Estimating Readability with the Strathclyde Readability Measure

George R S Weir and Calum Ritchie

Department of Computer and Information Sciences
University of Strathclyde
Glasgow G1 1XH

Abstract
Despite their significant limitations, readability measures that are easy to apply have definite appeal. With this in mind, we have been exploring the prospects for more insightful measures that are computer-based and, thereby, still easily applied. The orthodox reliance on intrinsic syntactic features is an inherent limitation of most readability measures, since they have no reference to the likelihood that readers will be acquainted with the constituent words and phrases. To accommodate this feature of 'human familiarity', we have devised a metric that combines traditional factors, such as Average Sentence Length, with a measure of word 'commonality' based upon word frequency. This paper details the derivation, nature and application of the Strathclyde Readability Measure (SRM).

1. Introduction

Readability measures are heuristic-based metrics that commonly derive a readability value from 'intrinsic' textual characteristics of a document. The usual presumption (the heuristic) is that texts with short sentences and short words are more easily read and comprehended by the average reader. With this in mind, readability measures usually focus on such document features as average sentence length (ASL) and average word length (AWL). Although for many texts this assumption is appropriate, there are also many instances of texts with short but complex words and short but obscure sentences. For such examples, most fog indices would award 'good' readability scores, although most human readers may consider these texts 'difficult'.

According to Connaster (1999), 'readability formulas fail to predict text difficulty' and he cites an experiment to demonstrate 'that "text difficulty" is 'a perception of the reader and therefore cannot be objectively calculated by counting syllables, word length, sentence length, and other text characteristics'. This caution is repeated by Oakland and Lane (2004), who advise that the use of readability measures 'should be confined to paragraphs and longer passages, not items'.

For humans, the readability of any text is primarily a function of the textual content and the reader's knowledge and language experience. The apparent need to accommodate the human reader's perspective in such measures, leads Oakland and Lane (op. cit.) to recommend the use of 'Readability methods that consider both quantitative and qualitative variables and are performed by seasoned professionals'. In our approach, we aim to include a factor within our readability measure that goes some way to accommodating the likely familiarity of the words in any sample text. This is based upon the frequency or commonality of words and represents a move toward the 'human reader's perspective' by going beyond purely syntactic features toward the semantic impact of textual content.

2. Deriving the Strathclyde Readability Measure

As a starting point for our readability measure, a number of subjects agreed to complete a survey questionnaire. A key feature of this questionnaire addressed the perceived relative readability of sample passages. Although this 'text levelling' process is traditionally performed by experts in readability (DuBay, 2004, p.45), our informal approach used a small group of volunteers from varied backgrounds.

Passages were selected from several classic novels, each with a different level of reading complexity. The test subjects were asked to rank the texts in order of readability. This provided a comparison against which to consider existing readability measures. Nine text samples were ranked by the human readers and gave the result shown in Table 1 (with more difficult to the top of the table).

9	A Tale of Two Cities
8	The Brothers Karamazov
7	Notes from the Underground
6	Great Expectations
5	Crime and Punishment
4	Moby Dick
3	Around the world in 80 days
2	Alice's Adventures in Wonderland
1	Little Women

Table 1: Ranking according to perceived reading complexity

To these same test passages we applied two common readability measures, Flesch-Kincaid Grade Level (FKGL) and Flesch-Kincaid Reading Ease (FKRE). The results of these tests produced two further rankings for the sample texts. Table 2 shows the text rankings (in decreasing order of complexity) as indicated by the Flesch-Kincaid Grade Level measure (the FKGL score is also shown).

9	A Tale of Two Cities	20.99
8	The Brothers Karamazov	15.65
7	Around the world in 80 Days	11.16
6	Alice's Adventures in Wonderland	9.25
5	Moby Dick	8.96
4	Crime and Punishment	8.13
3	Great Expectations	7.32
2	Notes from the Underground	7.03
1	Little Women	6.78

Table 2: Ranking according to Flesch-Kincaid Grade Level

While this ranking agrees partially with the human readers' perceived reading complexity, there is significant discrepancy.

We next applied the Flesch-Kincaid Reading Ease measure to these same test passages. Table 3 shows the text rankings (in decreasing order of complexity). The FKRE score is also shown. As well as showing variation from the rankings derived from the human readers, the FKRE also differs from the FKGL ranking.

