Some comparisons between
Basic Elementary Reading Vocabularies
and other word lists

ALBERT J. HARRIS
The City University of New York
and
MILTON D. JACOBSON
University of Virginia

Describes 13 reading vocabulary lists and draws comparisons between the Harris-Jacobson Basic Elementary Reading Vocabularies and the other 12 lists. A high degree of agreement was found on the most common 2,000 words, corresponding to the H-J words for the first 3 grades. For grades 4 through 6 the overlapping among lists drawn from children's materials was a little lower, but still high. However, nearly half of the words in a sample of the Kucera-Francis adult list at frequencies corresponding to upper sixth grade are quite uncommon in elementary reading materials. There is considerably less agreement on the reading levels and relative frequencies of specific words than on their presence or absence. The data presented are discussed both for what they show about each of the 12 comparison lists, and as casting light on the validity of the H-J list which served as the criterion.

Quelques comparisons entre les Vocabulaires Fondamentaux de la Lecture Élémentaire et d'autres glossaires

DECRIT 13 LISTES de vocabulaire et compare le Harris-Jacobson Vocabulaire de la Lecture fondamentale élémentaire à 12 autres glossaires. On constate un haut degré de conformité parmi ces divers listes lorsqu'il s'agit des 2,000 mots les plus communément usités. Ces mots correspondent à ceux que l'on trouve dans le glossaire de H-J et on les rencontre surtout dans les classes primaires (1-3). Quant aux classes qui vont de la 4ème à la 6ème, la conformité reste importante bien que sensiblement moins grande. Cette observation a été faite d'après une étude
comparative du vocabulaire tiré des lectures effectuées par ce groupe. Cependant, presque la moitié des mots tirés d'un échantillon du vocabulaire pour adultes de Kucera-Francis (à des fréquences conformes au niveau avancé de la 6ème année) n'apparaissent que rarement dans les textes de lectures élémentaires. Il y a considérablement moins de conformité au sujet des mots spécifiques par rapport aux niveaux de lecture et à leur fréquences relatives qu'il n'y avait lorsque l'on ne considérait simplement leurs présences ou leurs absences. L'étude entreprend de discuter ces faits parce qu'ils peuvent révéler des aspects importants de chacune des 12 listes et parce qu'ils démontrent la valeur du glossaire dressé par H-J, qui a servi de critère à cette étude.

**Algunas comparaciones entre el Vocabulario de Lectura Basic Primario y otras listas de palabras**

**Describe 13 listas de vocabulario de lectura** y establece comparaciones entre el "Harris-Jacobson Basic Elementary Reading Vocabularies" (Vocabulario de Lectura Básico Primario de Harris-Jacobson) y las otras 12 listas. Se halló un alto grado de acuerdo en las 2.000 palabras más comunes correspondientes a las palabras de H-J para los 3 primeros grados. Para los grados 4 al 6, la sobreposición entre las listas (basadas en materiales para niños) resultó un poco menor, aunque todavía elevada. No obstante, casi la mitad de las palabras de un modelo de la lista para adultos de Kucera-Francis (frecuentemente correspondientes al sexto grado superior), son bastante poco comunes en los materiales de lectura primaria. Existe un acuerdo considerablemente menor en las palabras específicas para los niveles de lectura y las frecuencias relativas, del que existía simplemente entre su presencia o ausencia. Los datos presentados son discutidos tanto por lo que muestran sobre cada una de las 12 listas de comparación, como por la claridad que arrojan sobre la validez de la Lista de H-J, utilizada como criterio.
Since the publication of the *Teacher's Word Book of 10,000 Words* (Thorndike, 1921) over 50 years ago, the development of vocabulary lists has stimulated a large volume of research. In a revised *Bibliography of Vocabulary Studies* (Dale, Razik, and Petty, 1973), over 3,000 references are listed.

One of the major areas for application of vocabulary lists has been in readability research; such widely used readability formulas as those by Dale and Chall (1948), Lorge (1944), and Spache (1953) have each used a vocabulary list as their most important component. Lists have also been used in many comparisons of reading materials in regard to vocabulary control, vocabulary difficulty, and overlapping of vocabulary among books intended to cover the same ground. Another major use of vocabulary lists has been as references for authors, especially those trying to write for readers of a particular level of reading competence. A list may also be used as a source for vocabulary test items.

It seems significant, then, to compare the recently published *Basic Elementary Reading Vocabularies* (Harris and Jacobson, 1972) with a number of existing vocabularies. Some of the comparisons are computerized comparisons of entire lists of many thousands of words. Others are comparisons of selected small samples. The lists involved in the comparison include some based on adult, secondary, elementary, and primary reading materials. Both comprehensive lists and very short lists for special purposes, such as minimal core lists for beginning reading or remedial work, have been examined.

### Computerized comparisons of total word lists

A computerized procedure was developed for comparing large word lists. First each list was punched onto data cards, with such information as reading level punched after the word. Each of the word lists was entered onto memory tapes, and then the words were automatically alphabetized. Thus a single alphabetized list of the words contained in all of the lists was created—in effect, merging the lists. The printout shows the complete alphabetical list down the left-hand side. To the right of this is a column for the criterion list; words from that list are printed there along with such data as assigned level; if words in the complete list are not in the criterion list, the space beside them in the latter list is blank. To the right, each comparison list in which the word appears is shown along with
information such as the word's placement or frequency as indicated by its compilers.

