Computational Text Analysis: A More Comprehensive Approach to Determine Readability of Reading Materials

Anealka Aziz, Chan Yuen Fook and Zubaida Alsree

doi:10.7575/aiac.alls.v.1n.2p.200

Introduction

Reading materials are considered having high readability if readers are interested to read the materials, understand the content of the materials and able to read the materials fluently. In contrast, reading materials with low readability discourage readers from reading the materials, create difficulties for readers to understand the content of the materials and prevent readers to read the materials fluently.

Studies on readability have started since the early 1920s. These studies seek for measures that can best predict the readability level of reading materials, so that readers are able to comprehend and learn new information from these materials (Harris-Sharples, 1983). If the measures could be identified, the difficulty level of reading materials could be determined. Once the readability level of reading materials is determined, at least half of the matching problem can be solved. Hence, it is important to ensure a match between readers and reading materials as this match determines how much readers can benefit from the materials they read (Gilliland, 1972).
Readers with limited language ability can easily be discouraged to continue their reading if they are given reading materials beyond their language ability. Similarly, competent readers may soon be discouraged from reading, if their choices of reading materials are restricted to simple repetitive ones. Readers in both cases may not benefit as much from the reading materials they read because the materials are poorly matched to their language ability.

Many of the factors that affect readability of reading materials have yet to be quantified. Nevertheless, Bailey (2002) reports that many studies have shown that readability of reading materials is highly correlated with two factors that can be easily measured: sentences and words. Chavkin (1997) identifies that the most strongly associated factors to readability are word difficulty and sentence length. These two factors or variations of these two factors can be found in all readability formulas currently in use (Chavkin, 1997). Studies have confirmed that inclusion of other factors in the formula contributes more work than it improves the results (Stephens, 2000). It shows that readability of reading materials can sufficiently be measured using word difficulty, sentence length and variations of the two. There is no need to include factors other than word difficulty, sentence length or the variations of the two.

As mentioned earlier, readability of reading materials is related to sentence and word factors of the materials. One of the measures of sentence difficulty is length of sentences (Gunning, 1971; MacGinitie & Tretiak, 1971; Klare, 1985; Grabe, 1993; Shehadeh & Strother, 1994; Chavkin, 1997; Johnson, 1998; Bailey, 2002; Thornbury, 2005; Mesmer, 2008). Long sentences contain many relationships
which require learners to infer more information than shorter sentences (Mesmer, 2008). Although not all long sentences are difficult to understand, reviewing is useful as length and difficulty tend to be related (Klare, 1985). This is because longer sentences require the mind to hold more information in suspense before the mind can make sense of the meaning of these words together (Flesch, 1979).

Sentence construction is another measure of sentence difficulty (Gunning, 1971; Klare, 1985). Complex sentence structures may contain more embedded sentences and more word depth, which have the tendency to be misinterpreted (Klare, 1985). It is uncommon in English language for sentence constructions to have more than two embedded clauses (Klare, 1985). The use of modifiers may reduce the difficulty of the sentences caused by these embedded clauses.

However, sentences with too many modifiers increase the word depth of the sentences. Word depth, which refers to the ‘commitments the words have as part of sentences’ (Klare, 1985, p. 103), can make a sentence difficult. One way to reduce the word depth is by breaking a sentence into several shorter sentences.

Besides sentence difficulty, word difficulty is another contributor to materials readability. As claimed by Chall (1958), Laufer (1997), Nation and Coady (1998), and Carter and McCarthy (1988), word difficulty in reading materials is the most significant predictor of overall materials difficulty. Word difficulty can be determined by looking at the word frequency, word familiarity, word length (Gunning, 1971; Chall, 1981; Klare, 1985; Nation & Coady, 1998).
High frequency words tend to be short and learners are likely to encounter these words more often than the low frequency words (Gunning, 1971; Thornbury, 2002; Gunning, 2003). These words make up the majority of tokens in any discourse (Schmitt, 2000) and in fact, the knowledge of the first 2000 most frequent words in the language allows learners to access to approximately 87% of any ordinary text (Nation, 1990). In the case of the second language learners, they need to know the 3,000 high frequency words of the language (Waring & Nation, 1997) as knowing these words enable them to begin reading authentic texts (Nation, 1990; Schmitt, Schmitt & Clapham, 2001). Knowing these words, also, enable them to “make accurate guesses about the meanings of the remaining less frequency words which are likely to be unknown” (Schmitt, 2000).

Aim of the Study
This study intends to propose a more comprehensive approach to analyze reading materials so that not only the overall readability of the materials can be determined, but information about sentence and word difficulty as well.

