: Word Lists for the 1970's
Friday, May 12, 3:00-4:00 P.M.
Crystal Ballroom, Sheraton-Cadillac

Rationale and Description of Basic Elementary Reading Vocabularies

In 1968 the time seemed ripe for a new elementary school reading vocabulary list. The lists in use were getting old. The Thorndike-Lorge and Dale lists had been issued in the 1940's. The Botel list, issued in 1962, was based on reading material of the 1950's and covered only the primary grades. Furthermore, the availability of natural-language computer technology made the undertaking seem practicable, since the computer could alphabetize words, merge lists, delete classes of words from a list, and perform many other operations that would be very time-consuming and costly if done by hand. With the computer, one could record and analyze the entire vocabulary of a book, and not have to rely on small samples. The resulting new vocabulary lists would be available on computer tapes as well as in print, and could be employed in a wide variety of research projects and practical applications.

The new vocabulary was planned to be basic, containing the words of greatest importance in elementary school reading material. To achieve this, a number of policy decisions were made. The list was to be based on textbooks, not on trade books or other optional reading matter. The textbooks were to include both basal readers and textbooks in English, mathematics, science, and social studies. The lists were to contain only words found in a substantial proportion of the series. Except at first grade levels, the lists were to consist of root words only, since the common inflected forms of nouns and verbs are learned today during the first and second grade reading levels. For the convenience of users, the vocabulary was to be presented both in a sequence of reading levels, starting with pre-primer, and in a total alphabetical order. It was decided to use three categories of words, core words, additional words, and technical words. These will be explained below.

Fourteen series of textbooks were chosen. There were six basal reader series, published by Allyn and Bacon, American Book Co., Ginn and Co., Houghton Mifflin Co., Macmillan, and Scott Foresman. There were eight series of content textbooks, two each in English, mathematics, science, and social studies. Each series covered grades one through six. The total number of books in the fourteen series was 127, and these 127 books contained about 4,500,000 running words.

The series to be used were chosen during the summer of 1969. The series were in wide use, were judged likely to remain popular for several years, and with one exception had been published during the preceding five years; the exception was a series for which the publisher was able to make available the vocabulary lists of a 1970 revision.

For all six of the basal reader series, lists of the new words that were introduced at each level were available for grades one through three, and for three of the series, for grades four through six also. For the eight content series, and for three of the intermediate reading series, lists of new vocabulary were not available. Each of those books was retyped word for word on IBM cards. Dr. Jacobson will explain the processes by which the words from 127 books were combined and reduced to the desired lists of basic words.

Since to the computer any group of characters preceded and followed by a space was a word, a combination of hand and machine operations was used to correct errors, and to delete several categories of entries. Certain inflected forms, such as plurals ending in -s or -es and verb forms ending in -s, -ed, and -ing, were combined with their root words. A root word and the inflected forms merged with it are called a unique word. Proper nouns, most hyphenated words, numerals, word parts, and misspelled words were eliminated. These steps took a great deal of time, since much of the work was done by personal inspection and had to be carefully checked.

The Basic Reading Vocabularies

The results of this project were published in April, 1972 under the title Basic Elementary Reading Vocabularies, by Harris and Jacobson. This contains several word lists.

The Core List is based entirely on the combined vocabulary of the six basal reader series. It contains the unique words which occur in at least three of the six series. The Core List is arranged by reading levels, and at each level the arrangement is alphabetical. There are three levels for the first grade (preprimer, primer, and first reader), and one level for each grade, two through six. The computer printout showed, for each unique word, all levels at which the word was used in each series. Each word was placed at the lowest level at, or below which, the word appeared in at least three different series.

It was anticipated that there would be quite a number of words used in less than half of the reader series, but found in enough of the content series to make them important for elementary reading. These make up the Additional List. Additional words are unique words found in fewer than three reader series, but which do appear in at least four of the fourteen series. An Additional word may appear in two reader series and two or more content series, in one reader series and three or more content series, or in no reader series but four or more content series. Additional words were placed at reading levels using the same criterion as for Core words. The Additional List, like the Core List, is arranged by levels and alphabetically within each level.