The printout thus records the unique words of each list and the words which appear in more than one list. It also shows where they are matched. This printout can easily be read and the nature of the matched and unmatched words observed. The program also tallies information such as the number of words in both of 2 lists, the number in each list but not in the other, the number of matched words assigned to the same levels in 2 lists, etc. The program can also print a list of matched words or a list of unmatched words. This procedure was used in comparing the Harris-Jacobson Basic Elementary Reading Vocabularies with 3 other lists.

The Harris-Jacobson Basic Elementary Reading Vocabularies is based on a computerized analysis of the total vocabularies of 14 series of textbooks for grades one through 6; of 6 series in reading, and of 2 series each in English, mathematics, science, and social studies. These series contained 127 books and about 4,500,000 running words. All of the words in all of these books were entered into memory tapes and combined into one total alphabetical list. After eliminating errors, inflected forms—such as plurals ending in s or es and verbs ending in d, ed, s, or ing—were merged with their root words. Three sets of lists were then extracted. The Core List contained the words present in 3 or more of the 6 basal reader series. The Additional List contained the words not in the Core List but present in 4 or more series, including content series. Each Core and Additional word was placed at the lowest grade level at which or below which it appeared in 3 series. Four Technical Lists were set up, one for each content area; these contained words not in the Core List but present in both content series and judged to have a technical meaning in that curriculum area. All words not qualifying for any of these lists were deleted. Also deleted were proper nouns, most abbreviations, and non-word items such as alphabet letters and digits. These deletions reduced the original 80,000-plus different items to 5,167 Core words, 1,699 Additional words, and 805 Technical words, giving a Total Alphabetical List of 7,613 words (plus their 9,236 inflected forms). For the computerized comparison each inflected form was listed separately, giving a total of 16,849 entries. For brevity, this list will be referred to below as the H-J List.
The Dale List of 3,000 Familiar Words (Dale and Chall, 1948) has been one of the most widely used vocabulary lists because it provides the most important component in the Dale-Chall Readability Formula. "The technique used for constructing the list was crude. When 80 per cent of the fourth graders questioned indicated that they knew a word, that word was included in the list." (Dale and Chall, 1948, p. 18) "Knew" seems to have meant both recognition and understanding of a meaning of the word. The list actually contains 2,946 words, in one alphabetical order. Inflected forms were merged with root words following rules quite similar to those used by Harris and Jacobson. For brevity, this list will be called Dale 3,000.

The Bucks County 1185 Common Words list was compiled in 1955 (Botel, 1962), and provides the basis for Botel's method of determining reading levels of children's books. It is based on a study of 5 major basal reader programs at each level from pre-primer through high third. A word was assigned to the lowest level at which it was used in 3 of the 5 series. As published, the original 1,185 words are amplified by 557 inflected forms that are identified by asterisks, giving a total of 1,742 words. There is some inconsistency in the treatment of derived forms. For example, the entry word arrive includes arrives and arrived, but arriving is a separate entry. On the other hand, ask also represents asks, asked, and asking. For the present study all of the inflected forms of the 1,185 words were listed separately, giving an expanded total of 3,266 words. For brevity, this will be identified as the Botel List.

The Revised Core Vocabulary (Taylor, Frackenpohl, and White, 1969) is the fifth edition of a list first developed in 1949. It will be referred to as the Taylor List. It is based in part on a 1968 study of the vocabulary of 9 series of basal readers, and in part on the Rinsland List of children's writing vocabulary (Rinsland, 1945) and the Thorndike-Lorge List (Thorndike and Lorge, 1944). First a tentative decision was made about the desirable number of items to have at each reading level, and then criteria were chosen which gave approximately those totals. For grades one through 3, a word was placed at the lowest level at which it appeared in 3 series. For grades 4 through 6, "Words occurring in at least three basals or below were judged to be suitable for that level. Words occurring twice at the same level or below were checked against the Rinsland List to confirm student
knowledge of the word, at that level, and next against the Thorndike-Lorge List to determine whether a word was used frequently enough in written material (G listing) to justify its inclusion.” (Taylor, Frackenpohl, and White, 1969, p. viii). For grades 7 and 8, “The balance of the words suggested by the basals in grades four to six and not yet included in the vocabulary list (on the basis of too low a frequency) were rechecked against the Rinsland and Thorndike-Lorge lists and added to the list if their frequencies then warranted. The remainder of the word load for these two levels was derived almost exclusively from the Rinsland and Thorndike-Lorge lists, using the same general approach as previously described.” A separate list of 2,411 words for grades 9 through 13 was chosen mainly from the Thorndike-Lorge List.

Comparison of the 4 lists

The quantitative results of the computerized comparisons of the H-J List with the other 3 lists are shown in Table 1. Of the 2,946 words in the Dale 3,000, 2,744 (or 93 per cent) are also H-J words. Of the 3,266 words in the Botel List, 3,095 (or 95 per cent) are also in the H-J List. Thus the Dale words known to most fourth graders and the Botel primary-grade words are nearly all included in the H-J List, which spans grades one through 6. The overlapping with the 2 Taylor lists is lower. Of the 6,714 Taylor words for grades one through 8, 5,473 (or 81 per cent) are also in the H-J List. The Taylor List for grades 9 through 13 shows only 7 per cent overlapping with the H-J List.

