1. INTRODUCTION

The great many of our efforts are centered on the transmission of information. Without that information being understood by the target audience, it is the same as that information never being communicated at all. When texts are too hard to read, readers stop reading. Readability is what makes some texts easier to read than others. George Klare defines readability as “the ease of understanding or comprehension due to style of writing.” [1] It is often confused with legibility, which concerns typeface and layout. [2] While readability can be determined by good old fashioned inspection of the text, there are algorithms that have been devised to determine the readability of a piece of text without having to read through the entire document.

The usage of analysis to determine readability has found itself within several industries. The medical industry has found that poor readability in their medication instructions affects their patients who are not able to properly take their medicine. When ensuring informed consent for legal and medicinal procedures the document must be comprehensible to the layman. Businesses lose millions a year due to the low readability of their documents, prompting inefficiency and waste. Software readability ensures ease of program maintenance and easier programming. The military seeks to streamline their technical aspects. Legal professions suffer from low readability in legal documents.

Despite the benefits of readability algorithms, there are some drawbacks as well. Most of the algorithms provide only a review of surface linguistics without judging the difficulty of the concepts being conveyed. It is also beyond the scope of the algorithms to assess the proper audience and judge the text in relation to them. That will require human judgment. Despite their faults, they provide an objective measurement for a text. [3]

This paper reviews the research and algorithms in the field of readability.

2. GUIDELINES FOR READABILITY

- Short, familiar words.
- Avoid jargon.
- Culture and gender neutral language.
- Correct grammar and punctuation.
- Active voice and present tense.
- Use imperative.
- Use visual aids.

2.1 FLESCH KINCAID TEST

One of the most prominent readability tests, the Flesch Kincaid readability test is designed to indicate the comprehension difficulty when scoring a piece of academic text. There are two versions: the Flesch Reading Ease test and the Flesch-Kincaid Grade Level. [4] Both measure the readability of the text but use somewhat different weighting so the answers correlate differently.
In the Flesch Reading Ease test, higher scores indicate that the text is easier to read. (Higher readability.) The lower the number the more difficult the passage is to read.

<table>
<thead>
<tr>
<th>SCORE</th>
<th>FOR:</th>
<th>% OF POP.</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 – 100</td>
<td>5&lt;sup&gt;th&lt;/sup&gt; grade</td>
<td>93</td>
</tr>
<tr>
<td>80 – 90</td>
<td>6&lt;sup&gt;th&lt;/sup&gt; grade</td>
<td>91</td>
</tr>
<tr>
<td>70 – 80</td>
<td>7&lt;sup&gt;th&lt;/sup&gt; grade</td>
<td>88</td>
</tr>
<tr>
<td>60 – 70</td>
<td>8&lt;sup&gt;th&lt;/sup&gt; and 9&lt;sup&gt;th&lt;/sup&gt; grade</td>
<td>83</td>
</tr>
<tr>
<td>50 – 60</td>
<td>10&lt;sup&gt;th&lt;/sup&gt; to 12&lt;sup&gt;th&lt;/sup&gt; grade</td>
<td>54</td>
</tr>
<tr>
<td>40 – 50</td>
<td>13&lt;sup&gt;th&lt;/sup&gt; to 16&lt;sup&gt;th&lt;/sup&gt; grade</td>
<td>33</td>
</tr>
<tr>
<td>0 – 30</td>
<td>College graduate</td>
<td>4.5</td>
</tr>
</tbody>
</table>

Note that this refers to the reading level of the person, not their grade level; most people have a reading level significantly lower than their grade level. The 8<sup>th</sup> grade reading level is the average the United States adults.

The formula for the FRE score is:

\[
206.835 - 1.015 \left( \frac{\text{total words}}{\text{total sentences}} \right) - 84.6 \left( \frac{\text{total syllables}}{\text{total words}} \right)
\]

*Reader's Digest* has a readability of 65. *Time* magazine scores 52. *Harvard Law Review* has a score in the low 30s. This is the same readability scale that Microsoft Word uses.

### 2.2 FLESCH-KINCAID GRADE LEVEL TEST

The second test, the Flesch-Kincaid Grade Level test, translates the readability of the text into a grade level suitable for students. It can also mean the number of years of education best for understanding the text.

The formula:

\[
0.39 \left( \frac{\text{total words}}{\text{total sentences}} \right) + 11.8 \left( \frac{\text{total syllables}}{\text{total words}} \right) - 15.50
\]

The result is a number that corresponds to grade level. A score of 11 would mean that the person would need to be at least in 11<sup>th</sup> grade to have a good chance of understanding the text. [5]

### 2.3 SMOG

SMOG, or Simple Measure of Gobbledygook, is another well-known readability test that summarizes the years of education required to understand a piece of writing. It is considered more accurate than most other readability tests.

To use the SMOG formula:
1. Count the number of sentences.
2. Count the words with 3 or more syllables in those sentences.
3. Calculate using:

\[
\text{grade} = 1.0435 \sqrt{\frac{\text{number of polysyllables}}{\text{number of sentences}}} + 3.1291
\]

### 2.4 GUNNING FOG INDEX

The Gunning Fog index is particularly suited at evaluating the readability of a text for a wide audience. It measures the readability in the number of years of formal education it would take to understand it. For example, a rating of 12 would indicate that someone would need to have the education level of at least a high school senior to easily comprehend the text upon the first reading.

The formula consists of the following algorithm:
1. Take a full passage that is around 100 words.
2. Find the average sentence length.
3. Count words with three or more syllables. (Not counting proper nouns or compound words.)
4. Add the average sentence length and percentage of complex words.
5. Multiply the result by .4.

It is a good formula but it has limitations. For example, not all words with three or more syllables are difficult, like “entertain.”
Furthermore, the algorithm has been modified to a less accurate model to make it easier to run on computers.