Procedure: Assembling the Composite Computational Tools for Text Analysis
This study is interested to analyze reading materials at three levels: text, sentence and word levels. At text level, the study looks at the readability scores of the materials as the overall text difficulty. At sentence level, the study looks at average sentence length, the use of simple and compound sentences and the use of complex and compound-complex sentences as predictors of sentence difficulty. At word level, the study looks at average word length and the coverage of the first
2000 high-frequency words as predictors of word difficulty. Information related to text, sentence and word difficulty is available within the reading materials themselves.

Three computational tools are used to extract information related to readability of reading materials at the three levels. A readability formula is used to estimate materials difficulty at text level, writing enhancement software is used to estimate materials difficulty at sentence level and concordance software is used to estimate materials difficulty at word level. Several readability formulas, writing enhancement software and concordance software are compared to determine the best possible computational tools for this study.

Readability Formulas

A comparison of several readability formulas used in Hamsik (1984), Brown (1998) and Greenfield (1999) studies is done before deciding on the formula to be used for the study. Readability formulas in these three studies are chosen as candidate formulas because these studies have tested the validity of these formulas on ESL/EFL learners. The common readability formulas found in at least two of these three studies are the Flesch Reading Ease Formula, the New Dale-Chall Readability Formula, the Fry Readability Graph and the Flesch-Kincaid Grade Level Formula. The Fry Readability Graph, however, is excluded from the comparison as it uses a graph in estimating passage difficulty level. Only the Flesch Reading Ease Formula, the Flesch-Kincaid Grade Level Formula and the New Dale-Chall Readability Formula are compared in detail.
The New Dale-Chall formula, despite its popularity in estimating reading grade of written materials, is not as accessible as the Flesch Reading Ease and Flesch-Kincaid Grade Level formulas. These two formulas are available automatically in Microsoft Word application once activated. This reason has excluded the New Dale-Chall Readability Formula from being shortlisted. Besides that, the formula uses the count of ‘hard’ words which refer to words outside the 3000 familiar words known to the U.S. fourth graders, which is very specific. These 3000 words may consist of words which are not familiar to the ESL learners.

The Flesch-Kincaid Grade Level Formula is a modified version of the Flesch Reading Ease Formula and it is best used to estimate readability of technical documents. The scale used to measure readability is based on the US grade level scale which may not be significant to ESL learners. These reasons have made the Flesch Reading Ease formula the best candidate to estimate readability of reading materials in this study. The formula has been validated to be used with ESL learners and is also available in Microsoft Word application. The scale used to measure readability is based on scores between zero and one hundred, which is more adaptable than using the grade level scale. Table 1 summarizes features of each readability formula.
Table 1: Features of the Flesch Reading Ease, Flesch-Kincaid Grade Level and the New Dale-Chall Readability Formulas

<table>
<thead>
<tr>
<th>Formula</th>
<th>Flesch Reading Ease Formula</th>
<th>Flesch-Kincaid Grade Level Formula</th>
<th>New Dale-Chall Readability Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Year Developed</td>
<td>1948</td>
<td>1976</td>
<td>1995</td>
</tr>
<tr>
<td>Created By</td>
<td>Rudolf Flesch</td>
<td>Rudolf Flesch &amp; John Kincaid</td>
<td>Edgar Dale &amp; Jeanne S. Chall</td>
</tr>
<tr>
<td>Predictive Variables</td>
<td></td>
<td>Average Sentence Length</td>
<td>Average Sentence Length</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average Syllable Length</td>
<td>Percentage of words not found in the list of 3,000 words</td>
</tr>
<tr>
<td>Scale Type</td>
<td>0-100 scale</td>
<td>US Grade Level Scale</td>
<td></td>
</tr>
<tr>
<td>License</td>
<td></td>
<td>Open System. No license required.</td>
<td></td>
</tr>
<tr>
<td>Operation Type</td>
<td>Automatic Calculation (Available in Microsoft Word)</td>
<td>Manual Calculation</td>
<td></td>
</tr>
</tbody>
</table>

**Writing Enhancement Software**

At sentence level, the study requires information related to sentence length and the use of different types of sentences in the materials. The best option is to use writing enhancement software as it usually provides suggestions to writers on how to improve the quality of their writing through revision. Revision requires changes to be done mostly at sentence level and sometimes at word level. The three top ranking enhancement software in the Writing Software Review (http://writing-enhancement-software-review.toptenreviews.com), Writer’s Workbench, Editor and WhiteSmoke, are evaluated to determine the software that can fulfill the need of the study.