The Core List anu the Additional List together make up the General Vocabulary. There is also a Technical Vocabulary with sections for the four content areas. These are words which, although not in the Core list, are found in both series in a particular content area and are judged to have a technical meaning in that area of the curriculum. The Technical lists are small mainly because a large number of technical words appear in euough basal reader series to be Core words. For example, all of the following are Core words: appreciate, author, composition, descriptive; additional, diameter, fifteenth, triangle; apparatus, archaeologist, astronomer, atmosphere; citizen, civil, constitution. If Core words like these had been ruled eligible for the Technical lists, those lists would have been substantially larger.

The longest of the lists is the Total Alphabetical List, which contains in one alphabetical sequence all of the words in the separate lists. It also provides several kinds of information about each entry; the inflected forms merged with it; the other list or lists in which it appears; the reading level at which it is placed; and for each of the fourteen series of books, whether or not the word occurs in that series, and if so, the lowest level at which it is employed.

One example will show the kinds of information in the Total Alphabetical List. The unique word crack includes the inflected forms cracked, cracking, and cracks. It is in the Core List at third grade level. It appears in all six basal reader series, two at second grade level, two at third, one at fourth, and one at fifth. It also appears in both social studies series, one at third grade and one at fourth; in both science series, one at third and one at fourth; and in both English series, one at first grade and one at sixth.

Using the data given in the Total Alphabetical List one could, of course, set up other word lists using different criteria; for example, the words in all six of the reader series, or the words appearing in seven or more of the fourteen series. All of the information needed for applying such different standards are supplied in the Total Alphabetical List.

Number of Words in the Lists

There were more than 80,000 different entries in the original 127 computer printouts. The computer recognized as an entry any character or group of characters preceded and followed by a space. These entries were reduced in number in successive stages as errors were corrected, several kinds of ineligible entries were deleted, inflected forms were merged with root words to form unique words, and the criteria for the several lists were applied. The final result is a total of 7,613 unique words, with a Core List of 5,167 words, an Additional List of 1,699 words, and a Technical vocabulary of 805 words. The details by grade level are given in Table I. There are also 9,236 inflected forms.

Insert Table I about here

Considering the lists by levels, and combining the Core and Additional Lists, there are 58 preprimer words, 62 primer words, and 215 first reader words, totalling 335 words for the first grade. There are 577 second grade words and 1,012 third grade words, giving a total of 1,924 primary words in the General Vocabulary. There are 1,598 fourth grade words, 1,682 fifth grade words, and 1,662 sixth grade words, for a total of 4,942 intermediate grade words. Primary and intermediate combine to give a total General Vocabulary of 6,866 words. It can be seen, therefore, that the tendency in the basal reader series has been to develop reading vocabulary slowly in the first grade, accelerate during the second and third grades, and sustain a fairly steady rate of over 1,600 words a year in the fourth, fifth, and sixth grade reading materials.

Comparisons with Other Reading Vocabulary Lists

The Dale List of 3,000 Words, compiled in the 1940's, contains the 2,946 words which were marked as known by 80 percent or more of the fourth grade pupils tested with them. This list has been very widely used as it is one of the two components of the Dale-Chall Readability Formula. The Dale List merges inflected forms with root words much as the Harris-Jacobson Lists do. A computerized comparison of the Dale List with the Harris-Jacobson List shows that 93 percent of the Dale words are in the Harris-Jacobson General Vocabulary at or below the fourth grade level. Many of the differences are due to the obsolescence of some older words and the emergence of new widely used words. For example, words no longer widely used include afar, candlestick, codfish, fret, lass, sleigh, washtub. Current words not in the Dale List include TV, tractor, camera, experiment, astronaut, committee, hamburger. Reading vocabulary changes somewhat as new scientific terms become widely used and as ways of living and working change. However, the 93 percent agreement is impressive in showing that the main part of the reading vocabulary remains quite stable.