9	A Tale of Two Cities	30.67
8	The Brothers Karamazov	37.97
7	Around the world in 80 Days	55.31
6	Moby Dick	65.58
5	Crime and Punishment	66.38
4	Notes from the Underground	70.23
3	Alice's Adventures in Wonderland	70.68
2	Great Expectations	74.90
1	Little Women	77.09

Table 3: Ranking according to Flesch-Kincaid Reading Ease

These small-scale tests indicate that measures such as FKGL and FKRE do not fully reflect the human readers' impressions for the sample texts. Furthermore, two interesting points emerged from our interviews. Firstly, the extract from 'Around the World in 80 Days' scored high in complexity on the two traditional measures, as the text has a longer average sentence length than the rest of the texts. However, readers felt that the text was not difficult to understand as the words used were familiar and the text was written in a clear manner.

Secondly, the extract from 'Notes from the Underground' earned a low score on the traditional measures, indicating an easy passage. This results from a low average sentence length. In contrast, readers felt that this characteristic combined with fairly complex words in the text, made the document harder to read. Neither of these factors, identified by the human readers, is accounted for by the traditional measures.

3. Characteristics of the Strathclyde Readability Measure

Our new measure aimed to satisfy a number of criteria. The output should be easily understood. Since sentence length affects readability, this should be included in the measure. A method of measuring word complexity is desirable and usually this is based on word length – relying on the view that more common words tend to be shorter. While it is reasonable that common words tend to be short, there are likely to be many instances of short words that are not common and, contrariwise, long words that are very familiar. In consequence, we employ a commonality measure that is based not on word length but on frequency of occurrence relative to the British National Corpus.

Output from the new measure is similar to other approaches, wherein a numerical score is generated between 0 and 100. In this instance, the score position on the scale directly indicates the relative complexity of the considered text. This means that no external reference for complexity - such as an educational grade level - is presumed.

Instead of using average sentence length as a variable in our measure, we use a constant based on ASL. This ensures that passages with similar word complexities and with an ASL that differs slightly will generate scores that are closely related. Our assumptions are:

- if the ASL is low or high and the word complexity is high, then this should generate a harder score than a text with the same word complexity but with a more readable average sentence length;
- if the ASL is low and the word complexity is low, then an easier score should be generated than a text with the same word complexity but a longer ASL;
- if the ASL is high, then the measure should generate a harder score, regardless of word complexity.

The complexity of words is derived using a frequency list. This provides an indication of word commonality, which we regard as indicative of the word's likely perceived difficulty. Our frequency list contains approximately one million words and is based on the British National Corpus (cf. Leech et al, 2001). Two measures have been created that are similar but differ in their use of the frequency list in determining word complexity.

3.1 Strathclyde Readability Measure 1

{ log(AWF x 2) x k } – 80

Where:

- AWF = the average word frequency, only counting words with a frequency less than 100,000.
- k = a constant based on the average sentence length (ASL)
    - 15: if the ASL is greater than or equal to 17 and less than 25, or the ASL is less than 17 and the AWF is greater than 95000.
    - 13: if the ASL is less than 17 or greater than or equal to 25.

3.2 Strathclyde Readability Measure 2

100 – { [(cc x 100) x k] x 3 }

Where:

- cc = number of complex words divided by the total number of words in the passage. Where a complex word is a word with a frequency less than 100.
- k = a constant based on the average sentence length (ASL)
    - 3: if the ASL is greater than or equal to 17 and less than 25, or the ASL is less than 17 and the AWF is greater than 95000
    - 5: otherwise.

4. Estimating readability

The use of these two measures is similar, the key difference being their intended application to different sizes of text samples (SRM2 is intended for application to small groups of sentences or even single sentences). Scoring on the 100 point scale is the same for each measure and we have estimated degrees of readability based on this range (see Table 4).

< 30	Extremely difficult
30 – 40	Difficult
40 – 50	Hard
50 – 65	Average
65 – 80	Easy
> 80	Very easy

Table 4: Estimates of readability on the SRM scale

The main operational difference between our two methods lies in the ways that the measures use the frequency list. The first measure totals the frequency of words whose frequency is less than 100,000, and then divides that total by the total number of words in the passage. If the passage is short (less than 100 words), then it may contain no words with a frequency less than 100,000. This means that the AWF value is 0. This would indicate an extremely hard passage, even though it is not.

However, this measure has proved the most accurate out of the two for larger passages.

The second measure overcomes this problem by using a count of the number of complex words (a frequency less than 100), so the higher the number of complex words, the less readable is the passage.

