

An Analysis to Identify the Most Suitable Readability Formula for Sri Lankan Context: Case Study Computer Science Students

Choolangika Sirisuriya¹, Lochandaka Ranathunga², Shironika Karunanayaka³ and Aniza Abdullah⁴

¹*Department of Computer Science, Faculty of Computing, General Sir John Kotelawala Defence University, Sri Lanka*

²*Department of Information Technology, Faculty of Information Technology, University of Moratuwa, Sri Lanka*

³*Department of Secondary and Tertiary Education, Faculty of Education, The Open University of Sri Lanka, Nugegoda, Sri Lanka*

⁴*Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, Malaysia*

Email: ¹*choolangika@yahoo.com*

²*lochandaka@uom.lk*

³*spkar@ou.ac.lk*

³*noraniza@um.edu.my*

Abstract: E-Learning is learner centric therefore content readability is an important factor when designing E-Learning content. Readability is the quality of writing that makes it easy or difficult to read and understand text. Readability tests or readability formulas are formulae for evaluating the readability of text, usually by counting syllables, words, and sentences. Researchers and writers have been using readability formulas since 1920 and, over the years, they have spent a lot of time devising the most accurate and scientific formulas to assess readability. At the moment, there are more than 200 readability formulas developed by various scholars but only some of these formulas are reliable to determine the reading-level of a text. The aim of our study is to investigate what are the most suitable readability formulas for Sri Lankan context since Sri Lankans are nonnative English speakers who have basic knowledge of English. This study focused on a survey with the participation of undergraduate students in Sri Lanka who read and evaluated the readability of three paragraphs. We selected the undergraduates who are studying in the Computer Science stream therefore the paragraphs are related to Computer Science field with different complexity levels. Then we calculate the readability of the three given paragraphs using selected readability formulas and compare the results with the students' feedbacks to select the most suitable formulas for Sri Lankan context. Factors that we consider when selecting the readability formulas are those suitable for college level readers, using English as Second/Foreign Language, popularity of the formula and suitability of the formulas design for technical text. After selecting suitable formulas design, a tool is developed to measure the text readability.

Keywords: Readability, Readability Formulas, Sri Lankan context

Introduction

Readability is the measure of how easy to read and comprehend a document. Readability tests or readability formulas are formulae for evaluating the readability of text, usually by counting syllables, words, and sentences. Readability tests were first developed in the 1920s in the United States. They are mathematical formulas, designed to determine the suitability of books for American students at a certain age or grade level. The tests were intended to help educators, librarians and publishers make decisions about purchase and sale of books. They were also meant to save time, because before the formulas were used those decisions were made based on recommendations of educators and librarians who read the books. They studied the existing books to determine the appropriate reading groups.

Today readability formulas are usually based on one semantic factor (the difficulty of words) and one syntactic factor (the difficulty of sentences). Being mathematically based, readability tests are unable to determine the likelihood that the document is comprehensible, interesting, or enjoyable. Layout and design are also important factors to the readability of a document that cannot be determined using readability tests. Many of these factors cannot be measured using readability tests. Instead, readability tests provide a prediction of the reading ease for a document but not the only method for determining readability and they do not help us evaluate how well the reader will understand the ideas in the text.

Presently, there are more than 200 readability formulas developed by various scholars but only a handful of these formulas are reliable to determine the reading-level of a sample text ("How To Choose The Best Readability Formula For Your Document," n.d.). When we studied the pedagogical evaluation process for web based E-Learning content, we discovered that pedagogical reviewers are more concern about the readability of the text since E-Learning is a learner centered learning method (de S Sirisuriya et al., 2013a)(de S Sirisuriya et al., 2013b). The discovery leads to the need to select the most suitable readability formula for Sri Lankan content. The aim of our study is to investigate what are the most suitable readability formulas for Sri Lankan context since Sri Lankans are nonnative English speakers who have basic knowledge of English. We conducted a survey with the participation of undergraduate students in Sri Lanka who read and evaluated the readability of three paragraphs. Then we calculate the readability of the three given paragraphs using selected readability formulas and compare the student feedbacks to select the most suitable formulas for Sri Lankan context. Factors that we consider when selecting the readability formulas are those that we believe suitable for college level readers, using English as Second/Foreign Language, popularity of the formula and the suitability of the formulas design for technical documents. This paper examines the readability formulas like Flesch Kincaid Grade Level, Flesch Kincaid Reading Ease, Gunning Fog Index, SMOG, Coleman Liau Index and Automated Readability Index (ARI). We selected the undergraduates who are studying in the Computer Science stream therefore the selected paragraphs are Computer Science related with different levels of complexity. After selecting suitable formulas design, a tool is developed to measure the text readability.