Table 1  Comparisons of the Harris-Jacobson List with the Dale 3,000, Botel, and Taylor lists

<table>
<thead>
<tr>
<th></th>
<th>Dale 3,000 List</th>
<th>Botel List</th>
<th>Taylor Lists 1-8</th>
<th>Taylor Lists 9-13</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Words</td>
<td>2,946</td>
<td>3,266</td>
<td>6,714</td>
<td>2,426</td>
</tr>
<tr>
<td>Number also in</td>
<td>2,744</td>
<td>3,095</td>
<td>5,473</td>
<td>179</td>
</tr>
<tr>
<td>H-J List</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number not in</td>
<td>202</td>
<td>171</td>
<td>1,241</td>
<td>2,247</td>
</tr>
<tr>
<td>H-J List</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Percent in</td>
<td>93%</td>
<td>95%</td>
<td>82%</td>
<td>7%</td>
</tr>
<tr>
<td>H-J List</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

While these raw tallies are interesting, the printed comparisons of the lists provide a means of discovering the more specific
differences in the lists. The effects of aging, for example, are evident in the comparison of the Dale 3,000 with the H-J List. The Dale words not in the H-J List include the following which seem obsolete or of diminished frequency now: afar, apiece, bedbug, bookkeeper, bran, buttermilk, candlestick, christen, codfish, cooper, fib, fret, goody, henhouse, jig, lard, lass, lice, Overall, reap, schoolmaster, sleigh, snuff, trolley, washtub. Conversely, the H-J List contains words which have come into common use since the Dale 3,000 was developed, such as TV, elevator, tractor, traffic, battery, camera, detective, experiment, helicopter, strike, astronaut, bargain, committee, concrete, hamburger, satellite. Vocabulary evolves as new scientific terms come into general use, as current events bring words to the forefront in news reports and conversation, and as public attitudes change.

The computer printout allows a quick evaluation of factors accounting for differences between lists. Comparison of the H-J and Botel lists revealed that most of the differences were the result of different criteria for inclusion. The Botel List includes only 4 root words that are not in the H-J List: berry, excite, fairground, and linesman. It also includes Indian and Christmas, excluded from the H-J List as proper nouns. It has some plurals which occurred too infrequently to be in the H-J List, such as bedrooms, buses, neckties, and schoolrooms (in the H-J List plurals were merged with the root word only if they were found in the materials). Of the words in both lists, approximately 1,700 were placed at the same level and approximately 1,400 at different levels.

Differences in compilation criteria also seemed to account for much of the difference between the H-J List and the Taylor List. Of the approximately 1,200 Taylor words (grades one through 8) not in the H-J List, 14 are placed at or below fourth grade, 65 are fifth grade, 132 are sixth grade, and almost 1,000 are at seventh- and eighth-grade levels. These last were words that did not meet the Taylor criteria for grade 6, and also did not meet the H-J criteria for grade 6; thus the difference between the lists is not as great as it would at first seem.

Comparisons of samples

Random samples of 100 words each were taken of the Dale 3,000, Botel and Taylor 1-8 lists; and the H-J level placement was determined for each word. The results are shown in Table 2.
Table 2  Harris-Jacobson reading level placements for 100-word samples of the Dale 3,000, Botel, and Taylor lists

<table>
<thead>
<tr>
<th>H-J Level</th>
<th>Dale 3,000</th>
<th>Botel 3rd Grade</th>
<th>Taylor 3rd Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre-primer</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Primer</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>First Reader</td>
<td>10</td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>Second Reader</td>
<td>20</td>
<td>30</td>
<td>52</td>
</tr>
<tr>
<td>Third Reader</td>
<td>20</td>
<td>52</td>
<td>59</td>
</tr>
<tr>
<td>Fourth Reader</td>
<td>21</td>
<td>13</td>
<td>24</td>
</tr>
<tr>
<td>Fifth Reader</td>
<td>8</td>
<td></td>
<td>5</td>
</tr>
<tr>
<td>Sixth Reader</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Technical</td>
<td></td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

No Entry

Eligible Words not in H-J List

<table>
<thead>
<tr>
<th>Proper Nouns</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Elite Words</td>
<td>7</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Proper Nouns</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
</tbody>
</table>

Total

|               | 100 | 100 | 100 |

The Dale words are words that were known to 80 per cent or more of fourth graders in the 1940's. As can be seen in Table 2, 91 of the 100 words were found in the H-J List, a close approximation to the 93 per cent for the entire Dale List. The 91 words in common range in H-J placement from pre-primer to sixth grade, with 77 per cent at or below fourth grade. These 77 Dale words are fairly evenly divided among first-, second-, third-, and fourth-grade levels. There were 8 fifth reader words, 6 sixth reader words, and 9 words not in the H-J List; the latter included 2 proper nouns and armful, baa, boo, christen, coffeepot, fret, and jacks.

The Botel sample of 100 words was taken from the words identified by Botel as belonging at the lower and upper third reader levels. As shown in Table 2, 97 of these words are also in the H-J List, a close approximation to the 95 per cent for the entire Botel List. Slightly more than half are placed at third reader level in both lists, and 83 per cent are placed at or below third reader level in the H-J List. The 3 words not in common are milkmen, overhear, and sticker.

The Taylor 100-word sample was drawn from the words listed by Taylor as third reader level. Only 2 of these words are not in the H-J List, and both of them are proper nouns. Ten words have H-J ratings of second reader, 59 are rated third reader, 24 are rated fourth reader, and 5 are rated fifth reader. Thus there is agreement between the 2 lists on the placement of the majority of the words. Of the words on which there is divergent placement, the Taylor List tends
Comparing word lists

HARRIS AND JACOBSON

95

to place them higher than H-J, while the Botel List tends to place them lower.

Comparisons with other computerized word lists

The Kucera-Francis List

Kucera and Francis (1967) made a computerized analysis of the words in the Brown University Corpus. That corpus is a one-million-word sample of adult reading material, containing 500 samples of about 2,000 words each, drawn from 15 different genres or types of material, using a combination of stratified and random sampling. All of the materials sampled were published in 1961. An "individual word" was defined as "a continuous string of letters, numerals, punctuation marks, and other symbols (e.g., of graphemes), uninterrupted by space . . ." (Kucera and Francis, 1967, p. xxi). Thus not only were inflected forms of a word treated separately, but the list also includes thousands of formulas, arabic numbers, combinations of initials, and abbreviations. The list is presented twice: in a descending order of frequency and alphabetically.