5. The SRM software application

To ease application of the SRM, a Java-based application was implemented. This program has a text editor style interface that allows multiple texts to be open at the same time by arranging them into tabs. Figure 1 shows a screen shot of the main view for the system.

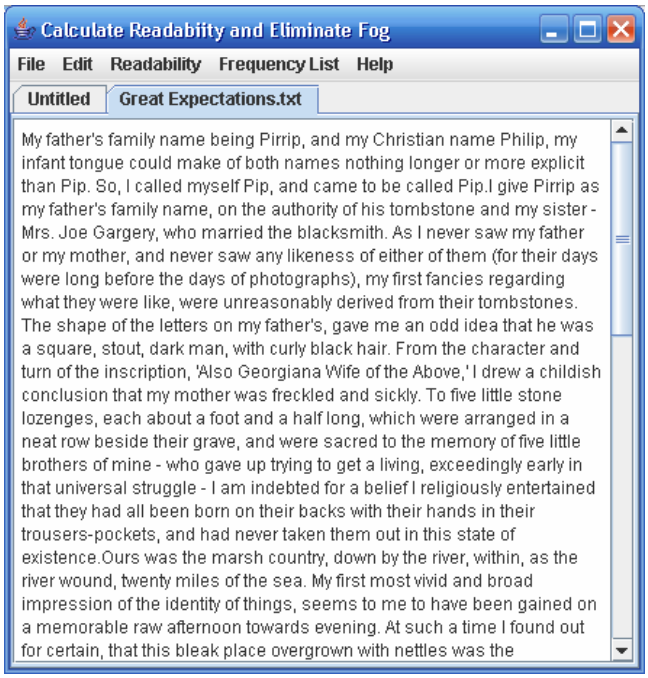

Figure 1 – Screen shot of main SRM window

The tabs provide an easy way of selecting multiple texts to compare the difficulty of. The functionality of the system will be accessed using the menus at the top of the screen. The menus are structured as follows:

- File menu
    - New – will open a new tab with a text area and a title entered by the user.
    - Save – will save the contents of the currently selected tab's text area to a location on disk specified by the user.
    - Load – will set the contents of the currently selected tab's text area to the contents of a text file on disk selected by the user.
    - Close Tab – will close the currently selected tab, asking for confirmation before doing so.

- Quit Application – will quit the application. If there are still tabs open, then the system will ask for confirmation before exiting.

- Edit menu
  - Copy – Copy the selected text in the current tab to the OS clipboard
  - Paste – Paste the text in the OS clipboard to the current tab's textpane at the cursor position

- Readability menu
  - Calculate Readability – this is a submenu that will contain the measures. Clicking on the measure in the sub menu will perform the measure for the text in the currently selected tab and displays the result in a message dialog similar to the one shown in Figure 2.

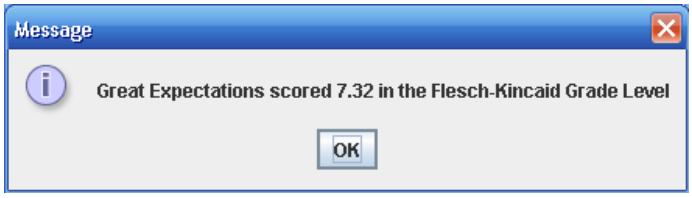

Fig 2 –Readability measure result

  - Compare texts – will compare the difficulty of multiple texts and rate them in a table.
  - Check Sentences With Poor Readability – will indicate sentences in the text that have poor readability.

- Frequency List menu
  - Switch list – will allow the user to switch the type of frequency list that is being used by the application.
  - Get Word Frequency – will get a word from the user via an input dialog then search for the word in the frequency list. The frequency of the word will be displayed if it is found; otherwise an error dialog is displayed.

5.1 Comparing texts

Using the SRM system, it is possible to load multiple texts and directly compare their readability. This is done by opening multiple tabs and populating them with text then selecting the Compare Texts option from the Readability menu. For example, see Figure 3.