Review of Literature

1. Flesch–Kincaid Readability Test

The Flesch Reading Ease and the Flesch–Kincaid Grade Level are two readability tests which are designed to specify the comprehension difficulty when reading a passage of modern academic English ("Flesch–Kincaid readability tests," 2015).

Word length and sentence length are the core measurements of these two formulas but they have different weighting factors. The results of these two tests are inversely related to each other that means a text with a comparatively high score on the Reading Ease test should have a lower score on the Grade Level test ("Flesch–Kincaid readability tests," 2015).

- Flesch Reading Ease

Flesch Reading Ease Formula is one of the oldest and most accurate readability formula developed by Rudolph Flesch in 1948 (Si and Callan, 2001). Assessing the easiness of the text to read is the core concept in Flesch Reading Ease formula. Because of that it becomes one of the few accurate measures and we can rely on without too much scrutiny. The Flesch Reading formula, the score ranges from 0 to 100 (Table I). Higher scores indicate the lowest reading difficulty and lower numbers corresponding to the highest reading difficulty. Generally, College or University graduates can understand documents with a score of 0-30. Equation (1) is the formula for the Flesch Reading Ease Score test (Zamanian and Heydari, 2012).

$$\text{Flesch Reading Ease} = 206.835 - 1.015 \left(\frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left(\frac{\text{total syllables}}{\text{total words}} \right) \quad (1)$$

Table 1 Score value of the Flesch Reading Ease

Score	Notes
90.0 – 100.0	Easily Understood by an average 11-year-old student
60.0 – 70.0	Easily Understood by 13-to-15-year-old students
0.0 – 30.0	Best understood by university graduates

The Flesch Reading Ease Formula is simple but it has certain following ambiguities. Periods, explanation points, colons and semicolons are considered as sentence delimiters, each group of continuous non-blank characters with in punctuation are not counted as a word, each vowel in a word is considered one syllable subject to: (a) -es, -ed and -e (except -le) endings are ignored; (b) words of three letters or shorter count as single syllables; and (c) consecutive vowels count as one syllable.

- Flesch–Kincaid Grade Level

Flesch-Kincaid Grade Level Readability formula develops by improving the Flesch Reading Ease Readability Formula. Rudolph Flesch is the co-author of this formula along with John P. Kincaid because of that this formula is called as Flesch-Kincaid Grade Level Readability formula. In 1976, this formula was originally formulated for US Navy purposes but this Formula is best suitable in the field of education. The Flesch–Kincaid Grade Level Formula translates the Reading Ease Score to a U.S. Grade School Level, making it easier for teachers, parents, librarians, and others to judge the readability level of various books and texts (Zamanian and Heydari, 2012). The formula for the Flesch Kincaid Grade Level test is given in Equation (2).

$$\text{Flesch Kincaid Grade Level} = 0.39 \left(\frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left(\frac{\text{total syllables}}{\text{total words}} \right) - 15.59 \quad (2)$$

The result is a number that corresponds with a grade level.

2. Gunning Fog Index

The Gunning fog index is another formula that measures the readability of English writing. The index estimates the years of formal education needed to understand the text on a first reading. A fog index of 12 requires the reading level of a U.S. high school senior (around 18 years old). This formula was developed by Robert Gunning, in 1952 (Zamanian and Heydari, 2012). The Fog Index, which became popular because of its ease of use. It uses two variables, average sentence length and the number of words with more than two syllables for each 100 words.

The Gunning fog index is calculated with the following algorithm:

- Select a passage (such as one or more full paragraphs) of around 100 words.
- Determine the average sentence length. (Divide the number of words by the number of sentences.);
- Count the "complex" words: those with three or more syllables.
- Add the average sentence length and the percentage of complex words; and
- Multiply the result by 0.4.

The complete formula is given in Equation (3).

$$\text{Fog Index} = (\text{Average no of words per sentence} + \text{Percentage of words with three or more syllables}) * 0.4 \quad (3)$$

3. SMOG (Simple Measure Of Gobbledygook)

The SMOG (Simple Measure Of Gobbledygook) developed by G. Harry McLaughlin in 1969, is another measurement of readability that estimates the years of education needed to understand a piece of writing. The SMOG grade produces a 0.985 correlation with a standard error of 1.5159 grades with the grades of readers who had 100% comprehension of test materials.

This formula was developed as a more accurate and more easily calculated substitute for the Gunning fog index.