The Kucera-Francis List contains 50,406 items, of which over 23,000 appeared only once and nearly 7,000 appeared only twice. Because of the length of the list and the inclusion of many thousands of items that had been excluded from the H-J List as ineligible, it did not seem feasible to employ the kind of computerized comparison employed with the Dale 3,000, Botel, and Taylor lists. Instead, 2 samples of 100 words each were drawn, using the descending frequency listing. Since the H-J List contains 1,924 words for grades one through 3, the first sample included the words ranked from 1,875 to 1,974 in the Kucera-Francis List. The words in this sample have frequencies of 56 to 58 per million words, occurring in 5 to 15 genres, and in 7 to 55 samples. The second sample included the words ranked from 7,514 to 7,613, corresponding to the final 100 words in the H-J List. The words in the second sample all had a frequency of 11 occurrences per million words, and had occurred in from one to 8 of the 15 genres, and in one to 11 samples.

The first sample, corresponding to end of primary, contained 88 words that are in the H-J List. Of the 12 words not in common, 7 were proper nouns and one was an abbreviation, leaving only 4 words that were eligible for the H-J List but not in it. These words
(sexual, ultimate, completion, and academic) seem clearly more appropriate for an adult vocabulary than for an elementary school vocabulary. The 88 words shared by the 2 lists ranged in H-J placement from primer level to sixth reader level, as shown in Table 3, with a median at third reader level. Fifty-five words were rated primary, and 33 words were rated at levels 4 through 6. Thus there was agreement between the lists that the words were among the 2,000 most common words for a little over half of the sample.

Table 3 Harris-Jacobson reading level placements for 100-word samples of the Kucera-Francis and American Heritage lists

<table>
<thead>
<tr>
<th>H-J Level</th>
<th>Kucera-Francis</th>
<th>American Heritage</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1875-1974</td>
<td>7514-7613</td>
</tr>
<tr>
<td>Pre-Primer</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>Primer</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>First Reader</td>
<td>19</td>
<td>22</td>
</tr>
<tr>
<td>Second Reader</td>
<td>23</td>
<td>16</td>
</tr>
<tr>
<td>Third Reader</td>
<td>23</td>
<td>23</td>
</tr>
<tr>
<td>Fourth Reader</td>
<td>17</td>
<td>9</td>
</tr>
<tr>
<td>Fifth Reader</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Sixth Reader</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Technical</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Eligible Words not in H-J List</td>
<td>4</td>
<td>42</td>
</tr>
<tr>
<td>Proper Nouns</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>Letters or Digits</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Abbreviations</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>

The second sample, corresponding to end of sixth grade, contained only 52 words that were also in the H-J List. Of the other 48 words, 6 were proper nouns and 42 were words like comprise, dimensional, expectation, and forthcoming. For the 52 words in common, the median H-J placement was fifth reader level, with 13 words of second and third reader levels, 39 words rated fourth to sixth reader levels, and 2 technical words (enzyme and finite). Thus at a level corresponding to the end of the H-J List there was only a moderate degree of overlapping.

The American Heritage List

The American Heritage Word Frequency Book (Carroll, Davies, and Richman, 1971) provides a word list similar in construc-
tion to the Kucera-Francis List except that the samples were based on materials for grades 3 through 9, and totaled over five million words. The list was based on 1,045 samples of about 500 running words each, taken from textbooks, workbooks, kits, novels, poetry, general non-fiction, encyclopedias, and magazines. The definition of a word or "type" corresponds to that used by Kucera and Francis; it "does not discriminate against proper nouns, numerals, formulas, nonce words, nonsense words, and other real forms not usually considered lexical in a formal sense . . ." (Carroll, Davies, and Richman, 1971, p. xiii). The words are presented in 2 arrangements, alphabetically and in descending order of frequency adjusted for dispersion. For brevity, this list will be referred to below as the AHI (American Heritage Intermediate) List.

Comparison of the AHI List with the H-J List is made difficult by a number of factors. Although both are elementary school word lists, the AHI List covers grades 3 through 9, while the H-J List covers grades one through 6. The AHI List is based on a large number of small samples, while the H-J List is based on the entire vocabularies of a limited number of textbook series. The AHI List found 86,741 different entries and kept them all, including 35,000 that occurred only once; the H-J List found over 80,000 entries and boiled them down to 7,613 basic words (16,849 when inflected forms are counted in). The AHI List treats each inflected form of a word as a separate entry, while in the H-J List the inflected forms actually found were merged with the root word. The AHI List arranges the words alphabetically and in order of frequency, while the H-J List contains separate Core and Additional Lists arranged by levels, Technical Lists, and a Total Alphabetical List.

Despite these differences, comparisons can be made using the same procedure as for the Kucera-Francis List. The results are shown in Table 3.

A 100-word sample consisting of the AHI words rated from 1,875 to 1,974 in the descending frequency listing was found to contain 88 words that are also in the H-J List. Of the 12 items not in common, 8 were proper nouns and 4 were letters or digits, so that none of the 12 was eligible for the H-J List. In other words, every word that met the H-J List rule requirements was actually in the H-J List. Their H-J placements ranged from pre-primer to sixth reader level, with a median at third reader level. Fifty of the words were at
or below third reader level, and 38 words were placed at fourth to sixth reader levels. The words rated sixth reader level were somewhat, element, cell, appearance, angles, and progress.

