DOCUMENT RESUME

ED 062 102 RE 004 120

AUTHOR	Jacobson, Milton D.
TITLE	Developing and Comparing Elementary School Word Lists by Computer.
PUB DATE	May 72
NOTE	14p.; Paper presented at the meeting of the American Educational Research Association, Chicago, Ill., Apr. 1972
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	Basic Reading; *Computational Linguistics; Information Processing; Instructional Materials; Textbooks; Vocabulary; *Word Lists

ABSTRACT

The procedures for compiling a new elementary word list using computers are described. Words were taken from 127 books in fourteen series of widely used elementary textbooks. The compilation procedures consisted of (1) input: putting the lists into the computer, (2) processing of the vocabulary into compiled lists, (3) output: production of the actual word lists. Rules set up to determine whether inflected forms of words would be included are described. Capitalized proper nouns, abbreviations, word parts, and hyphenated words were deleted. Scanning programs were used to correct and proofread initial lists. The processing of the words resulted in four kinds of lists: (1) the Core List (words which were included in three or more of the six reader series), (2) the Additional List (words found in four or more different series excluding Core words), (3) four Technical Lists, and (4) a Total Alphabetical List in which all the lists were merged and put in alphabetical order. A comparison between this list and four other word lists is made. Sample printouts, tables of data, and references are included. (AL)

Milton D. Jacobson
Bureau of Educational Research
106 Peabody Hall
University of Virginia
Charlottesville, Virginia 22903

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

Developing and Comparing Elementary School Word Lists by Computer

I. Compilation Procedures

The Harris-Jacobson word list (1972) is based on a computerized analysis of the total vocabulary content of 127 books in fourteen recently published and widely used series of elementary school textbooks. Since the fourteen series include six in reading, and two each in English, mathematics, science, and social studies, the vocabulary constitutes a rich variety of wordstock providing large numbers of general and technical vocabulary words which do not occur in most existing word lists. In addition, the inclusion of all of the books of six newer reading series which reflect the trend toward less exacting control over basal reader vocabulary increased the likelihood of obtaining words not in existing word lists. Thus the lists derived from these 14 series should have many words in common with other word lists but should also have many new and different words which the less comprehensive or older lists do not have.

The words determined to be the basic essential vocabulary for elementary reading were organized into a General List, a Technical List, and a Total List through a series of computer processes. These procedures may be defined conceptually as

1) input, getting the lists into the computer, 2) processing of the vocabulary into compiled lists, and 3) output, or production of the actual word lists.

Before work compiling the lists could proceed, two sets of rules had to be established. One set governed the situations in which inflected forms were or were not to be merged with their root words, the other set established which words were deleted. At the preprimer level roots were combined with plural inflections (root word plus *s*). Words at the primer level included root words plus -*s*, -*es*, -*'s*, -*d*, -*ed*, -*er* (comparative). At the first reader level, the rule was the same as that for the primer level with the addition that -*ing* and -*est* endings were listed with root words. At the second grade level all first grade variants were listed plus variants with the endings -*ed*, -*ing*, -*er*, and -*est* which follow a doubled consonant, variants which change *y* to *i* before adding -*ed*, -*er*, -*es*, or *est*, and variants ending in -*y*, -*ey*, and -*ily*. Variants at levels three and up were the same as those included at grade two. Variants occurring at a level lower than the level at which such variants were procedurally included were included according to the frequency criteria of root words. Variants dropping -*e*,before adding -*y* (bone, bony; rose, rosy) were treated as unique words. Variants ending in -*er* were classified as comparatives, agents, or root words by personal judgment.

The other set of rules established which classes of words were deleted. Capitalized proper nouns were deleted, as were abbreviations and word parts which appear in textbook reader and

English lessons. Hyphenated words were deleted except where their meaning can not be easily inferred from the meaning of the joined root words (good-by, tom-tom).