In this formula, the word length and sentence length should be multiplied rather than added. By counting the number of words of more than two syllables (polysyllable count) in 30 sentences, McLaughlin provided this simple formula:

$$SMOG\ Grading = 3 + \sqrt{Polysyllable\ count} \quad (4)$$

The formula for the SMOG grading is given in Equation (4).

4. The Coleman–Liau Readability Formula

The readability formula designed by Meri Coleman and T. L. Liau is called the Coleman–Liau Readability Formula or the Coleman–Liau Index which is used to measure the understandability of a text. When compare to the other commonly used readability formulas, the Coleman–Liau Index estimates the U.S. grade level which can be understand the given text (McCallum and Peterson, 1982)

$$CLI = 0.0588L - 0.2965S - 15.8 \quad (5)$$

Where

L is the average number of letters per 100 words

S is the average number of sentences per 100 words

Meri Coleman and T. L. Liau thought that characters more easily and accurately understand by computerized assessments than counting syllables and sentence length (Pitler and Nenkova, 2008). Because of that the Coleman–Liau formula depends on characters instead of syllables per word. This concept is similar to the Automated Readability Index (Refer to section 5).

5. Automated Readability Index

The Automated Readability Index (ARI) is a readability test designed to measure the understandability of a text (Smith and Kincaid, 1970).

$$ARI = 4.71 \left(\frac{\text{characters}}{\text{words}} \right) + 0.5 \left(\frac{\text{words}}{\text{sentences}} \right) - 21.43 \quad (6)$$

The formula for the Automated Readability Index test is given in (6).

Where,

characters are the number of letters, numbers and punctuation marks

words are the number of spaces

sentences are the number of sentences

The ARI formula outputs a number which estimates the grade level required to understand the text.

Most of the readability formulas like the Flesch-Kincaid Grade Level, Gunning Fog Index, SMOG Index and Coleman-Liau Index, produce an approximate value of US grade level which is needed to understand the text. The ARI and the Coleman-Liau formulas trust on characters per word than syllables per word.

Methodology

Readability means the quality of written language that makes it easy to read and understand. Being able to measure the readability of a text with a simple formula is an attractive prospect, and many groups have been using readability formulas in a variety of situations where the estimation of text complexity are thought to be necessary (Bruce et al., 1981). After studying several readability formulas, we conducted a survey with the participation of eighty-five undergraduate students who studying in Computer Science stream in a leading Sri Lankan university who read and evaluated the readability of three paragraphs. We conduct the survey in a lecture hall with all the participants because in a noisy environment, participants may have difficulty to concentrate to read or comprehend the passage. Before starting the survey, we gave them a brief introduction about what to do before they start the survey. Paragraphs are also based on Computer Science stream with different complexity levels. Before going to select three paragraphs, we calculate the complexity levels of paragraphs by using the selected readability formulas. Then we select paragraphs with different complexity levels. Then compare the results with feedbacks from the student to select the most suitable formulas for Sri Lankan context. Factors that we consider when selecting the readability formulas are as follows: the suitability of the formulas for college level readers, using English as Second/Foreign Language readers, the popularity of the formula and the suitability of the formulas design for technical text or documents. In this survey, the undergraduates were required to rank the readability of the three paragraphs, using a 5-point Likert scale. Additionally, we asked participants a simple question from each paragraph to check their understanding of the document.

After selecting suitable formulas, we developed a tool to measure the text readability presented in figure1. We used php to develop this small tool. We have to paste or type our text that we want to measure its readability. When ‘Calculate Readability’ button pressed, a pop window will appear to display readability value of the given text according to the Flesch Kincaid Grade Level, Flesch Kincaid Reading Ease, Gunning Fog Index, SMOG, Coleman Liau Index and Automated Readability Index (ARI).

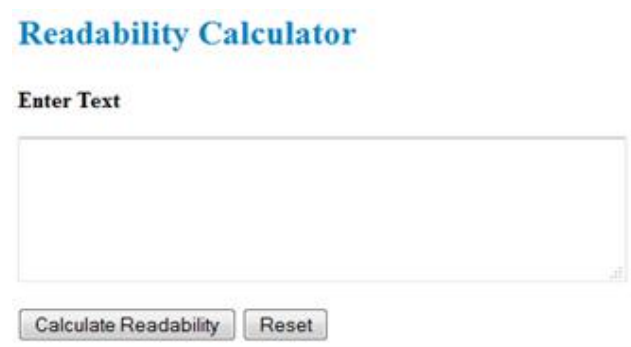
The image shows a web-based interface for a 'Readability Calculator'. At the top, the title 'Readability Calculator' is displayed in a blue, stylized font. Below the title, there is a label 'Enter Text' in a bold, black font. Underneath this label is a large, empty text input area with a light gray border. At the bottom of the interface, there are two buttons: 'Calculate Readability' and 'Reset', both with a light gray background and a thin border.

