# COMPUTER-BASED READABILITY INDEXES

Douglas R. McCallum
Computation Center and
Department of Computer Sciences

James L. Peterson
Department of Computer Sciences

The University of Texas
Austin, Texas 78712

## 1.0 INTRODUCTION

Computer-based document preparation systems provide many aids to the production of quality documents. A text editor allows arbitrary text to be entered and modified. A text formatter then imposes defined rules on the form of the text. A spelling checker ensures that each word is a correctly spelled word. None of these aids, however, affect the meaning of the document; the document may be well-formatted and correctly spelled but still incomprehensible.

The general problem of checking correct syntax and semantics of English is still very much a research problem. However, one form of assistance in producing a readable document can be very easily provided now: a readability index. A readability index is a measure of the ease (or difficulty) of reading and understanding a piece of text. These generally give the grade level (1 to 12) of the material or an index, from 0 (hard) to 100 (easy). Several readability formulas to compute readability indexes have been defined and are in fairly wide use. We give some of these formulas in Table I.

Readability formulas were originally developed mainly by educators and reading specialists. One of the first, the Flesch formula, was published in 1948 and is still in wide use. The primary application was in defining the appropriate reading level for elementary and secondary school text books.

More generally, readability indexes can be used to assist a writer by pointing out possible grammatical and stylistic problems and by helping to maintain a consistent audience level throughout a document. Particularly now that more and more newspapers, books and reports are created with the aid of computer-based document systems, the use of readability indexes should become quite common.



Readability indexes were originally computed by hand. As early as 1963, however, attempts were being made to computerize their computation. Several formulas have been developed with machine computation in mind, while improved computer algorithms have helped with others. As we show in Section 4, the computation of these formulas is relatively easy. Readability indexes can be computed by stand-alone programs, or they could be added to existing text processing programs, such as an editor, formatter, or spelling checker.

Most recently, the value of readability formulas as one part of a system for helping writers has been recognized. The PWB UNIX Writers Workbench [Cherry 1981] provides many types of document analysis, including calculation of 4 readability indexes (Kincaid, Automated Readability Index, Coleman-Liau, and Coke-Rothkopf). The STAR system developed by General Motors calculates the Flesch formula. The Computer Readability Editing System developed for the U.S. Navy computes the Kincaid formula, which it indicates is a Department of Defense standard [Kincaid 1981]. Barry [1980] describes a system written in Fortran to compute the Flesch, Dale-Chall, Farr-Jenkins-Paterson, and Fog formulas. Programs written in BASIC to compute readability indexes for personal computer systems have been published [Irving and Arnold 1979].

## 2.0 UNDERLYING PRINCIPLES

Defining and selecting a readability formula requires some attention to the underlying question: What constitutes a readable document? Specifically, what features of text play an important role in determining readability? A survey of the features used in various readability formulas reveals the following list of features which have been suggested and used:

1. length of words (in characters)
2. number of words of 6 or more letters
3. number of syllables
4. number of words which are monosyllables
5. number of words of 3 or more syllables
6. number of affixes (prefixes or suffixes)
7. number of words per sentence
8. number of sentences
9. number of pronouns
10. number of prepositions

Klare [1968, 1974] and others [McLaughlin 1966] have shown that the two most common variables in a readability formula are: (1) a measure of word difficulty, and (2) a measure of sentence difficulty. Clearly a sentence with a number of unusual and uncommon words will be more difficult to understand than a sentence with simpler, more common words. Similarly a sentence with contorted and complex syntactic structure is more difficult to read than a sentence with a simple structure. The problem is mainly how to measure, and combine these two variables.

## 2.1 Measures Of Word Difficulty

The most direct measure of the difficulty of a word is its frequency in normal use. Words which are frequently used are easy to read and understand; words which are uncommon are more difficult to read and hence understand.

Analysis of English language text has shown that a small number of words occur very frequently while many words are quite uncommon. In the Brown Corpus [Kucera and Francis 1967] for example, a body of 1,014,232 words yielded only 50,406 unique words. Of these the 168 most frequently occurring words accounted for half the occurrences of the total number of words, while about half the unique words were used only once. This skewness in frequency of use means that the best measure of the commonness of a word is probably the logarithm of its frequency of use.