In the 100-word sample occupying AHI ranks from 7,514 through 7,613, corresponding to the total number of H-J words, 82 were in the H-J List and 18 were not. The 18 words not in the H-J List included 13 capitalized words, 2 numbers, and only 3 eligible words according to H-J rules: acute, literary and management. Thus there were 85 words eligible by H-J rules, and 82 of them (or 96 per cent) were actually in the H-J List. The H-J placements for the 82 words in common ranged from primer to sixth reader level and included 2 Technical words: quotient and establishment. The median H-J placement was fourth reader level, with 28 words rated as primary and 54 rated above primary.

One may conclude that there is far greater agreement between the AHI List and the H-J List than between the Kucera-Francis List and the H-J List. AHI and H-J agree almost completely on which words belong among the 7,600 most frequent elementary school words, the differences being due mainly to the AHI inclusion of items barred from the H-J List. There is somewhat less agreement on placement of the words, which is due at least in part to the differing criteria used for sequencing the words. Even so, the majority of the words in the first sample, corresponding to end of primary, were rated primary in H-J, and the majority of words in the second sample, corresponding to the H-J total, were of intermediate level. Despite the many differences in materials covered, criteria for inclusion, and procedures for sequencing the words, there is substantial agreement between the AHI and H-J lists.

It may be noted that in the primary-level samples, the Kucera-Francis analysis and the AHI analysis give quite similar results. Both have 88 words in common with the H-J List, and the distributions of H-J placements are very similar, as can be seen by comparing the first and third columns in Table 3. At the 7,514-7,613 frequency level, the 2 samples are quite different, as can be seen by comparing the second and fourth columns. The AHI sample has 82 words in the H-J List, while the Kucera-Francis List has only 52 words in that list. Of the words in H-J, the AHI List has a median at fourth reader level, while the Kucera-Francis median is at fifth reader level. Most of the AHI words not in H-J were excluded by the application of
Comparing word lists

HARRIS AND JACOBSON

general rules, while 42 of the Kucera-Francis words were excluded from H-J because of insufficient frequency.

From these comparisons it seems reasonable to conclude that the 2,000 most frequently used words in American English printed matter are pretty much the same, whether the materials analyzed are for grades one through 6, for grades 3 through 9, or for adults. At around the 7,500 mark in descending frequency, or upper sixth grade, the degree of commonality between elementary school vocabularies and adult vocabulary is much lower. At frequencies between these 2, one would probably find results shifting toward less agreement as the frequency rating moves toward less common words.

Comparisons of short reading word lists

This section considers 6 short reading word lists, ranging in length from 220 words to 769 words, analyzes their contents in terms of H-J ratings, and discusses their suitability. Data about these lists may be found in Table 4.

Dolch Basic Sight Vocabulary versus the Johnson List

The Dolch Basic Sight Vocabulary (Dolch, 1936) has probably been the most widely used of all short word lists, especially in corrective and remedial reading programs. It was compiled by taking the words common to three lists published before 1931 and, according to Dolch, consisted of 50 per cent or more of the running words in most elementary reading materials.

The Dolch BSV has recently been criticized by Johnson (1971a), who asserts that it is substantially out of date. His criterion is degree of correspondence to the first 220 words of the Kucera-Francis List, an adult list described above. Johnson found that of the 220 Dolch words, 82 are not among the most frequent 220 Kucera-Francis words. He also proposed a list of 306 words to take the place of the Dolch BSV (Johnson, 1971b). These were words that met a double criterion: being among the 500 most frequent Kucera-Francis words and also used 50 or more times by kindergarten or first-grade children (Murphy, 1957).

It seemed appropriate to analyze both the Dolch BSV and the Johnson List in terms of reading level placement in the H-J List. The results are shown in the first and second columns of Table 4. Because
the number of words varied from list to list, the results are stated in percentages.

Table 4 Harris-Jacobson reading level placements for 6 short reading word-lists

<table>
<thead>
<tr>
<th>H-J Level</th>
<th>Dolch BSV</th>
<th>Johnson</th>
<th>Otto-Chester</th>
<th>Thorndike-Lorge 1st 500</th>
<th>Fry 1st 300</th>
<th>2nd 300</th>
<th>Stone-Dale</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preprimer</td>
<td>24.0%</td>
<td>17.6%</td>
<td>13.8%</td>
<td>10.2%</td>
<td>18.0%</td>
<td>0.3%</td>
<td>8.0%</td>
</tr>
<tr>
<td>Primer</td>
<td>20.5%</td>
<td>14.1%</td>
<td>11.4%</td>
<td>9.6%</td>
<td>17.0%</td>
<td>2.0%</td>
<td>5.0%</td>
</tr>
<tr>
<td>First reader</td>
<td>36.4%</td>
<td>35.3%</td>
<td>34.4%</td>
<td>26.8%</td>
<td>40.0%</td>
<td>14.3%</td>
<td>31.0%</td>
</tr>
<tr>
<td>Total 1st Grade</td>
<td>80.9%</td>
<td>67.0%</td>
<td>59.6%</td>
<td>46.6%</td>
<td>75.0%</td>
<td>16.6%</td>
<td>44.0%</td>
</tr>
<tr>
<td>Second reader</td>
<td>18.2%</td>
<td>26.8%</td>
<td>31.0%</td>
<td>32.4%</td>
<td>22.7%</td>
<td>48.0%</td>
<td>46.0%</td>
</tr>
<tr>
<td>Third reader</td>
<td>0.9%</td>
<td>4.3%</td>
<td>3.8%</td>
<td>13.8%</td>
<td>2.0%</td>
<td>24.7%</td>
<td>7.0%</td>
</tr>
<tr>
<td>Fourth reader</td>
<td>0.3%</td>
<td>1.4%</td>
<td>5.6%</td>
<td></td>
<td>8.3%</td>
<td></td>
<td>1.0%</td>
</tr>
<tr>
<td>Fifth reader</td>
<td>0.9%</td>
<td>1.7%</td>
<td>24.7%</td>
<td></td>
<td>0.3%</td>
<td></td>
<td>7.0%</td>
</tr>
<tr>
<td>No entry</td>
<td>1.6%</td>
<td>4.2%</td>
<td>1.6%</td>
<td></td>
<td>0.7%</td>
<td></td>
<td>1.0%</td>
</tr>
<tr>
<td>Total</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
<td>100.0%</td>
</tr>
</tbody>
</table>

