Creating Consumer Friendly Health Content: Implementing and Testing a Readability Diagnosis and Enhancement Tool

Joshua Proulx  
University of Utah  
joshua.proulx@utah.edu

Sasikiran Kandula, MS  
University of Utah  
sasi.kandula@utah.edu

Brent Hill, RN, MS  
University of Utah  
brent.hill@hsc.utah.edu

Qing Zeng-Treitler, PhD  
University of Utah  
q.t.zeng@utah.edu

Abstract

In the era of patient centered care, creating consumer friendly health content is an important task. Manual content development is labor intensive and could benefit from a readability assessment and enhancement tool. Building on our prior work, we developed and evaluated such a tool called ReDE. In testing, a clinician was asked to use the tool to simplify ten full-length medical documents with 9573 words. In order to assess inter-rater agreement, a second clinician simplified four of the ten documents. The results show that 77% of the clinicians’ revisions were for concepts identified as difficult by ReDE, which validates the readability assessment made by the tool. However, a much smaller percentage (33%) of the ReDE suggested replacements were accepted by either of the clinicians which indicates further improvement is warranted.

1. Introduction

According to a report from the Office of Disease Prevention and Health Promotion “Only 12 percent of U.S. adults had proficient health literacy... Although half of adults without a high school education had below basic health literacy skills, even high school and college graduates can have limited health literacy.” [1] Even people with strong health literacy skills experience functional health deficits when, “They are not familiar with medical terms or how their bodies work. They have to interpret numbers or risks to make a health care decision. They are diagnosed with a serious illness and are scared or confused. They have complex conditions that require complicated self-care.” [1] In an effort to address this almost universal struggle to understand complex health information, researchers have been working to develop more ‘readable’ health material.

In the content development process, traditional formulae such as the Simplified Measure of Gobbledygook (SMOG) [2] and the Flesch-Kincaid Grade Level (FKGL) [4] are still commonly used. These instruments, however, are not well suited to measure the readability of health content as they overestimate the readability of certain health information such as clinical notes [5-8]. We developed a readability assessment and enhancement (ReDE) tool that is built on our prior work [9-13].

In this paper, we describe the design, implementation, and evaluation of the ReDE tool. We tested the tool with five clinical notes, two scientific journal articles, and three scientific study descriptions. The first clinician used the tool to simplify all ten documents while the second clinician used it on four of the documents. The majority of revisions the clinicians selected (77%) were difficult concepts identified by ReDE, which provides validation of the tool’s diagnostic capability. Conversely, only a small percentage of ReDE suggested replacements were accepted by the clinicians, which indicates a need for additional development of concept simplification alternatives.

2. Background

Traditional readability formulae such as the Simple Measure of Gobbledygook (SMOG) [2] and the Flesch-Kincaid Grade Level (FKGL) [4] are still in use, but are not well suited for health texts [5-8]. The FKGL and SMOG formulae are the most widely used of these older formulae. They use features of average sentence length and syllable count to assign a grade level to the text that corresponds to the minimum grade level of education that is required to comprehend the associated text. For instance SMOG is calculated as [10]:

\[ SMOG = 1.043 \times \sqrt{pw} + \frac{30}{sentences} + 3.1291 \]

where pw is the number of three or more syllable words. FKGL is similar and is defined as [10]:

The characteristics of health texts make sentence length and syllable count poor measures to rate readability. For example, clinical notes often have very short sentences and many short abbreviations. They are very difficult for even highly educated users to understand.

Recently, newer readability tools have been developed and tested [9, 14-16]. Coh-Metrix is a newer readability tool that provides measures of text cohesion and text difficulty by using computational linguistic techniques to extract information from text automatically [16]. Leroy et al. developed and evaluated metrics that indicate the difficulty level of health related documents with the goal of developing a software toolkit of metrics to provide laypersons and professionals with specific indicators of readability [17]. Gemoets et al. analyzed factors that predict the readability of consumer health texts and incorporated them into their Readability Analyzer tool [14]. Our team has also done several readability studies [7, 9-13].

We have also developed readability measure and simplification algorithms in several projects that have created a foundation for the ReDE readability tool. In 2007, Kim et al. reported a new health specific readability measure designed to assess and improve the readability of consumer health materials [9]. The measurement was evaluated with four types of health documents and compared to three existing readability formulae. It was found to provide a more accurate assessment of the readability of medical records than older tools. In 2010, Clauson et al. used the aforementioned measure to assess the readability of patient and health care professional targeted dietary supplement flyers [12]. In 2010, Kandula et al. described a semantic and syntactic algorithm for health content that substitutes difficult terms with simpler synonyms or related terms and splits long sentences into shorter grammatical sentences. We found that this algorithm led to a statistically significant improvement in comprehension [11]. A comprehensive explanation of the readability formula within ReDE is outside the scope of this paper but is covered in detail by Zeng-Treitler et al. [13].