All the three software can perform editing functions like scanning the text for mis-spelt words, checking grammar related problems and giving suggestions to correct the problems. Besides that, the software can evaluate ambiguous statements and the meaning of words to determine whether the sentence or selection makes sense...
or not. The software can also highlight phrases that use more words than what is needed to convey a message, check word redundancy, point out passive sentences and offer suggestions to change them to active sentences. Not only that, the three software can also give additional adverb or adjective suggestions to add character or variety to the sentences.

In terms of feedback, *Writer’s Workbench* and *Editor* outdo *WhiteSmoke* in providing explanation for editing and detecting syntax and subject/verb agreements. Feedback given by *Writer’s Workbench* is in the form of numbers, percentages and descriptive suggestions. These types of feedback make it more objective in analyzing reading materials as opposed to *Editor* and *WhiteSmoke*. Feedback for the other two software is in the form of a comparison with other databases and it requires writers to make the final decision whether to accept or reject the suggestions.

In terms of referencing tools, *Writer’s Workbench* outdoes the other two by having one extra feature, the *Grammar Guide*, besides a built-in dictionary and thesaurus. The *Grammar Guide* includes basic grammar information as well as advanced grammar or style guides for writers to refer while writing.

From this comparison, *Writer’s Workbench* seems to be able to provide the analysis needed by the study. As mentioned earlier, the study needs software that can provide quantifiable information on the average sentence length and the types of sentences in used in the reading materials. This extra ability of the software has
made it the best candidate to analyze the materials. Table 2 shows the summary of some of the features of *Writer’s Workbench*, *Editor* and *WhiteSmoke*.

<table>
<thead>
<tr>
<th>Software</th>
<th>Features</th>
<th>Reference Tools</th>
<th>Types of Feedback</th>
<th>Software Compatibility</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Address</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writer’s Workbench</td>
<td>EMO Solutions.com</td>
<td>Serenity Software.com</td>
<td>NA</td>
<td>Microsoft Word</td>
</tr>
<tr>
<td>Editor</td>
<td>Check misspelt words, grammar use, wordy phrases, word redundancy, passive verbs and overused words in text.</td>
<td>Dictionary and Thesaurus</td>
<td>Comparison with numerical standard</td>
<td></td>
</tr>
<tr>
<td>WhiteSmoke</td>
<td>Offer grammar and word choices.</td>
<td>Grammar Guide</td>
<td>Comparison with database</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Provide explanation for editing and able to detect syntax and subject/verb agreements</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Concordance Software**

At word level, information related to the average word length and the coverage of the first 2000 high-frequency words is required. This present study involves comparing the corpus of the materials and the list of the first 2,000 high-frequency words in the *BNC World* (2000). Therefore, the study needs software that can perform a comparison between at least two sets of corpora. Concordance software would be best to serve the purpose of this study. Furthermore, the use of concordance software in text analysis is not new as it makes the evaluation of texts more objective and less dependent on subjective judgment (Berber-Sardinha, 1999).
A comparison of three concordance software which are marked as ‘suitable’ for text analysis purposes in a review by Mukundan (2004), is performed. The three software are Concordance 3.0, TextQuest 1.37 and WordSmith Tools 3.0. The comparison, however, uses the later version of the software: Concordance 3.2, TextQuest 3.0 and WordSmith Tools 4.0.

The three software, Concordance 3.2, TextQuest 3.0 and WordSmith Tools 4.0 are capable of generating text statistics, performing frequency analysis and displaying concordance lines. However, Concordance 3.2 lacks the ability to provide readability analysis, KWIC analysis and vocabulary growth analysis. Wordsmith Tools 4.0 has an extra advantage over Concordance 3.2 and TextQuest 3.0 as it is able to display the concordance plot and perform a comparison of different wordlists at the same time.

As mentioned earlier, the study requires software that could perform a comparison between at least two sets of corpora. From this comparison, WordSmith Tools 4.0 seems to be the best candidate that can provide the type of analysis required by the study. Table 3 summarizes some of the important features of Concordance 3.2, TextQuest 3.0 and WordSmith Tools 4.0.
Table 3: Features of Concordance 3.2, TextQuest 3.0 and WordSmith Tools 4.0.