### 2.5 Coleman Liau Index

It was designed by Meri Coleman and T. L. Liau to gauge the understandability of a text. Like the previous formulas mentioned, it outputs the approximate US grade level necessary to comprehend a piece of text.

However, unlike most other tests, the Coleman Liau Index relies on characters, not syllables. Although there is a range of opinion on whether or not this provides a more accurate analysis, it is more computer friendly in that characters are more easily and accurately counted by computer programs than syllables and words.

To calculate the Coleman Liau Index:

\[
CLI = 5.89 \left( \frac{\text{characters}}{\text{words}} \right) - 0.3 \left( \frac{\text{sentences}}{100 \text{ words}} \right) - 15.8
\]

1. Divide the number of characters by the number of words, then multiply by 5.89.
2. Take the number of sentences in a fragment of 100 words, and multiply by .3.
3. Subtract #2 from #1 and subtract 15.8

### 2.5 Automated Readability Index

The Automated Readability Index is similar to the Coleman Liau Index and different than other readability formulas in that it focuses its algorithm on characters rather than syllables per word.

The formula is similar:

\[
ARI = 4.71 \left( \frac{\text{characters}}{\text{words}} \right) + 0.5 \left( \frac{\text{words}}{\text{sentence}} \right) - 21.43
\]

Surveys done by the military in their efforts to constantly find the best soldiers has also revealed that the ability for a reader to comprehend pieces of text correlates to their general breadth of knowledge, and a higher IQ. This means that to increase the readability level of a reader that their knowledge in general must be increased.

1. Divide the number of characters by the number of words, then multiply by 4.71.
2. Divide the number of words by the number of sentences, then multiply by .5.
3. Add the two together then subtract 21.43.

The output approximates the US grade level required to understand a piece of text, like the others.

### 3. Research

Much effort has gone into surveying the scope and breadth of readability in modern society and the impact and influence it has within it.

It has proven to be an integral part of many industries, such as:

- Businesses: Millions of dollars is lost each year between the shuffling of paperwork throughout the bureaucracy in obtusely written material and in the inefficiency of clarifying matters.
- Medical: For medication to be properly taken and have the desired effect, the readability must be high. [6] [7]
- Legal: In signing contracts and ensuring informed consent, the legal industry finds itself in need of clear and concise communication.
- Programming: The readability of a program is a key part of its maintenance in the SDLC. [8]
Those who read at a higher level were also found to exhibit better work habits, be more efficient, and generally more intelligent. Increasing the readability level of the user also increased these characteristics. However, performing such a colossal undertaking, while being performed in schools to a certain degree of success, is one method. The more efficient and surefire method of ensuring understanding is to bring the text to the reader’s level to ensure they understand it, since not everybody can be raised up.

4. PROPOSAL

I wish to create a readability formula that can assess the difficulty in reading a topic when taking into account topical complexity. The formula will take a text as an input like a normal readability formula and examine the readability of the text. However, it will not stop there. It will have access to a database that stores a dictionary size allotment of words. Each word is weighted in relation to its topical complexity. For example, the word “apple” will receive a relatively low weight whereas the word “quark” will have one that is significantly higher despite being a simple word. Thus a paper that is superficially easy to read can be assessed as more difficult when the concepts it is dealing with are brought to light. Once it takes the weights of these words the readability formula then finds an average weight so that represents the average readability of the paper. The database can be tuned to represent a certain field instead of a breadth of knowledge. For example, if the text in question is one dealing with Chemistry, then the database the formula uses can be one limited to Chemistry terms to increase efficiency and reduce the workload.

Readability is an important problem in all fields. Being able to assess the topical complexity in addition to the linguistic sophistication of a piece of text will aid proper communication in all areas. In terms of software, the readability formulas can even be tweaked to assess the complexity of code, using a database weighted to analyze the difficulty level of the program presented before it. Common practices would rate high readability while obscurely constructed code would rate low.

Readability formulas have been produced and tweaked for decades. [9] They are all adept at performing readability exercises determining the readability of a piece of text on the basis of the makeup of the words. It is acknowledged that they lack the ability to assess topical complexity and so the previous works provide a convenient basis upon which to improve.

I will conduct a study of the currently existing readability algorithms and determine the ones best suited to this particular task. After determining the most efficient and most appropriate formula I will create a database for it to access as well as modify the code or create my own algorithm to utilize it. Then I will run texts of various difficulty and topics through the algorithms I have made and test them for how closely they relate to the actual difficulty of the topics they discuss.

After conducting a thorough research of the readability field I am knowledgeable of how the algorithms work and the principles behind readability as well as the readability formulas that would be best suited to the project. Furthermore, I have taken many computer science courses that have given me an understanding of the mathematics and programming that is required for each readability formula. As such I will be able to deconstruct, understand, and reconstruct each formula in a way to benefit the project and realize the final goal. I am also an avid reader so I have a broad view of the range of the English language and will be able to easily
evaluate the difficulty of various concepts in weighting them.

For the first month I will focus on evaluating current readability formulas and seeing how they could conceivably be changed to accommodate a database from which to access words to assign a readability score in terms of topical complexity to a piece of text. I will also explore the weighting system to use in a database and how large the database can be without compromising speed and efficiency to an unacceptable degree. During the second month I will implement the readability algorithm to assess the complexity of a topic and interface it with the database. By the end of that month there should be a formula that can accept a piece of text as an input, and give the readability score on the superficial linguistic level as well as the difficulty of the subject.

REFERENCES


[6] "Readability Assessment of British Internet Information Resources on Diabetes Mellitus Targeting Laypersons,” Bath University, 2004


[8] Emilio Collar, Ricardo Valerdi, “Role of Software Readability on Development Cost,” Massachusetts Institute of Technology