The Botel Bucks County List, published in 1962, contains 1,185 words found in at least three of six basal reader series for grades one through three. The series used were popular in the 1950's. Comparing the Botel List with the Harris-Jacobson words for grades one through three, one finds that 94 percent of the Botel words are in the Harris-Jacobson General Vocabulary. Thus although there is high agreement on the continued importance of nearly all of the Botel words, the Harris-Jacobson List contains 62 percent more unique words. This seems to indicate that between the 1950's and 1970 there has been a quite substantial increase in the number of different words used in primary grade reading programs. Most of this increase has been at second and third grade levels.

The Taylor List is a graded list published in 1969, which depends partly on word counts in nine series, and partly on words taken from older vocabulary lists such as the Thorndike-Lorge and Rinsland lists. The Taylor List includes two separate grade lists, one for grades one through eight, the other for grades nine through thirteen. They list 5,327 words for grades one through six as compared to the Harris-Jacobson total of 6,866 words for the same grades. A computerized comparison showed that 81 percent of the Taylor words for grades one through eight are also in the Harris-Jacobson General Vocabulary. Most of the differences are due to the presence of words in the Harris-Jacobson List that are not in the Taylor List, and some are due to differences in ways of treating inflected forms.

Two other recent vocabulary lists will be briefly discussed and some comparisons made, although computerized comparisons have not yet been made.

Kučera and Francis made a computerized analysis of the words in the Brown University Corpus, a sampling of one million running words of adult reading material. They found 50,406 items, a large number of which appeared only once. The words in their list are arranged in two ways, alphabetically and in order of decreasing frequency, and represent adult rather than child vocabulary. Nevertheless, it is interesting to investigate the degree of overlapping between the 1924 primary Harris-Jacobson words and the most frequent 2,000 words in the Kučera-Francis list.

Of the first 200 words in the Kučera-Francis list, 191 are in the Harris-Jacobson primary vocabulary (Core list through level three). Of the nine words unique to the Kučera-Francis list, _Mr_., _Mrs_., _American_, _F_, and _H_ were excluded from the Harris-Jacobson list because they are abbreviations, proper nouns, or non-word characters. The remaining words unique to Kučera-Francis, _public_, _however_, _general_, and _being_, begin to reveal the qualitative differences between these lists despite the 98 percent degree of overlap in these first 200 words. The Kučera-Francis list is based on adult reading materials in which words such as _general_ and _however_ could be expected to appear quite frequently, while the Harris-Jacobson list, based on elementary reading materials, reveals that these words are introduced at level four in primary reader series.

As the frequencies of the words in the Kučera-Francis list diminish, the differences between the lists become greater. In the last 400 words of the Kučera-Francis 2000 most frequent words, 48 percent are found in the Harris-Jacobson primary vocabulary. Around the 500 level of frequency rank in the Kučera-Francis list, unique words such as system, per, political, development, economic, and individual appear. In these levels of diminishing frequency, the decreasing degree in overlap between the lists reveals the qualitative differences between lists based upon, and describing, different types ot textual material. The overall degree of overlap between the lists is 59.3 percent, attributable to the difference between the adult vocabulary on which the Kučera-Francis list is based, and the elementary reading vocabulary on which the Harris-Jacobson primary vocabulary is based.