However, since the computation of this measure requires a dictionary of words plus their commonness, this measure is generally not used directly. Rather, indirect measures, such as features (1) to (6) above, are used in an attempt to approximate word difficulty.

Most of these measures are based on the fact that common words tend to be short, while uncommon words tend to be longer: Zipf's law [Zipf 1935]. Thus, measures such as (1), the number of characters per word, or (3), the number of syllables per word, are indirect measures of the frequency of a word in the language.

## 2.2 Measures Of Sentence Difficulty

The difficulty of a sentence would seem to depend mainly upon its syntactic structure. However, again, this is not readily computable or quantifiable. Thus, more easily computed measures, such as (7), sentence length, are more commonly used, based upon the statistically valid assumption that long sentences are more complex than short sentences.

## 2.3 Combining The Features

Once the features to be measured are selected, they must be combined to produce a composite readability value. The existing formulas are generally derived in the following manner:

1. An independent assessment of the reading difficulty of a collection of texts is made, either by a panel of judges, a standard set of material (such as the McCall-Crabbs [1925] Standard Test Lessons in Reading) or cloze tests. A cloze test replaces every fifth word from a passage with a blank space and then determines the difficulty of the passage by the percent of deleted words which can be correctly guessed by a reader. (More difficult passages mean that fewer words can be guessed.)

2. The values of the chosen features (number of words, number of sentences, and so on) are computed for each of the texts in the collection.

3. (Linear) regression analysis is applied to produce the coefficients of a linear equation combining the features and computing the independently derived readability score.

Thus the coefficients in readability formulas are empirically determined.

## 3.0 EXAMPLE READABILITY FORMULAS

Many readability formulas have been developed. A survey paper by Klare in 1963 listed over 30 different formulas for determining readability; an update in 1974 listed over 30 more new or updated formulas. Many of these vary only slightly from others and all are highly correlated.

One of the earliest (1948) and most popular formulas is the Flesch formula [Flesch 1948]. Designed for general adult reading matter, it is based upon the McCall-Crabbs Lessons. The formula yields a readability index in the range 0 (hard) to 100 (easy).

$$R = 206.835 - 84.6 * S/W - 1.015 * W/T$$

where,

S = total number of syllables
W = total number of words
T = total number of sentences

This formula is based upon the average number of syllables per word (S/W), a measure of word difficulty, and the average number of words per sentence (W/T), a measure of sentence difficulty.

At about the same time as the publication of the Flesch formula, the Dale-Chall formula [Dale and Chall, 1948] was published. It is one of the more accurate general purpose formulas. There are two major differences between the Dale-Chall formula and the Flesch formula.

1. The Dale-Chall formula does not compute a number from 0 (hard) to 100 (easy), but computes the grade level (1 through 12) of a pupil who can answer correctly at least half of the test questions asked about a text passage.

2. The measure of word difficulty is computed directly from the Dale Long List [Dale and Chall, 1948]. The Dale Long List is a list of 3000 general words known to 80 percent of fourth grade children (in 1948, of course). Words on the Dale Long List are considered easy; words not on the Dale Long List are considered hard.

The original formula, published in 1948, was based upon the 1925 McCall-Crabbs Lessons.

$$G = 19.4265 - 15.79 * D/W + .0496 * W/T$$

where,

D = total number of words on the Dale Long List
W = total number of words
T = total number of sentences

In 1950, the McCall-Crabbs Lessons were revised, and Powers, Sumner, and Kearl [1958] recalculated the Dale-Chall formula for the new McCall-Crabbs Lessons. This produced the following formula.

$$G = 14.8172 - 11.55 * D/W + .0596 * W/T$$

Again in 1961, the McCall-Crabbs Lessons were revised, and again the Dale-Chall formula was recalculated [Holquist 1968]. This time the formula was,

$$G = 14.862 - 11.42 * D/W + .0512 * W/T$$

Table I lists 11 readability formulas. These formulas were selected from a survey of the literature on readability formulas. Table I includes most of the more commonly used readability formulas. Of particular note are the Automated Readability Index and Coleman-Liau formula. These two readability formulas were designed specifically to be easy to compute.