Figure 1: Tool to measure Readability

Data Analysis

The survey was conducted on eighty-five undergraduates. Out of the eighty-five feedbacks, paragraph1 and paragraph3 were evaluated by eighty-three undergraduates and paragraph2 was evaluated by eighty-two undergraduates. Percentage values of the readability of the three paragraphs obtained from the feedbacks are presented in Table 2.

Table 2 Percentage value of readability three paragraphs

Paragraph	Very Easy	Easy	Moderate	Difficult	Very Difficult
Paragraph 1	39.76%	34.94%	24.10%	1.20%	-----
Paragraph 2	30.49%	40.24%	26.83%	2.44%	-----
Paragraph 3	28.92%	49.39%	21.69%	-----	-----

Table 3 represents the overall readability level of three paragraphs obtained from the feedbacks of the undergraduates.

Table 3 Reading levels of three paragraphs

Paragraph	Reading Level
Paragraph 1	In-between Very Easy and Easy
Paragraph 2	Easy
Paragraph 3	Easy

Readability scores calculated for the three given paragraphs by using readability formulas are shown in Table 4.

Table 4 Readability scores calculated by using readability formulas

Paragraph	Flesch Reading Ease score	Flesch-Kincaid Grade Level	Gunning Fog Index	The Coleman-Liau Index	The SMOG Index	Automated Readability Index
Paragraph 1	22.8 Very difficult to read	14.7 College	20.9 Very difficult to read	15 college	14.5 college	14.4 college level
Paragraph 2	51.5 Fairly Difficult	10.7 Eleventh Grade	15 Hard to read	13 College	11 Eleventh Grade	12.1 Twelfth Grade
Paragraph 3	70.7 Fairly Easy	7.7 Eighth Grade	11.5 Fairly Easy	8 Eighth Grade	8.3 Eighth Grade	7.9 Seventh & Eighth Grade

According to the above given six readability formulas, readability levels (Very Easy, Easy, Moderate, Difficult and Very Difficult) for the three given paragraphs are shown in Table V. All of these readability formulas are design to measure the readability of native English speakers. But we use these formulas for Sri Lankans who are considered as nonnative English speakers. When analyzing the results shown in Table IV, we realize that these scores based on the U.S. Grade School Level. According to the U.S. Grade School Level, college level is similar to the university graduates in U.S (“Education in the United States,” 2015). Readability of U.S university graduates does not similar to readability of Sri Lankan university graduates because we are nonnative English speakers. So, we considered college level as difficult based on the Sri Lankans readability. Eleventh or twelfth grades match with easy and seventh and eighth grades match with very easy.

Table 5 Readability Levels for three paragraphs according to the readability formulas

	Flesch Reading Ease score	Flesch- Kincaid Grade Level	Gunning Fog Index	The Coleman- Liau Index	The SMOG Index	Automated Readability Index
Paragraph 1	Very difficult	Difficult	Very difficult	Difficult	Difficult	Difficult
Paragraph 2	Difficult	Easy	Difficult	Difficult	Easy	Easy
Paragraph 3	Easy	Very Easy	Easy	Very Easy	Very Easy	Very Easy

Analysis and Discussion

To do a quantitative analysis, we allocate values to match the results in Table 2 and Table 4. Values of Table 2 based on the survey results and values of the Table 4 based on the readability formulas, because of that we cannot directly compare the Table 2 and Table 4. Table 6 represents the values for the deviation and Table 7 represents the values get for the deviation when comparing Table 2 and Table 4.

Table 6 Values for the deviation

	Very Easy	Easy	Moderate	Difficult	Very Difficult
Very Easy	1	+0.75	-0.5	-0.75	-1
Easy	+0.25	1	+0.25	-0.5	-0.75
Moderate	-0.5	+0.25	1	-0.25	-0.5
Difficult	-0.75	-0.5	-0.25	1	-0.25
Very Difficult	-1	-0.75	-0.5	-0.25	1

Table 7 Values for the deviation for the three paragraphs based on the readability formulas

	Flesch Reading Ease score	Flesch- Kincaid Grade Level	Gunning Fog	The Coleman- Liau Index	The SMOG Index	Automated Readability Index
Paragraph 1	-1	-0.75	-1	-0.75	-0.75	-0.75
Paragraph 2	-0.5	1	-0.5	-0.5	1	1
Paragraph 3	1	+0.25	1	+0.25	+0.25	+0.25
Final Score	-0.5	+0.5	-0.5	-1	+0.5	+0.5

According to the final score value in Table VII, we can conclude Flesch-Kincaid Grade Level, The SMOG Index and Automated Readability Index are the most suitable readability formulas for Sri Lankan context.