The American Heritage Word Frequency Book, published late in 1971, provides a word list based on 1,045 samples of 500 running words each, taken from a wide variety of materials including textbooks, workbooks, kits, novels, poetry, general nonfiction, encyclopedias, and magazines, for grade three through grade nine. The omission of materials for grades one and two and the inclusion of samples from grades seven, eight, and nine make a direct comparison with the Harris-Jacobson List difficult. The two lists not only cover different ranges of grades, but also are based on different principles. The American Heritage List is based on a large number of small samples; the Harris-Jacobson List is based on the entire vocabularies of a limited number of series of textbooks. The American Heritage List found 86,741 different entries and kept them all; thus it includes letters of the alphabet, numerals, chemical formulas, archaic and misspelled words, hyphenated words, proper nouns, and other kinds of items that have been eliminated from the Harris-Jacobson List. The American Heritage List also lists each inflected form of a word as a separate entry--crack, cracks, cracked, and cracking are four entries, while in the Harris-Jacobson List these have been merged into one unique word represented by the root word crack. The American Heritage List arranges the words alphabetically and according to descending frequency corrected for dispersion, while the Harris-Jacobson List arranges words according to reading grade levels.

Despite these differences, and the great difference in the number of items in the two lists, a quick comparison can be made. The Harris-Jacobson List contains 1,924 primary-level words. The fifty words occupying places 1875 to 1924 in the American Heritage List were studied. Only six of them were missing entirely from the Harris-Jacobson List. These included four proper nouns, one alphabet letter, and one three-place number. Of the remaining 44 words, 25 are Harris-Jacobson root words and 19 are inflected forms not listed separately in Harris-Jacobson. Thus 88 percent of these fifty words are common to the two lists. Another sample of 50 American Heritage words was taken preceding the word ranked 7,613, which is the rank corresponding to the total number of Harris-Jacobson words. Forty of these, or 80 percent, were also in the Harris-Jacobson List, 27 as root words and 13 as inflected forms. Of the ten words not in common, two were numerals, five were proper nouns, and three were words not found. Thus even at this level, which represents an average frequency of only one appearance per 125,000 or so words, there is a fairly high overlapping between the two lists.

Summary

The Harris-Jacobson Basic Elementary Reading Vocabularies contain 7,613 words found to be basic in fourteen series of textbooks for grades one through six. They include a Core List, an Additional List, a Technical

Vocabulary, and a Total Alphabetical List. Characteristics of these lists have been described. Comparisons with the Dale List and the Botel List show very high degrees of overlapping with the appropriate part of the Harris-Jacobson List. Comparisons have also been made with the Taylor List, the Kučera-Francis List, and the American Heritage List. Despite marked differences in the techniques used in compiling some of these lists, there is substantial agreement about the words that are basic for reading in the elementary school.

References

Botel, Morton. Botel Predicting Readability Levels. Chicago: Follett Publishing Co., 1962.

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

Dale, E., and Chall, Jeanne S. "A Formula for Predicting Readability," Educational Research Bulletin (Ohio State University) 27 (Jan. 21 and Feb. 17, 1948), pp. 11-20, 37-54.

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

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

Taylor, S. E., Frackenpohl, Helen, and White, Catherine E. A Revised Core Vocabulary: A Basic Vocabulary for Grades 1-8; An Advanced Vocabulary for Grades 9-13. Research and Information Bulletin No. 5 (revised). Huntington, N.Y.: Educational Developmental Laboratories, 1969.

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

Table I

Number of Words in the Harris-Jacobson Lists

List	Words
General Vocabulary	
Core, Preprimer	58
Core, Primer	62
Core, First Reader	211
Additional, First Reader	4
Core, Second Reader	552
Additional, Second Reader	25
Core, Third Reader	881
Additional, Third Reader	131
Core, Fourth Reader	1,196
Additional, Fourth Reader	402
Core, Fifth Reader	978
Additional, Fifth Reader	704
Core, Sixth Reader	1,229
Additional, Sixth Reader	433
Total, Core	5,167
Total, Additional	1,699
Total, General Vocabulary	6,866
Technical Vocabulary	
English	106
Mathematics	86
Science	359
Social Studies	254
Total, Technical Vocabulary	805
Total Alphabetical List	7,613