The first step in compiling the lists was input, or getting the words from the books into the computer. When the publisher provided a list of the words new to the series, the list was typed in sequence on IBM cards. This was true for all of the primary-grade readers and half of the intermediate-grade readers. When such lists were not available (the other half of the intermediate readers, and all of the content textbooks), every word in the book was typed .1 sequence either on IBM cards or on photosensitive, machine-readable paper in machine-readable type. From the cards or paper the data were fed into a computer and registered in memory tapes. A comparative study showed the IBM card procedure to be the less costly, because the photosensitive paper required several intermediate machine operations which were expensive.

The word listsfor each book was alphabetized by the computer. The resulting printout was then corrected by a series of four procedures which ensured that erroneous entries were reduced to an absolute barest minimum. Initial text corrections were made by a single oral proofreading, found to be much faster than machine verification on a keypunch verifier and capable of discovering 2/3 of the errors in the first reading. Since this oral proofreading process required 27 hours of clerical time per 100,000 word book, and there were 127 books, repetitions of such proofreadings were considered inefficient.

The second correction procedure utilized a new computer program which greatly reduced the manual labor required. This program is based on the existing Key Word In Context (KWIC) programs. As it is a specialized, abbreviated adaptation it was entitled "Quickie."

The Quickie program scans=input text and produces a reedited and sequenced file consisting of IBM card images (these images are two-thirds the length of a line of 120 spaces of ordinary computer printouts). This file is printed by the computer. Every line on the computer printout is numbered in sequence and consists of the exact textual data as punched on one IBM card.

Once the card image printouts have been printed, the Quickie program uses this file to reduce to a fraction the material to be proofread.

The body of unique words subject to proofreading and correction can be further reduced by comparing, by computer, the text to a core-memory dictionary of common words stored in the computer. Approximately 60% of the running words in textual material are among Thorndike's 1000 most common words. If these words include variants to make a 3000 word dictionary, a single scanning operation by the computer will reveal that only 5% of the 100,000 running words in the fifth-grade text are not in the dictionary and thus require visual verification. Of these 5,000 words approximately 250 were identified as possibly incorrect and were referred to in context. Almost all of the 250 words required correcting.

The third correction operation was a visual scanning of corrected texts, after which the word lists were generated. Finally, the lists were scanned by the authors and odd-looking words were verified or corrected.

Though the input text was punched on IBM cards, the processing system is able to accept data on paper tapes, magnetic tapes, or photosensitive paper, enabling researchers to use packaged instruction programs, or other texts such as AP-UPI tapes available on such input media, in studies which implement the processing procedures used in compiling this wordlist.

After correction of all of the input data, the second or processing stage was conducted. The computer merged all the words from all the basal readers, from pre-primer through grade six, into one alphabetical sequence. This is done by a scan-and-sort computer operation which alphabetizes the words and indexes their frequencies and levels of appearance into one list of unique words. Each word was accompanied by information which showed each book in which it appeared, making it easy to note the lowest book in which it first was used in each series.

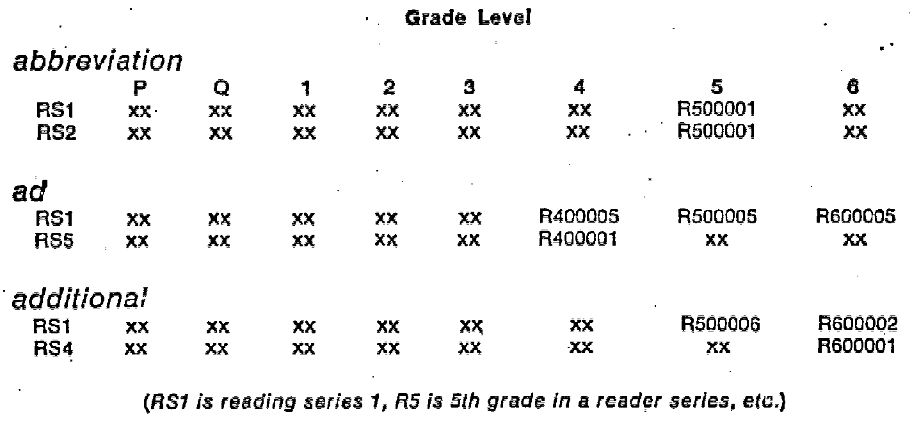
These listings were then printed to obtain a master file of all unique words found in the reading series. This file gave unique words and listings for over 2,000,000 running words. Figure 1 illustrates these listings.

