The purpose of this paper was to investigate letters-per-syllable constants as predictors of true syllable counts for basal readers and trade books at five instructional levels. One hundred fifty language samples of approximately one hundred words each were selected and keypunched for computer analysis to determine average word length and total number of running words. Manual syllable counts for each sample were also performed. Their values were then used to calculate letters-per-syllable predictors for each level. The use of letters-per-syllable constants predicted between 73 percent and 90 percent of the variance of the actual syllable count. It was concluded that this procedure for calculating the number of syllables makes it possible to program a computer to compute readability of language samples using readability measures which rely on syllable counts. (TS)
THE VALIDATION OF A SYNTHETIC SYLLABLE COUNT
APPROPRIATE FOR COMPUTER-DETERMINED READABILITY ESTIMATES

Carol J. Hopkins and Alden J. Moe
Purdue University


The word readability has been used to refer to characteristics of reading materials which help determine how easily a passage can be understood, how fluently it can be read, and how much interest it generates for the reader. Factors typically considered in determining readability are used in various ways in readability formulas which results in some grade level designation being given to the passage examined. The factors examined in readability calculations include a measure of syntactic complexity and a measure of vocabulary difficulty. Syntactic complexity is usually determined by examining sentence length and vocabulary difficulty is determined by noting the presence of hard word as defined by some list (the Dale lists, for example) or by a syllable count.
The Lorge (1944), Dale and Chall (1948) and Spache (1953) readability formulas use two word lists compiled by Dale to determine the percentage of "hard words" in a sample and the recently developed Harris-Jacobson Primary Readability Formulas (1973) are also based on the percentage of words included on graded word lists.

The Flesch Reading Ease Formula (1948), the Gunning 'Fog Index' (1952), the Fry Readability Formula (1968), and McGlaughlin's 'Smog Grading' formula require finding the number of syllables contained in the passage for which the readability is being calculated.

More than 50 readability formulas have been developed to investigate the level or difficulty of various written language samples (Harris, 1974) and reviews by Chall (1958), Klare (1963), Seels and Dale (1971), Gilliland (1972), and Harris (1974) provided information on the values, limitations, and uses of the various formulas.

Unfortunately, the application of these formulas is often tedious and time-consuming. However, the availability of computer technology has now made it possible to analyze large amounts of language. Through computer analysis of text it is easy to obtain the following information:

1. the total number of running words in the text
2. the total number of different words
3. an alphabetical listing of all words used
4. a sorted listing of all words used based on frequency of use
5. a percentage of use figure for each word
6. average word length and standard deviation
7. number of words of one letter, two letters, etc.
8. total number of sentences
9. average sentence length and standard deviation
10. number of sentences of one word, two words, etc.

It is difficult however, to develop a computer program which will accurately identify the number of syllables in words because of problems encountered in defining syllables objectively.

Fang (1968) reported the difficulty in developing a computer program using existing rules of syllabication. Coke and Rothkopf (1970) developed three computer algorithms for estimating the average number of syllables per word in text from letter counts of vowels per word, consonants per word, and letters per word. Readability scores based on actual syllable counts and the computer-estimated syllable counts yielded a correlation of .92 for vowels, .78 for consonants and .88 for letters. Their study, however, was based on only 28 one-hundred-word samples.

Felsenthal, Shamo, and Bittner (1971) concluded that the number of letters per syllable could be accurately estimated by dividing the number of letters in the text by the constant 3.1127. This constant enabled them to use a computer to determine a syllable count. In their analysis of 40 language samples, correlations were computed for an actual syllable count with both a letter count and vowel count. The use of the constant 3.1127 characters per syllable to determine syllables with the computer, correlated .98 with the actual syllable count.

Use of the 3.1127 constant is also reported in studies to assess the readability of language samples (Felsenthal and Felsenthal, 1972), to measure the readability of selected social studies texts (Felsenthal, 1973), and the readability of Newbery award books (Moe and Arnold, 1973).
Other than the original calculations conducted to determine the 3.1127 constant, however, no further validation has been done.

Purpose

The purpose of this study was to investigate letters-per-syllable constants as predictors of true syllable counts for basal readers and trade books at five instructional levels.

The identification of a new constant or the validation of the 3.1127 constant will permit quick computer analysis of language samples and subsequent calculations of readability estimates for all formulas which require a syllable count.