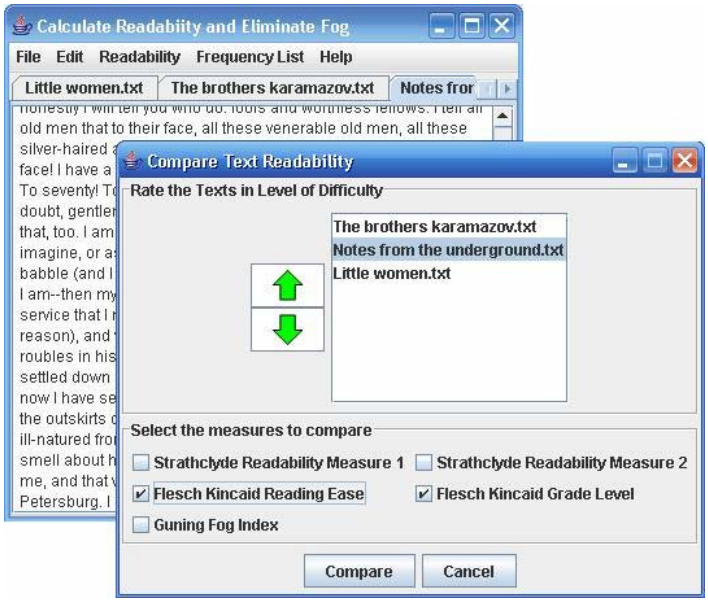

Figure 3: Compare text readability screen

The user may use the arrows to arrange the texts in order of difficulty. They will then select the measures that they wish to apply to the texts and click Compare which will perform the comparison and generate the output shown in Figure 4.

Figure 4: Test rating table

The list is arranged in the order specified by the user and the position that the measure placed the text, is indicated in the columns, where 1 is the hardest. If the cell is

- Green indicates that the measure rated the text in the same position as the user;
- Yellow indicates that the measure was close to the user in its rating of the text;
- Red indicates significant distance from the user rating of the text.

Since this is intended to allow an easy visual comparison of text complexity measures, the actual values derived for each measure is not shown in this display. The actual value for any of these tests is obtained outwith the text comparison by applying a selected readability measure to a specific text sample (via the readability menu option).

6. Conclusions

This paper describes the rationale behind our development of the Strathclyde Readability Measure and accompanying software application. We believe that our approach goes beyond conventional fog index measures of readability by employing an extrinsic factor in deriving our results. This factor is based upon word frequency,

relative to the British National Corpus, and aims to provide an estimate of 'likely familiarity' for the words contained in sampled texts.

The SRM formula appears to work well for most sampled texts, with higher scores indicating easier passages. On this measure, an intermediate text will score around 50. Table 4 shows the sample texts and their SRM values.

Text	**SRM1**
A Tale of Two Cities	43.61
The Brothers Karamazov	46.33
Notes from the Underground	47.39
Great Expectations	64.85
Crime and Punishment	44.32
Moby Dick	64.92
Around the World in 80 Days	62.46
Alice's Adventures in Wonderland	69.85
Little Women	66.88

Table 4: Ranking according to Strathclyde Readability Measure

In our small evaluation, compared to traditional approaches, such as Flesch-Kincaid Grade Level and Flesch-Kincaid Reading Ease, our measure shows a better degree of match to rankings of readability provided by human readers. Nevertheless, there is still a marked variation between the ranking of the readers and of the SRM. This is likely to result from factors sensible to the readers but invisible to the readability measure, e.g., writing style and subject matter. In addition, since our reader survey was rather small, this may also skew the human ranking result. This indicates a need for larger scale empirical tests to determine with greater confidence the degree of match between readers and the Strathclyde Readability Measure.

Finally, we should note that the accompanying software implementation, allows for easy application of the Strathclyde Readability Measure, as well as comparisons with several standard fog indices. Thereby, we maintain the convenience of easy application commonly associated with traditional intrinsic readability measures, whilst adding the greater credibility afforded by the extrinsic factor of word commonality.

References

Connaster, B.R., 'Last rites for Readability Formulas in Technical Communication', *Journal of Technical Writing and Communication*, Vol. 29, No. 3, 1999, 271– 287.

DuBay, W.H., The Principles of Readability. Costa Mesa, CA: Impact Information, 2004. URL: http://www.impact-information.com/impactinfo/readability02.pdf Last Accessed 13 Jul. 2006

Leech, G., Rayson, P. and Wilson, A., *Word Frequencies in Written and Spoken English: based on the British National Corpus*, 2001, Longman, London.

Oakland, T. and Lane, H.B., 'Language, Reading, and Readability Formulas: Implications for Developing and Adapting Tests', International Journal of Testing, 2004, Vol. 4, No. 3, 239-252