| | |
|---|---|
| $R = 206.835 - 84.6 * S/W - 1.015 * W/T$ | Flesch |
| $R = -31.517 + 159.9 * M/W - 1.015 * W/T$ | Farr-Jenkins-Paterson |
| $R = 235.87 - 84.44 * V/W - 1.015 * W/T$ | Coke-Rothkopf |
| $R = -37.95 + 116.0 * M/W + 148.0 * T/W$ | Coleman |
| $G = 14.862 - 11.42 * D/W + .0512 * W/T$ | Dale-Chall |
| $G = 3.068 + 9.84 * P/W + .0877 * W/T$ | Fog |
| $G = -21.43 + 4.71 * L/W + 0.50 * W/T$ | Automated Readability Index |
| $G = -15.8 + 5.88 * L/W - 29.59 * W/T$ | Coleman-Liau |
| $G = -15.59 + 11.8 * S/W + 0.39 * W/T$ | Kincaid |

W = total number of words
T = total number of sentences
L = total number of letters
V = total number of vowels
D = total number of words on the Dale Long List
S = total number of syllables
M = total number of one-syllable words
P = total number of words with 3 syllables or more

Table I -- Common Readability Formulas and Their Variables

## 4.0 IMPLEMENTATION

Constructing a program to compute a readability index is fairly straightforward. First one (or more) of the readability formulas are selected for computation. The text of the input file is read, accumulating the necessary counts. Finally the readability formula is used to compute and print an index for the specific input file.

The statistics needed for computing most formulas are easily accumulated in one pass through the document. The number of letters, vowels, words (sequences of letters separated by blanks or punctuation), and sentences (sequences of words separated by period, exclamation point or question mark) are easy to compute. The number of words on the Dale Long List can be computed directly by reading in a copy of the list and searching it for each word, or it can be approximated such as suggested by Irving and Arnold [1979].

The most difficult statistics are probably those dealing with the number of syllables per word. These can be exactly determined by a dictionary look-up, or they can be approximated by the approach of Fang [1968] or Coke and Rothkopf [1970].

One other point in implementation concerns the amount of text over which the index is computed. While we certainly want an index for the entire input file, we probably also want readability measures for smaller pieces of the file. Thus, we may want to compute a separate index for each section, each page, or each paragraph. This allows an author to quickly scan a document looking for sections which are more (or less) difficult than appropriate for the intended audience. These portions may be rewritten to bring them more into line with the author's intentions.

## 5.0 CONCLUSIONS

A very simple program can be written to compute a readability index for a document. Readability formulas have been developed by reading specialists to allow easy determination of the reading level of a document. With the new ability of computers to store large dictionaries of words, and their properties on-line, we expect that even better readability indexes can be produced and can help to improve the quality of documents produced with the aid of a computer system.

As an example, applying the formulas listed above to this paper results in the following readability indexes:

| Readability: 0 (hard) to 100 (easy) |
|---|
| 81.7 Coke-Rothkopf |
| 55.2 Farr-Jenkins-Paterson |
| 53.0 Flesch |
| 47.3 Coleman |

| Grade Level: 1 (easy) to 12 (hard) |
|---|
| 6.5 Fog |
| 8.0 Coleman-Liau |
| 9.1 Automated Readability Index |
| 10.6 Dale-Chall |
| 11.0 Kincaid |
| 13.6 SMOG |

Acknowledgements: We would like to thank Carol Engelhardt for her assistance in this work.

## 6.0 REFERENCES

1. J. G. Barry, "Computerized Readability Levels", IEEE Transactions on Professional Communications, Volume PC-23, Number 2, (June 1980), pages 88-90.

2. Lorinda Cherry, "A Toolbox for Writers and Editors", Proceedings of the AFIPS Office Automation Conference, (March 1981), pages 221-227.

3. Lorinda Cherry, "Computer Aids for Writers", Proceedings of the ACM SIGPLAN/SIGOA Symposium on Text Manipulation, (June 1981), pages 61-67.