Methods

The concept of readability is based on the premise that there are differences in the difficulty levels of language. Therefore, for the purpose of this study, basal readers and trade books were selected to reflect increasing language complexity.

Five instructional levels were identified; they were:

1. primary (grades one through three)
2. intermediate (grades four through six)
3. junior high (grades seven through nine)
4. senior high (grades ten through twelve)
5. college (designated as grades thirteen through fifteen for the purposes of this study).

Five basal readers and five trade books were selected for each grade at every instructional level, resulting in ten first-grade books, ten second-grade books, etc. Therefore, a total of 150 books were used in this study.
The basal readers selected for use for all grades at the primary and intermediate levels were the most recent editions of the Ginn; Harper & Row; Holt, Rinehart and Winston; Houghton Mifflin; and Lyons and Carnahan basal reading series.

Because there are generally no basal readers as such at the junior high and high school levels, the investigators selected the most recent editions of literature textbooks for those levels published by Ginn; Houghton Mifflin; Holt, Rinehart and Winston; McGraw-Hill; and Scott Foresman.

Textbooks commonly used in introductory and advanced college English courses comprised the basal reader component for the college level.

Trade books used in this study were also chosen to reflect increasing language complexity. The books chosen were recommended in the following sources:

1. Adventuring with Books (Root, 1973)
2. Books and the Teen-Age Reader (Carlsen, 1971)
4. Children and Books (Arbuthnot and Sutherland, 1972)
5. Children's Catalog (Fidell, 1971 and supplements)
6. Children's Literature in the Elementary School (Huck and Kuhn, 1968)
7. Elementary School Library Selection (Gaver, 1972)
9. Junior High School Library Catalog (1965 and supplements)
10. Literature for Adolescents: Selection and Use (Meade and Small, 1973)
A complete listing of the books selected appears in the bibliography of trade books.

A one hundred word sample from each of the 150 books selected was keypunched for computer analysis of the text. The samples were taken from a page near the middle of the book. Each sample began with the first word in a sentence and ended with the last word in the sentence in which the 100th word occurred.

A computer program developed at Purdue University performed all calculations described earlier in this paper (p. 2). Pertinent to the analyses to be performed in this study were average word length, and the total number of running words in the text. The total number of syllables in each sample was determined by an actual syllable count. These results were recorded for synthetic letter-per-syllable calculations to be performed later in the study.

New letters-per-syllable constants were calculated to determine the best predictor of true syllable counts for basal readers and trade books. Using the formula:

\[
\text{New Constant} = \frac{\text{sum of the average word length}}{\text{sum of the actual syllables}} \times \frac{\text{sum of the total words}}{\text{number of books at level}}
\]

A new constant was calculated for:

1. primary level basal readers and trade books (N = 30)
2. intermediate level basal and trade books (N = 30)
3. junior high level basal and trade books (N = 30)
4. senior high level basal and trade books (N = 30)
5. college level (N=30)
6. All primary through college level basal readers (N = 75)
7. All primary through college trade books (N = 75)
8. All primary through college basal readers and trade books (N = 150)
In addition to the constants listed in Table I, the constant 3.1127 was included for purposes of validation. The investigators also chose to include 3.0000 as a constant, reasoning that this was an easier value to manipulate in calculations.

Table I about here

To determine which of these new letters-per-syllable constants was the best predictor of actual syllable counts, a stepwise regression was performed for each level.

Results

Results obtained from stepwise multiple regressions performed for all levels listed above indicated the best predictors for each grade level. The amount of variance accounted for by each of the predictor variables using the various constants accounted for from 73% ($R^2 = .72843$) to 90% ($R^2 = .89715$) of the variance.

The amount of variance accounted for by the best predictors for each level was compared to the amount of variance accounted for by the 3.1127 Felsenthal, Shamo and Bittner constant and the 3.0000 constant included by the investigators. Results of the regression analyses appear in Table II.

Table II about here

These comparisons indicate that the difference in amount of variance accounted for the best predictor of each level, by 3.1127, and by 3.0000 predictor of the number of letters-per-syllable. The following formula
may be used to determine synthetic syllable count.

\[
\text{Number of Syllables} = \frac{(\text{average word length}) \times (\text{total number of running words})}{\text{(constant used)}}
\]

Conclusions

Because the amount of variance accounted for by each of the predictor variables was substantial, one can accept the notion that a constant used to predict the actual number of syllables through the use of computer programming has been demonstrated.