At this point the rules for merging variants with roots, and for deleting certain classes of words were applied.

The criteria for inclusion in the Core List were then applied and the words which qualified were marked. Words which appear

Figure 1

An Example of the Information Contained in the Reading Series Master File Printout



Grade Level

	P	Q	1	2	3	4	5	6
abbreviation								
RS1	xx	xx	xx	xx	xx	xx	R500001	xx
RS2	xx	xx	xx	xx	xx	xx	R500001	xx
ad								
RS1	xx	xx	xx	xx	xx	R400005	R500005	R600005
RS5	xx	xx	xx	xx	xx	R400001	xx	xx
additional								
RS1	xx	xx	xx	xx	xx	xx	R500006	R600002
RS4	xx	xx	xx	xx	xx	xx	xx	R600001

(RS1 is reading series 1, R5 is 5th grade in a reader series, etc.)

in three or more of the six reader series were included in the Core List. The Core List was copied out, verified, typed on IBM cards, and entered into the computer.

The next step involved two operations, adding all of the words from the content books to the basal reader list, and deleting all Core words from that list. The resulting alphabetical list provided the raw material for the Additional List and the four Content lists. Variants were merged and deletions made again.

The Additional List, consisting of words found in four or more different series (excluding Core words), was then selected by research assistants and reviewed by the authors. With the Additional List available, the alphabetized word list for each content area was gone over and those words which satisfied the criteria for the particular content area were marked and verified. The four Technical Lists were copied out and entered into the computer.

At this point, all the data needed for the Total Alphabetical List had been assembled. A series of computer operations merged all of the separate lists into the Total Alphabetical List of 7,613 words, 16,849 when inflected forms are included. To do this, each word appearing in at least one of the component lists (Core, Additional and Content) was listed.

After completing the processing of the lists the third stage or computer printout was made. Figure 2 illustrates this printout. The Total List presents information about the list in which the word appeared such as Core, Additional, or Content and identified each series (reader or content) and level in which the word appeared. Because of the rules for inclusion of inflected forms, the Total Alphabetic List contains all unique words, lists their inflected forms, and lists the stipulated special inflected forms as unique words.

In addition to containing all of the unique words that are in each of the other lists, the Total Alphabetical List provides for each word all of the essential information used in assigning the words to the respective lists.