Conclusion and Further Developments

Readability is the quality of writing that makes it easy or difficult to read and understand (“8 Readability Web Tools to Test Your Writing Quality,” n.d.). By using readability formulas, we can measure the difficulty of words and sentences in any given materials. From several dozen readability formulas Flesch-Kincaid and Flesch Reading Ease, Fog Index, SMOG, Coleman–Liau Formula, and Automated Readability Index (ARI) are commonly selected. Readability formulas vary based on the purpose and the calculation because of that most writers use no more than three similar formulas to measure the readability of a document. Readability

formulas estimate text difficulty based on the length of words and sentences. In this study, we calculate the readability of the three given paragraphs using selected readability formulas and compare the student feedbacks to select the most suitable formulas for Sri Lankan context. Factors that we consider when selecting the readability formulas are as follows: the suitability of the formulas for college level readers, using English as Second/Foreign Language readers, the popularity of the formula and the suitability of the formulas design for technical text or documents. We selected three computer science related paragraphs with different complexity levels. We also selected eighty-five undergraduate students who studying in Computer Science stream in a leading Sri Lankan university to participate in our survey.

Although we are having the cultural barriers because we are not native English-speaking people, based on our findings, we can conclude that Flesch-Kincaid Grade Level, The SMOG Index and Automated Readability Index are the most suitable readability formulas for Sri Lankan context. The outcome of this study is considered when we going to automate the pedagogical reviewing process for evaluating web based E-Learning contents. When we studied the manual pedagogical reviewing process, we identified Easy-to-understand Instruction is one of the most important review factors in pedagogical reviewing process for evaluating web based E-Learning content understand (de S Sirisuriya et al., 2013b).

References

- 8 Readability Web Tools to Test Your Writing Quality [WWW Document], n.d. . MakeUseOf. URL <http://www.makeuseof.com/tag/writing-reader-friendly-check-8-readability-testing-web-tools/> (accessed 1.26.15).
- Bruce, B.C., Rubin, A., Starr, K.S., 1981. Why readability formulas fail.
- de S Sirisuriya, S.C.M., Ranathunga, L., Karunanayaka, S.P., Abdullah, N., 2013a. Can pedagogical evaluation process of web based E-Learning be automated?, in: 2013 8th International Conference on Computer Science Education (ICCSE). Presented at the 2013 8th International Conference on Computer Science Education (ICCSE), pp. 650–655. doi:10.1109/ICCSE.2013.6553989
- de S Sirisuriya, S.C.M., Ranathunga, L., Karunanayaka, S.P., Abdullah, N., 2013b. Calibration checklist to use in an automated pedagogical evaluation process of web based E-Learning, in: 2013 8th IEEE International Conference on Industrial and Information Systems (ICIIS). Presented at the 2013 8th IEEE International Conference on Industrial and Information Systems (ICIIS), pp. 162–167. doi:10.1109/ICIIS.2013.6731974
- Education in the United States, 2015. . Wikipedia Free Encycl.
- Flesch–Kincaid readability tests, 2015. . Wikipedia Free Encycl.
- How To Choose The Best Readability Formula For Your Document [WWW Document], n.d. . hicow.com. URL <http://backend2/SMOG-readability/spache-readability-formula/fry-graph-1.html> (accessed 2.27.15).
- McCallum, D.R., Peterson, J.L., 1982. Computer-based Readability Indexes, in: Proceedings of the ACM '82 Conference, ACM '82. ACM, New York, NY, USA, pp. 44–48. doi:10.1145/800174.809754
- Pitler, E., Nenkova, A., 2008. Revisiting Readability: A Unified Framework for Predicting Text Quality, in: Proceedings of the Conference on Empirical Methods in Natural Language Processing, EMNLP '08. Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 186–195.
- Si, L., Callan, J., 2001. A Statistical Model for Scientific Readability, in: Proceedings of the Tenth International Conference on Information and Knowledge Management, CIKM '01. ACM, New York, NY, USA, pp. 574–576. doi:10.1145/502585.502695
- Smith, E.A., Kincaid, J.P., 1970. Derivation and Validation of the Automated Readability Index for Use with Technical Materials. Hum. Factors J. Hum. Factors Ergon. Soc. 12, 457–564. doi:10.1177/001872087001200505
- Zamanian, M., Heydari, P., 2012. Readability of texts: State of the art. Theory Pract. Lang. Stud. 2, 43–53.