It can be seen that 80.9 per cent of the Dolch words are placed at first grade levels in the H-J List, as compared to only 67.0 per cent of the Johnson words. The difference is mainly at pre-primer and primer levels. Most of the remaining words in both lists are of second reader level, but the Johnson List contains 1.6 per cent (5 words) that are not in the H-J List.

The 82 Dolch words that Johnson criticized as obsolete were also examined. They included 63.4 per cent that have first grade ratings in the H-J List, with 35.4 per cent at second reader level, 15.9 per cent at third reader level, 7.3 per cent at fourth and fifth reader levels, and 3.7 per cent (3 items) not in the H-J List. The 82 corresponding words from Kucera-Francis include 32.6 per cent at first grade levels, 40.2 per cent at second reader level, 15.9 per cent at third reader level, 7.3 per cent at fourth and fifth reader levels, and 3.7 per cent (3 items) not in the H-J List. Thus the 82 Dolch words criticized by Johnson seem more appropriate than the 82 words used in his criterion. Furthermore, the 82 Kucera-Francis words include 25 words that are always or usually nouns and 6 words that are sometimes nouns. Since Dolch excluded all nouns from his list, it would seem that Johnson should have excluded nouns from his criterion.

Returning to the Johnson List, the 306 words include 14 words that are inflected forms, mainly noun plurals and past tenses of verbs. For 8 of these, the root is also within the 306 words; for the
other 6, the root word is not present. There are no inflected forms in the Dolch BSV.

If the main criterion for judging suitability of a brief word list is that knowing the words should help one to read easy elementary school material, the old Dolch BSV seems superior to the newer Johnson List on several counts.

The Otto-Chester List

A "Great Atlantic and Pacific Sight Word List" has recently been presented (Otto and Chester, 1972). This list contains, very simply, the 500 words that were the most frequent in the AHI third grade samples. The distribution of H-J placement of the Otto-Chester words is shown in the third column of Table 4. Almost 60 per cent of the words are at first-grade levels, mainly first reader. Of the remainder, 31 per cent are second reader, 3.8 per cent are third reader, 1.4 per cent are fourth reader, and 4.2 per cent (21 entries) are not in the H-J List.

The Otto-Chester List has recently been criticized by Harris and Jacobson (1973). "Any symbol or group of symbols that was preceded and followed by white space was identified as a 'word' by the computer and retained. Thus the 500 words selected by Otto and Chester include six arabic numerals (1, 2, 3, 4, 5 and 6; but not 7, 8, or 9). They also include twelve proper nouns, three words that are present twice (once beginning with a lower-case letter and once beginning with a capital), 35 words that are represented two or three times each (once by the root word, and also by one or two inflected forms—e.g., live, lived and living are separate entries), and 26 inflected forms whose root words are not in the list. Thus 82 words out of the 500, or about 16 per cent of the list, are of dubious value in a basic sight vocabulary list." (Harris and Jacobson, 1973)

The Thorndike-Lorge First 500 Words

The original Teacher's Word List of 10,000 Words (Thorndike, 1921) and its revisions, extending it to 20,000 words (Thorndike, 1931) and to 30,000 words (Thorndike and Lorge, 1944) represent the most extensive sampling of print to date, totaling about 18 million words. The original Thorndike list was based about equally on child and adult materials. To extend the list to 20,000 words, most of the new material was adult; and to extend it to 30,000 words, all of
the new material was adult material. Nevertheless, the most frequent 500 words in the Thorndike-Lorge List have often been used as a criterion of suitability for first grade reading material.

The H-J placement distribution for this list is shown in the fourth column of Table 4. The per cent of first grade words, 46.6, is lower than for the Otto-Chester List. There are 5.6 per cent of fourth reader level, and 1.6 per cent not in the H-J List. As pointed out by the list's authors, the first thousand of the original Thorndike List (1921) are "better to use in choosing very common words for children" than the first thousand of the Thorndike-Lorge List (Thorndike and Lorge, 1944, p. 270). The present writers are in full agreement with that statement.

The Fry Instant Words

Fry (1957) has prepared a list of "Instant Words" particularly for remedial teaching of sight vocabulary. His selection was based on the Thorndike-Lorge first 500, the Rinsland List based on children's writing (Rinsland, 1945), and the Faucett-Maki List (1932), which in turn was based on the original Thorndike List (1921) and the Horn (1926) adult writing vocabulary list.

Fry arranged his list in groups of 25 words for teaching, the first 25 being the most frequent and the last 25 the least frequent. He pointed out that the first 300 words in his list are much more frequent than the second 300 and that there is high agreement among all of the word counts for the first 300 words but considerably less agreement about the second 300.