<table>
<thead>
<tr>
<th>Software</th>
<th>Concordance 3.2</th>
<th>TextQuest 3.0</th>
<th>WordSmith Tools 4.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web Address</td>
<td><a href="http://www.concordancesoftware.co.uk">www.concordancesoftware.co.uk</a></td>
<td><a href="http://www.textquest.de">www.textquest.de</a></td>
<td><a href="http://www.lexically.net/wordsmith">www.lexically.net/wordsmith</a></td>
</tr>
<tr>
<td>Features</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Text Statistics</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Frequency Analysis</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concordance lines</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NA</td>
<td></td>
<td>NA</td>
<td>Concordance Plot</td>
</tr>
<tr>
<td>NA</td>
<td></td>
<td>NA</td>
<td>Detail Consistency Analysis</td>
</tr>
</tbody>
</table>

The Composite Computational Tools for Text Analysis

Based on the comparison performed earlier, three computational tools namely the Flesch Reading Ease formula, Writer’s Workbench 8.18 and WordSmith Tools 4.0 are selected to extract the relevant information from the materials.

At text level, the Flesch Reading Ease formula is used as a tool to obtain the overall materials readability indicated by the Flesch Reading Ease score (FRE) scores. At sentence level, Writer’s Workbench 8.18 is used to obtain information on the average sentence length (ASL) and the types of sentences such as simple sentences (S), compound sentences (Cd), complex sentences (Cx) and compound-complex sentences (CdCx) used in the materials. At word level, Wordsmith Tools 4.0 is used to obtain information on the average word length (AWL) and the coverage of high-frequency words (HFW) of English in the materials.
The Flesch Reading Ease Formula

The Flesch Reading Ease (FRE) formula is used in this study to analyze reading materials at text level, as it is the most widely used, most tested, reliable instrument of materials difficulty (Chall, 1958; Klare, 1969; Hamsik, 1984; Greenfield, 1999) and is available in any writing enhancement software and Microsoft Office word processor. Hamsik (1984), Greenfield (1999; 2004) and Shokrpour (2005) also confirm that FRE is valid and can be used to determine readability level of English language materials for ESL/EFL readers. The formula takes into consideration the average sentence length and the average syllables per word in determining the readability of a passage. In the FRE formula below, the FRE score generated by the Microsoft Word application is used in this study and no manual calculation of FRE is involved:

\[
\text{Flesch Reading Ease (FRE)} = \frac{206.835 - 1.015}{\left(\frac{\text{total words}}{\text{total sentences}}\right)} - \frac{84.6}{\left(\frac{\text{total syllables}}{\text{total words}}\right)}
\]

Table 4 shows the description of the scores and the estimated reading grade (Flesch, 1948). A reading material with a score between 90 and 100 is considered as ‘Very Easy’ and can be understood by a fifth grader. A reading material with a score between 80 and 90 is considered as ‘Easy’ and can be understood by a sixth grader. A reading material with a score between 70 and 80 is considered as ‘Fairly Easy’ and can be understood by a seventh grader. A reading material with a score between 60 and 70 is considered as ‘Standard or Plain English’ and can be understood by an eighth and ninth grader. This level is also appropriate for an average person with an average education level (Flesch, 1948). Reading material
with a score between 50 and 60 is considered as ‘Fairly Difficult’ and can be understood by a high school sophomore to senior. Reading material with a score between 30 and 50 is considered as ‘Difficult’ and can be understood by students studying in college. Finally, reading material with a score between 0 and 30 is considered as ‘Very Difficult’ and can be understood by those who have graduated from college. The FRE formula is used in this study to analyze reading materials at text level.

<table>
<thead>
<tr>
<th>Reading Ease Score</th>
<th>Style Description</th>
<th>Estimated Reading Grade</th>
</tr>
</thead>
<tbody>
<tr>
<td>90 – 100</td>
<td>Very Easy</td>
<td>5&lt;sup&gt;th&lt;/sup&gt; Grade</td>
</tr>
<tr>
<td>80 – 90</td>
<td>Easy</td>
<td>6&lt;sup&gt;th&lt;/sup&gt; Grade</td>
</tr>
<tr>
<td>70 – 80</td>
<td>Fairly Easy</td>
<td>7&lt;sup&gt;th&lt;/sup&gt; Grade</td>
</tr>
<tr>
<td>60 – 70</td>
<td>Standard /Plain English</td>
<td>8&lt;sup&gt;th&lt;/sup&gt; and 9&lt;sup&gt;th&lt;/sup&gt; Grade</td>
</tr>
<tr>
<td>50 – 60</td>
<td>Fairly Difficult</td>
<td>10&lt;sup&gt;th&lt;/sup&gt; to 12&lt;sup&gt;th&lt;/sup&gt; Grade (High School Sophomore to Senior)</td>
</tr>
<tr>
<td>30 – 50</td>
<td>Difficult</td>
<td>In College</td>
</tr>
<tr>
<td>0 - 30</td>
<td>Very Difficult</td>
<td>College Graduate</td>
</tr>
</tbody>
</table>