Figure 2

TOTAL WORD LIST

WORD+ENDINGS	LISTED	READFR	EN	MA	SC	SS
A	[CP]	PPPPPP	11	11	11	11
ABACUS	[MA]	..6...		23		
ABANDON ED ING	[C5]	545566	5.			56
ABBOT S	[C6]	6.5..3	14			
ABBREVIATION S	[A5 EN]	55....	22	.6	.6	66
ABILITY IES	[C4]	454554	55		44	56
ABLE ST	[C2]	323232	12	34	32	33
ABOARD	[C3]	3.34.3	12		53	54
ABOLISH ED	[SS]					55
ABOUND ING S	[A6]	5....6	6.			5.
ABOUT	[C0]	011100	11	11	11	21
ABOVE	[C2]	232422	11	32	22	22
ABREAST	[C6]	66..6.				
ABROAD	[C5]	4.55.6	35			55
ABRUPT LY	[C5]	6.5654	66		.5	5.
ABSENCE S	[C6]	6.56..	56	44	.5	.5
ABSENT LY	[C6]	6..66.	34	43		
ABSOLUTE LY	[C5]	4.3655	63		55	66
ABSORB ED S	[C5]	4.5.65	55		45	.6
ABSURD	[C6]	4.6665				
ABUNDANCE	[A5]	5.....			64	56
ABUNDANT LY	[C6]	4.6.65	6.		65	55
ABUSE D ING	[A6]	4....5	6.			6.
ACADEMY IES	[C6]	.6.6.5	64			55
ACCENT ED S	[C5]	66534.	54			
ACCEPT ED ING	[C4]	463653	15		55	55
ACCEPTABLE	[EN]		55			
ACCIDENT S	[C3]	34324.	23	.5	64	35
ACCIDENTAL LY	[A5]	5.4...	5.		6.	5.
ACCOMPANY IED IES	[C5]	445.65	.65			56
ACCOMPLISH ED	[C6]	4..65.	55		65	55
ACCOMPLISHMENT S	[C6]	.656.5	56		6.	6.
ACCORDING LY	[C4]	454.44	33	54	44	43
ACCOUNT ED S ING	[C4]	46463.	35	54	45	45
ACCURACY	[A6]	6.....	6.		65	6.
ACCURATE LY	[C5]	465.65	35	56	44	45
ACCUSE D S ING	[C4]	4..534	55			55
ACCUSTOM ED	[C6]	..5665	5.		55	5.
ACHE D S ING	[C4]	44344.	44		6.	64
ACHIEVE D S ING	[C6]	46.66.	.5		65	56
ACHIEVEMENT S	[C5]	5.4..4	66		64	54
ACID S	[A5 SC]	6.....	5.		54	55
ACKNOWLEDGE D ING	[C6]	4...66				6.
ACORN S	[C4]	2.4.3.			42	62
ACQUAINT ED	[C6]	46466.	66	.4	54	45
ACQUIRE D S ING	[C6]	6.5.65	5.		65	55
ACRE S	[C4]	54544.			66	35
ACREAGE	[C6]	.666..				
ACROBAT S	[C6]	..6.64		5.		
ACROSS	[C2]	212122		22	12	22
ACT ED ING S	[C2]	232232	13	.5	34	13
ACTION S	[C4]	4444.3	14		54	54
ACTIVE LY	[C5]	565..4	43		53	55
ACTIVITY IES	[C5]	5643.5	51		33	53

WORD+ENDINGS	LISTED	READER	EN	MA	SC	SS
ACTOR S	[C6]	6..43.	35		6.	15
ACTRESS ES	[C6]	5.56..	45			
ACTUAL LY	[C4]	464454	43	55	44	45
AD S	[A3]	4...4.	34	23		
ADAPT ED S	[C6]	44.6..	.2		44	66
ADD ED ING S	[C2]	243223	11	21	22	22
ADDEND S	[MA]			11		
ADDITION S	[C5]	4.535.	53	11	55	54
ADDITIONAL	[A5]	5..6..	53		45	6.
ADDRESS ED ES ING	[C3]	244352	12	.3	5.	53
ADEQUATE	[A6]	6	6.		6.	.5
ADJECTIVE S	[A5 EN]	5....6	43			
ADJOIN ING	[A6]	6.6...			.5	.6
ADJUST ED S	[C4]	454.65	.6		64	
ADMIRAL S	[C5]	56455.	55		64	55
ADMIRATION	[C4]	4644.3	1.			
ADMIRE D S ING	[C4]	444463	13			54
ADMIT S TED TING	[C4]	354434	65	.5	.4	54
ADOBE	[A5]	..6.5.				55
ADOPT ED ING S	[C5]	5.4.56	55		.6	55
ADORE D	[C6]	4.466.	4.			
ADULT S	[C4]	4.464.	4.	33	23	6.
ADVANCE D ING S	[C5]	45566.	55	.4	54	55
ADVANTAGE S	[C5]	465465	.6	6.	64	44
ADVENTURE S ING	[C3]	33.342	12		46	43
ADVENTUROUS	[A5]	4..5..	5.		4.	5.
ADVERB S	[EN]	5.....	45			
ADVERTISE D S ING	[C5]	5.5555	35		6.	45
ADVERTISEMENT S	[A5]	5.....	55			56
ADVICE	[C4]	443450	36		55	55
ADVISE D ING	[C5]	55566.	5.		55	65
ADVISER S	[A6]	5..6..	.6			65
AERIAL S	[A5]	4.3...	6.		.5	.5
AFFAIR S	[C4]	465445	66			55
AFFECT ED ING S	[C5]	5643.5	34	34	43	44
AFFECTION S	[C5]	565..4	56			6.
AFFECTIONATE LY	[C5]	3.5.4.	.6			5.
AFFIX ES	[EN]	5.....	65			
AFFLICT ED	[A6]	..5..6				66
AFFORD	[C4]	46434.	6.			45
AFLOAT	[A6]	5.....	65		6.	.6
AFRAID	[C2]	222222	18		4.	35
AFTER	[C1]	211102	11	21	12	11
AFTERNOON "S S	[C3]	332432	12	32	.2	33
AFTERWARD S	[C5]	5.34..	55			54
AGAIN	[C1]	211110	11	33	11	31
AGAINST	[C3]	232333	12	64	32	39
AGE D S	[C3]	444339	12	43	24	23
AGENCY IES	[A5]	4.....	3.		.5	55
AGENT S	[C5]	.45655	55			
AGILE	[A6]	6.6...	5.			
AGO	[C2]	1322.2	12	33		[illegible]
AGONY	[C6]	4666.4				
AGREE D ING S	[C3]	243233	13	63	34	33