Because of Fry's comments, separate distributions of placements in the H-J List were made for Fry's first 300 and second 300 words. These are shown in Table 4. For the first 300, 75.0 per cent are of first grade level, and only 2.3 per cent are above second reader level. For the second 300, only 16.7 per cent are at first grade level, 48.0 per cent are at second reader level, 24.7 per cent are at third reader level, 10.0 per cent are at fourth and fifth reader levels, and 0.7 per cent are not in the H-J List. Thus the first 300 Fry words seem well selected for their main purpose of filling gaps in basic reading vocabulary, while the utility of the second 300 seems much more doubtful.
The Stone Revision of the Dale Easy Words

Dale (1931) prepared a list of the 769 words that occurred both in the first thousand of the Thorndike List (1921) and in the International Kindergarten Union List (1926) of words known to five year olds. This became known as the Dale List of 769 Easy Words. It came into wide use when it was adopted by Spache as the main element in the Spache Readability Formula for primary grade materials (Spache, 1953). Stone offered a modernized version of the Dale 769 in which 173 of the Dale words were replaced by 173 easy words from Stone's own word counts (Stone, 1956). Spache accepted the Stone revision and has used it in recent expositions of his formula (Spache, 1972, pp. 194-207). For this reason, the Stone Revision of the Dale 769 Easy Words was added to the lists to be analyzed.

The H-J placement distribution for a 100-word random sample of the Stone-Dale List is shown in the last column of Table 4. There is an almost even distribution between first and second grade levels, with 44 per cent at first grade levels and 46 per cent at second reader level. Of the remaining 10 words, 7 are third reader level, one each at fourth and fifth, and one is an ineligible word. In view of the fact that the H-J List contains 912 root words of the first and second grade levels, the presence of 90 per cent of the Stone-Dale words within these levels shows a high degree of correspondence between the two lists.

Discussion

Although in the comparisons presented above the H-J List was used as the criterion for analyzing other word lists, the relationships discovered cast light on the H-J List also. The writers, as authors of that list, were interested in discovering what they could about the validity and usefulness of that list. Cronbach (1971, p. 444), in a recent discussion of the validity of psychological and educational tests, distinguished among criterion-related validity or predictive validity, content validity, and construct validity.

In giving evidence about criterion-related validity, testmakers usually present data on how the new test correlates with other tests of the same ability or trait. They hope for correlations that will be substantial but not too high; near-perfect correlations would indicate that
the new test is not significantly better than the predecessors, none of which is considered to be a truly satisfactory criterion. In comparing word lists, correlations are replaced by 2 kinds of numerical evidence: per cent of overlapping between lists, and degree of agreement on the placement of words in levels or in a frequency order.

The data presented above show that there are substantial to high degrees of overlapping between the H-J List and all of the lists with which it has been compared. Highest overlappings were with the Dale 3,000 List and Botel List, both over 90 per cent, for the longer comparison lists. Overlappings were in the 80 to 90 per cent range for the Taylor 1-8 List, both AHI samples, and the Kucera-Francis first sample. The lowest degree of overlapping was with the Kucera-Francis second sample, which seems to show that there are substantial differences between elementary school and adult vocabularies at about the upper sixth grade level, corresponding to about the 7,500 level in descending frequency.

In terms of placement, there was fairly high agreement with the Dale 3,000 List, the Botel third grade sample, the Taylor third grade sample, the Dolch Basic Sight Vocabulary, the Fry First 300 Instant Words, and the Stone Revision of the Dale 769 Easy Words. There was lower agreement on placement with the Johnson and Otto-Chester Lists and with the Kucera-Francis and AHI Lists from which they were derived; and also with the Thorndike-Lorge First 500 and the Fry Second 300. Degree of agreement seems to be highest when the comparison list is based on primary materials and tends to decrease in rough relationship to the proportion of secondary school and adult vocabulary in the comparison list.

Criterion or predictive validity of tests also involves correlating the test (usually given before instruction) with achievement or performance after instruction. Thus a readiness test given prior to reading instruction is usually correlated with achievement scores after one or more years of reading instruction. There does not seem to be any analogous procedure that can be applied to vocabulary lists.

The H-J List was intended to be a list of the most important or basic words in required reading materials for the first 6 grades of elementary school. In terms of content validity, the possible unreliability of small samples was avoided by utilizing the entire vocabularies of the 127 books in the 14 textbook series. The rules for merging inflected forms with root words and for eliminating certain entire
classes of entries were chosen after careful consideration of the rules used by Thorndike and Lorge (1944) and by Dale and Chall (1948). The criteria used for inclusion of a word (present in 3 of 6 reading series for the Core List, present in 4 of 14 series for the Additional List) provided lists of reasonable length and were quite similar to those used by Botel (present in 3 of 5 series) and by Taylor (present in 3 of 9 series). The criterion for placing a word at a particular reading level (the lowest level at which the word was present in at least 3 series) corresponded roughly to the mid-score of the word's first appearances. These facts should be helpful in judging the content validity of the H-J List.

Construct validity is the most difficult to evaluate, and the procedures described by Cronbach do not seem to apply to word lists. Perhaps the most appropriate way to attempt to evaluate the construct validity of a word list is to try to answer the question: To what extent are the design and content of the list congruent with the major purposes or uses for which the list is intended?

The main purpose for constructing the *H-J Basic Elementary Reading Vocabularies* was described as "to show which words—at each grade level—were widely used in elementary school textbooks in 1970." (Harris and Jacobson, 1972, p. 1) The information provided in this paper may aid the reader in answering the question as it applies to the H-J List.