*Table 4: Description of FRE Scores and Grade Level (Flesch, 1948)*

**Writer’s Workbench 8.18 (WWB)**

WWB is used in the study to assist the analysis of reading materials at sentence level. At sentence level, WWB *Style Statistics with Support* analysis tool is used because it offers numerical information and evaluation statements on average sentence length and the types of sentence used in the reading materials (S/Cd and Cx/CdCx). WWB suggests that the ASL of a good piece of writing is around 18 to 26 w.p.s. and the use of S/Cd should be less than 50%, while the use of Cx/CdCx should be more than 50% but less than 70% of the whole sentences in the piece of writing (WWB Manual, 2009). Table 5 shows the standard recommended by WWB 8.18.
Table 5: Standard of ASL, S/Cd, Cx/CdCx Recommended by WWB 8.18
(WWB Manual, 2009)

<table>
<thead>
<tr>
<th>Text Characteristics</th>
<th>WWB Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>ASL</td>
<td>18 – 26 w.p.s.</td>
</tr>
<tr>
<td>S/Cd</td>
<td>X &lt; 50%</td>
</tr>
<tr>
<td>Cx/CdCx</td>
<td>50% &lt; X &lt; 70%</td>
</tr>
</tbody>
</table>

Note: ASL = Average Sentence Length, S/Cd = Simple / Compound, Cx/CdCx = Complex / Compound-complex

WordSmith Tools 4.0 (WST)

WST is used in this study to analyze reading materials at word level. The WordList tool of WST is utilized as the study requires comparison between corpora. The WordList tool provides useful statistics on average word length, which are used to explain materials difficulty at word level.

The study also utilizes the Detailed Consistency Analysis function, which is one of the WordList tool sub-functions, to compare two or more word lists created. This function is used to compare reading materials with the first 2000 high frequency words in the BNC World (2000). The following formula is used to calculate the coverage of the high-frequency words in the materials.

\[
HFW = \frac{\text{Total number of words that is within the high-frequency list}}{x} \times 100
\]

Figure 1 summarizes the computational tools used in the study and the types of data obtained at text, sentence and word levels.
Reliability of Instruments

The FRE formula is one of the most tested and reliable formulae to measure readability of materials (Chall, 1958; Klare, 1969). Its validity in estimating readability of materials in an ESL/EFL context has also been proven. Hamsik (1984), Greenfield (1999; 2004) and Shokrpour (2005) state that the formula is valid and can be used to determine readability level of English language materials in a foreign language context.

The use of WWB in text analysis is rather new. So far, only one study has validated the reliability of WWB distinguishing the different types of sentences used in reading materials. Aziz (2010) conducted an inter-coder reliability check to verify the reliability of WWB in distinguishing different types of sentences – simple, compound, complex and compound-complex sentences. Results of the test showed an average Kappa value of .793. Based on Landis and Koch (1977), this value is substantial in terms of inter-rater reliability. It shows that WWB is
reliable in distinguishing the different types of sentences used in reading materials.

On the other hand, reliability for WST in analyzing texts has been verified by numerous studies such as Nelson (2000), Bondi (2001), Henry and Roseberry (2001), Scott (2001), Flowerdew (2003), Mukundan (2004) and de Klerk (2004; 2005). Mukundan (2004) also concludes that WST is the most capable tool in providing instant basic information about words at sentence and paragraph levels as compared to a few other text analysis software.

**Conclusion**

Conventional readability formulas usually provide estimates of overall readability of reading materials. However, the composite computational tools proposed in this study, are able to provide more information about readability of reading materials at sentence and word levels. These tools enable estimation of materials difficulty to be performed objectively and reliably.

The use of WWB enables information on the average sentence length and the use of different types of sentences in the materials to be extracted. Meanwhile, the use of WST enables information on the average word length and the coverage of high-frequency words in the materials to be extracted. This additional information, together with the overall readability of the materials, gives a better estimation of the difficulty level of reading materials. Language instructors can use this information to match reading materials with the learners. Besides that, language
instructors can also adjust the readability of the materials by making changes related to the ASL, S/Cd, Cx/CdCx, AWL and HFW of the materials.

The composite computational tools are not just reliable but comprehensive as the tools analyze reading materials at three different levels: text, sentence and word levels. Therefore, language instructors should consider this alternative way to measure material difficulty when selecting reading materials for their learners in order to ensure a good match between reading materials and the target learners.
References