II. Comparison Procedures

A computer program capable of comparison of word list content seems useful for a variety of reasons. Most obvious is facilitation of comparison of word list content according to criteria of range, scope, or form of words which should be included. A more subtle application might be the comparison of lists and the materials constructed with them in order to identify differences created by the passage of time, or some other factor.

Some of the lists in widespread use today were developed as many as fifty years ago. A computerized comparison procedure allows one to evaluate the differences between old lists and modern ones according to criteria of obsolescence in vocabulary. In effect, the process of aging can be isolated and identified, making the evaluation of the usefulness of old lists and the materials which they were used to develop a feasible task. As new lists are developed, their content can be compared, allowing users to evaluate the relative usefulness of one or another.

The procedure used to enable an automated comparison of word list content involved the punching of several lists onto IBM cards, then programming the computer to sort the words, compare them for correspondence, check for correspondence or variation in level assignment, and print out the results in verbal form. This has been done in a comparison of the Harris-Jacobson Basic Elementary Vocabularies (1) with the Dale list of 3,000 words (2), the Botel list (3), and the Taylor list for grades 1-8 and grades 9-13 (4). The words were punched sequentially, separated by commas or spaces and followed by level information.

The computer processing can be broken into two stages. The first stage receives and stores the raw data of the lists, automatically alphabetizing the words. This stage of the program forms a file constituting a single list of the words contained in all the lists, in effect merging the lists to be compared. Every word contained in the lists is recorded once in alphabetical order. Each word is accompanied by a mask 96 columns long, allowing the recording of 96 pieces of information for each word, such as the lists in which it appears. These columns could be alotted so as to record level assignments or other categorizations made by Harris-Jacobson and compilers of the other lists. For instance, the Harris-Jacobson list is composed of Core, Additional, and Content vocabularies, and the Core and Additional vocabularies are stratified by grade level. Thus, the columns of the mask could be alotted so as to indicate the composite list and/or the grade level in which a word appears.

The next group of bits could be alotted to the next list, broken down according to its assigned levels or categories and so on. The file composed by this first stage of the program incorporates facilities for generating new information, for updating, or for correction of the existing data.