It has been determined that the constant 3.0000 is as accurate a constant for use in calculating the number of letters per syllable as the 3.1127 constant proposed by previous researchers or any of the grade level predictors determined for this study. Regardless of which of these constants one chooses to use, it is possible to apply the constant in a formula to calculate the number of syllables per language sample which may be used for subsequent calculations of readability requiring a syllable count.

Computer programs can be developed which will use this constant to generate readability levels which can be printed directly on the printout. Readability can be determined merely by keypunching a language sample and then analyzing the sample by computer programming which will provide a readability estimate on the printout.
REFERENCES


Harris, Albert J. "Some New Developments on Readability." Paper presented at the Fifth IRA World Congress on Reading, Vienna, Austria, August 13, 1974.


McLaughlin, G. H. "Proposals for British Readability Measures." Third International Reading Symposium (Brown, A.L. and Downing, J., eds.) Cassell.


BIBLIOGRAPHY OF BASAL READERS

Grade 1


Young America Basic Reading Program, Level 6. Pasadena: Lyons & Carnahan, 1972, pp. 76-78.

Grade 2


Young America Basic Reading Program, Level 7. Pasadena: Lyons & Carnahan, 1972, pp. 136-137.

Grade 3


Grade 4


Young America Basic Reading Program, Level II. Pasadena: Lyons & Carnahan, 1972, pp. 222-223.

Grade 5


Grade 6


Young America Basic Reading Program, Level 13. Pasadena: Lyons & Carnahan, 1972, p. 244.

Grade 7


Grade 8

Grade 9

Grade 10


Grade 11


Grade 12


**Grade 13**


**Grade 14**


Grade 15


BIBLIOGRAPHY OF TRADE BOOKS

Grade 1


Grade 2


Grade 3


Grade 4


Grade 5


Grade 6


Grade 7


Grade 8


Grade 9


Grade 10


Grade 11


**Grade 12**


**Grade 13**


Grade 14


Grade 15


<table>
<thead>
<tr>
<th>Level</th>
<th>Constant</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>3.1968</td>
</tr>
<tr>
<td>Intermediate</td>
<td>3.2113</td>
</tr>
<tr>
<td>Junior High</td>
<td>3.1486</td>
</tr>
<tr>
<td>Senior High</td>
<td>3.1302</td>
</tr>
<tr>
<td>College</td>
<td>3.0770</td>
</tr>
<tr>
<td>All Trade</td>
<td>3.1664</td>
</tr>
<tr>
<td>All Basals &amp; Trade</td>
<td>3.1491</td>
</tr>
</tbody>
</table>
# TABLE II

SUMMARY OF VARIANCE ACCOUNTED FOR BY CONSTANTS
USED AS PREDICTOR VARIABLES
IN REGRESSION EQUATIONS

<table>
<thead>
<tr>
<th>Level</th>
<th>Best Predictor</th>
<th>3.1127</th>
<th>3.0000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Primary</td>
<td>.72843&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.72753&lt;sup&gt;a&lt;/sup&gt;</td>
<td>.71033&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Intermediate</td>
<td>.75587</td>
<td>.75540</td>
<td>.75547</td>
</tr>
<tr>
<td>Junior High</td>
<td>.78425</td>
<td>.78425</td>
<td>.77934</td>
</tr>
<tr>
<td>Senior High</td>
<td>.89715</td>
<td>.89629</td>
<td>.89711</td>
</tr>
<tr>
<td>Adult</td>
<td>.85245</td>
<td>.85193</td>
<td>.85172</td>
</tr>
<tr>
<td>All Basals</td>
<td>.85275</td>
<td>.84832</td>
<td>.85091</td>
</tr>
<tr>
<td>All Trade</td>
<td>.84765</td>
<td>.84765</td>
<td>.84719</td>
</tr>
<tr>
<td>All Basal &amp; Trade</td>
<td>.84351</td>
<td>.84184</td>
<td>.84213</td>
</tr>
</tbody>
</table>

<sup>a</sup>All numerals in the columns represent $R^2$ values.