4. Ester U. Coke and Ernst Z. Rothkopf, "Note on a Simple Algorithm for a Computer-Produced Reading Ease Score", Journal of Applied Psychology, Volume 54, Number 3, (1970), pages 208-210.

5. Edmund B. Coleman, "On Understanding Prose: Some Determiners of its Complexity", NSF Final Report GB-2604, (1965).

6. Meri Coleman and T. L. Liau, "A Computer Readability Formula Designed for Machine Scoring", Journal of Applied Psychology, Volume 60, Number 2, (1975), pages 283-284.

7. Edgar Dale and Jeanne S. Chall, "A Formula for Predicting Readability", Educational Research Bulletin, Volume 27, (February 1948), pages 11-20, 37-54.

8. Irving E. Fang, "By Computer: Flesch's Reading Ease Score and a Syllable Counter", Behavioral Science, Volume 13, (1968), pages 249-251.

9. James N. Farr, James J. Jenkins, and Donald G. Paterson, "Simplification of Flesch Reading Ease Formula", Journal of Applied Psychology, Volume 35, Number 5, (October 1951), pages 333-337.

10. Rudolf F. Flesch, "A New Readability Yardstick", Journal of Applied Psychology, Volume 32, (June 1948), pages 221-233.

11. Robert Gunning, The Technique of Clear Writing, McGraw-Hill, (1952).

12. John B. Holquist, "A Determination of Whether the Dale-Chall Readability Formula may be Revised to Evaluate More Validly the Readability of High School Science Materials", Ph.D. Thesis, Colorado State University, (1968).

13. Steve Irving and Bill Arnold, "Measuring Readability of Text", Personal Computing, (September 1979), pages 34-36.

14. J. Peter Kincaid, Robert P. Fishburn, Jr., Richard L. Rogers, and Brad S. Chissom, "Derivation of New Readability Formulas for Navy Enlisted Personnel", Research Branch Report 8-75, Naval Air Station Memphis, Millington, Tennessee, (February 1975), 40 pages.

15. J. Peter Kincaid, James A. Aagard, John W. O'Hara, and Larry K. Cottrell, "Computer Readability Editing System", IEEE Transactions on Professional Communications, Volume PC-24, Number 1, (March 1981), pages 38-41.

16. George R. Klare, The Measurement of Readability, Iowa State University Press, Ames, Iowa, (1963).

17. George R. Klare, "The Role of Word Frequency in Readability", Elementary English, Volume 45, (January 1968), pages 12-22.

18. George R. Klare, Paul P. Rowe, M. Gregory St. John, and Lawrence M. Stolurow, "Automation of the Flesch Reading Ease Readability Formula, With Various Options", Reading Research Quarterly, Volume 4, (Summer 1969), pages 550-559.

19. George R. Klare, "Assessing Readability", Reading Research Quarterly, Number 1, (1974-1975), pages 62-102.

20. H. Kucera and W. N. Francis, Computational Analysis of Present-Day American English, Brown University Press, (1967), 424 pages.

21. W. A. McCall and L. M. Crabbs, Standard Test Lessons in Reading, Columbia University Teachers College, 1926, 1950, 1961.

22. G. Harry McLaughlin, "What Makes Prose Understandable", Ph.D. Thesis, University College, London, (1966).

23. G. Harry McLaughlin, "SMOG Grading -- a New Readability Formula", Journal of Reading, Volume 12, (May 1969), pages 639-646.

24. R. D. Powers, W. A. Sumner, and B. E. Kearl, "A Recalculation of Four Readability Formulas", Journal of Educational Psychology, Volume 49, (April 1958), pages 99-105.

25. Edgar A. Smith and J. Peter Kincaid, "Derivation and Validation of the Automated Readability Index for Use with Technical Materials", Human Factors, Volume 12, Number 5, (October 1970), pages 457-464.

26. George K. Zipf, The Psycho-Biology of Language, Houghton Mifflin Co., Boston, (1935); reprinted by M.I.T. Press, Cambridge, (1965).