The second stage of the program reads through the file compiled by the first stage, and prints and tallies the merged lists. This printer stage of the program inputs a list of the potential titles to be sought in the mask of the stage-one file, checks the columns for the requisite information, and prints the words with the appropriate titles. The result is a listing with all the words contained in all the word lists appearing in alphabetical order along the left margin. Next is a space in which the presence or absence of the word in the master list can be noted. To the right the comparison list in which the word appears are shown. The print thus records the unique words of each list, the words which appear in more than one list and where they are matched, and records level information for each word if such information is provided by the compilers of the list. This print-out can be easily read, and the nature of the matched and unmatched words can be observed.

In addition to the print out of the merged and compared lists, the program tallies information about the results, such as the number of words in both of two lists, the number of words in one list not in the other, the number of matched words which have been assigned to the same level by both compilers, or similarly, different levels. Categorical information supplied by the compilers can be noted as criteria in the comparison. Further, the program can print out a list of matched words without unmatched words, or the unmatched words form any list without the matches.

The data for the study consisted of four word lists. The first was the Harris-Jacobson Basic Elementary Reading Vocabulary recently developed by Albert Harris and myself (1). The H-J computer list for this study includes foth the Harris-Jacobson 7,613 root words and 9,237 inflected forms, totalling 16,850 entries. This list was compared to three other word lists:

the Dale list of 3,000 common words developed by Edgar Dale (2), the Botel Bucks County list of 1,185 common words developed by Morton Botel (3), and the EDL vocabulary developed by Stanford Taylor and others (4). The EDL vocabulary was broken into two sublists which were compared independently, one for levels 1-8 and one for levels 9-13. The results of the comparison are shown in Table 1.

Of the 2,946 words in the Dale list, 2,744 or 93 percent also appear in the Harris-Jacobson List. Of the 3,266 words in the Botel List (including inflected forms), 3,095 or 94 percent are also in the Harris-Jacobson List. Thus the overlapping among these three lists is quite high. The degree of overlapping with the two Taylor lists is lower. Of the 6,714 Taylor words for grades one through eight, 5,473 or 81 percent are also in the Harris-Jacobson list. This is not surprising, since the Harris-Jacobson list stops at sixth grade and the Taylor list includes seven and eight. The Taylor high school list shows still less overlapping.

While these tallies are interesting, the output of this comparison program provides a means for a detailed content analysis to discover the reasons for differences or overlap between texts. The matched and mismatched words can be scrutinized to ascertain what factors or features of the various lists might explain the results of a comparison.

TABLE I

COMPARISON OF THE HARRIS-JACOBSON BASIC ELEMENTARY READING VOCABULARY WITH FOUR OTHER WORD LISTS

	LIST BEING COMPARED			
	Dale List	Botel List	Taylor (1-8)	Taylor (9-13)
Total Number of Words in Harris-Jacobson List	16,849	16,849	16,849	16,849
Total Number of Words in Comparison List	2,946	3,266+	6,714	2,426
Number of Words in Harris-Jacobson That Are Not in Comparison List	14,105	13,754	11,376	16,670
Number of Words in Both Lists	2,744	3,095	5,473	179
Number of Words in Comparison Not in Harris-Jacobson	202	171	1,241	2,247

*Harris and Jacobson, Basic Elementary Reading Vocabularies
Of the 16,849 entries, 7,612 are root words in the published lists and 9,237 are inflected forms not printed as separate entries.

+Basically 1,185 words. When separate entries are made for each variant form it consists of 3,266 words (example: beat, beats, beating).

NOTES

1. Harris, Albert, and Milton Jacobson, Basic Elementary Reading Vocabularies. New York, Macmillan, 1972.

2. Dale, Edgar and Jeanne S. Chall, "A Formula for Predicting Readability: Instructions, "Educational Research Bulletin, Vol. XXVII, No. 2, February 17, 1948.

3. Botel, Morton, Botel Predicting Readability Levels, Chicago: Follett, 1962.

4. Taylor, Stanford E., Helen Frackenpohl, and Catherine E. White, A Revised Core Vocabulary: A Basic Vocabulary for Grades 1-8, and Advanced Vocabulary for Grades 9-13. Huntington, New York: McGraw-Hill, Educational Development Laboratories, 1969.