The 2 other recently published computerized word lists differ in many respects from the H-J List. They are non-selective, including every item that could be distinguished by the computer as a "word." They used a large number of relatively small samples, rather than the total vocabularies of a limited number of complete books. While they collected and reported data on the number of different sources in which each word was found, they gave most weight to the total number of times the word occurred. Both the Kucera-Francis List and the AHI List contain many thousands of words that occurred just once.

These characteristics seem appropriate for lists which are intended primarily to provide a factual base for quantitative linguistic research. The Kucera-Francis book (1967) contain 149 pages of research data on such topics as the distribution of occurrence of the most frequent words among genres, type-token ratios, sentence-length variety and homogeny, lognormal word frequency distributions, etc. The AHI List seems to have been compiled with 2 main purposes: to
provide a research basis for a new dictionary, and to provide a basis for linguistic research; 110 pages of statistical analyses and interpretations are presented based on the list (Carroll, Davies, and Richman, 1971).

The inclusion of all symbols or groups of symbols set off by spaces, characteristic of both the Kucera-Francis and AHI lists, allowed certain types of research to be undertaken which would not have been possible if, for example, inflected forms had been merged with their root words; research such as comparisons of the relative frequencies of the singular and plural forms of key words, and the relative frequency of such pairs as Mother and mother, king and King. This decision, which gave priority to meeting the needs of quantitative linguistic studies, impaired the value of these 2 lists for checking the vocabulary difficulty of reading materials. Those lists which have merged most inflected forms with their root words (Thorndike-Lorge, Dale 3,000, Botel, Stone-Dale, Fry, and Harris-Jacobson) seem to be better adapted to that purpose.

No word list can be the best list for all of the various purposes for which word lists have been used. Comprehensive lists are needed for some purposes, brief lists for other purposes. Some lists are most appropriate for primary-grade applications, others for the intermediate grades, secondary years, or adult levels. For some purposes, a frequency sequence is most desirable; for other purposes, arrangement of words in levels is most helpful.

The Harris-Jacobson Basic Elementary Reading Vocabularies have characteristics that make them seem well suited for some applications and not appropriate for others: 1) they are based exclusively on textbook materials for the first 6 grades; 2) they provide a list of the most basic elementary words; 3) they are the result of the most recent large-scale word count, based on materials widely used in 1970; 4) each of the 6,866 Core and Additional words is assigned to a specific reading level; 5) the words are presented both in a total alphabetical sequence and by reading levels; 6) the placement of a word depends on its distribution, not its frequency; and 7) data provided for each word allow a user to employ other criteria than those utilized by the authors.

The relative stability of the vocabulary used in elementary reading materials is worthy of note. Despite some differences in criteria, a high degree of overlapping was found in all of the lists based
Comparing word lists  HARRIS AND JACOBSON 107

on children's materials. Even the Kucera-Francis adult list overlapped with the H-J List 88 per cent at the third grade level.

There is a high degree of agreement on which are the 2,000 commonest words, regardless of differences in materials analyzed and criteria employed, and regardless of whether assignment is based on total frequency of occurrence or on the number of sources in which the word occurs. There is even high agreement between assignment based on per cent of children knowing the meaning, and assignment based on commonness of occurrence. While there is some evidence of aging in the older lists, the 25-year-old Dale 3,000 List has 93 per cent overlapping with the H-J List and should not be considered obsolete.

The agreement on words for grades 4 through 6 is also high, although not as high as at primary levels. The 82 per cent overlapping between the Taylor 1-8 List and the H-J List would have been much higher if the Taylor words for seventh and eight grades had been deleted. The AHI second sample and H-J List overlapped 82 per cent despite the fact that the comparison was made at the equivalent of upper sixth grade; at fourth and fifth grade levels the agreement would have been higher than that. Thus one can conclude that the degree of agreement among lists drawn from children's materials is high for grades one through 6, corresponding to the most common 7,500 words.

The one list which did not have high overlapping at sixth grade level was the Kucera-Francis second sample, which had ranks just beyond 7,500. Over 40 per cent of that sample are not in the H-J List, and inspection of those words shows that most of them are highly abstract and far more suitable for an editorial than for a children's book. Thus while there is high agreement between adult and child word lists on the most common 2,000 words, the degree of agreement drops as one moves toward less frequently occurring words. The 7,500 level corresponds to a frequency of 11 occurrences in one million, or about once in 91,000 running words of adult material. These adult words occur much less frequently than that in materials for intermediate grade children.

There is far less agreement among lists on the placement of words in levels or in rank order than on overlapping. For example, while there is high overlapping between the AHI List and the H-J List, the AHI words in the first sample, corresponding to upper third grade,
were found in the H-J List at all levels from pre-primer to sixth, and the words in the AHI second sample occurred at all H-J levels from primer to sixth. This may be due to the different sampling procedures employed in the lists. While the same words tend to be found, certain words common in first grade materials (daddy, pony, etc.) are less common in upper-grade materials, and some words widely used in upper-grade materials do not occur at all in beginning reading materials. When one list excludes first and second grade materials and the other excludes materials for grades 7 through 9, differences in ranking may be expected.
REFERENCES


FRY, EDWARD. Developing a word list for remedial reading. Elementary English, November 1957, 36, 456-458.


HORN, ERNEST. A basic writing vocabulary. University of Iowa Monographs in Education, First Series, No. 4, 1926.


JOHNSON, DALE G. The Dolch list reexamined. The Reading Teacher, February 1971, 24, 449-457.(a)

JOHNSON, DALE G. A basic vocabulary for beginning reading. Elementary School Journal, September 1971, 72, 29-34.(b)


LORGE, IRVING. Predicting readability Teachers College Record, March 1944, 45, 404-419.