The emerging body of science in readability of health texts by our team and our colleagues is making significant progress towards creating more comprehensible health information. However, none of these new measures/tools have been directly used in health content development. The advancement that this project contributes is the development of an end user application that not only assesses readability but also helps to diagnose specific readability issues and supports the enhancement of text through the integration of editing capabilities.

3. Methods

3.1. Design/Architecture

The readability diagnosis and enhancement tool

### Figure 1. Readability tool architecture
ReDE was built on our previously developed readability assessment and translation methods [9, 12]. Clinicians and health educators are the intended users. ReDE provides a graphical user interface that supports the analysis of free text documents. It assesses an overall readability score for a document using a number of unit length, lexical, vocabulary, style, and coherence features as well as scores for the individual features, which allows the difficult/easy aspect of each concept to be displayed. Difficult terms are also highlighted for users with recommended revisions, which are generated based on knowledge from the UMLS vocabulary [18] and the Consumer Health Vocabulary (CHV) [19] to generate the recommendation.

3.2. Implementation

The ReDE tool is a Java application with a Swing graphical user interface (GUI). The application runs within a Java Virtual Machine, version 1.7, and accesses UMLS vocabulary data [18] and the Consumer Health Vocabulary [19] stored in a local MySQL database.

The application utilizes the GATE natural language processing library [20] for a number of medical text annotation steps. During initialization of the tool, the readability model creates a Pipeline class to interface with the GATE library, version 7.0. It also generates a StatisticsGenerator class that is used after document parsing to calculate the readability statistics on the document, in addition to a StatisticsWriter class which will store the document readability statistics back into a local MySQL database.

Once the readability model has initialized, the application then generates a view used to display the tool’s readability GUI (see Figure 2). A ReadabilityDocPanel class used to display the text being simplified is created and positioned on the left half of the main application window. To the right a POSChartPanel class is positioned to display the part-of-speech tag frequency distribution for both the currently loaded document along with a difficult and easy corpus for comparison. Below the part-of-speech tag distribution a DocumentStatistics pane is initialized to display the length statistics for paragraphs and sentences, along with the syllables per word for both the current document as well as the lengths for a difficult and easy corpus. Below the length graphs displayed within the DocumentStatistics pane are two additional boxes: one containing the readability score percentages for the document including the context-based term score, frequency-based term score, CUI (concept unique identifier) score, adjusted score, and cohesion; and another containing the readability measures including the Flesch-Kincaid grade level [4], Flesch reading score [5], SMOG [2], GFI [22], Bormuth grade [23], Dale-Chall grade [24], Cole-Liau grade [25], and the readability distance score [9]. Some document controls are displayed below the

Figure 2. Readability diagnosis and enhancement tool GUI
ReadabilityDocPanel in the bottom left to allow the user to set the readability score threshold that determines the cutoff score. To the right of the threshold slider are three checkboxes to control whether or not to highlight the easy, difficult, and unscored medical concepts. The easy terms consist of all identified medical concepts with a readability score above the currently selected threshold and are highlighted within the document panel in green. The difficult terms with readability scores lower than or equal to the currently selected threshold are highlighted in red while the medical concepts that were identified but did not have readability scores are highlighted in blue. The readability score used for each identified medical concept is controlled by the “Score to display” radio-based boxes allowing the user to select either the context-based term score, frequency-based term score, CUI score, or adjusted score. To the right of that box is a control to allow the user to specify which character is used within the document to delimit paragraphs which allows the tool to handle documents that may have been created in another operating system.

Once the model and view have been initialized, the application waits for callback functions tied to various event listeners on the GUI to run once the user clicks on a control or menu item. When the user chooses to open a document a standard file selector is displayed and the user is allowed to specify which text file to load within the application. Once a valid text file is selected the application begins to load and parse the document.

The tool loads documents with UTF-8 character encoding and internally converts the document character set to ISO-10646, more commonly known as Unicode. After loading the document, the Pipeline class calls the GATE library [20] to execute the natural language processing components on the document in the following order: Text Tokenizer, Sentence Splitter, Part-of-Speech Tagger, Noun Phrase Splitter, UMLS Concept Finder, and Text Statistics Generator. The first four components executed are part of the HiTEx library [26]. The Text Tokenizer identifies the whitespace characters within the document that separate words, sentences, and paragraphs. The Sentence Splitter component identifies which AnnotationSet objects are grouped together within the same sentence. The Part-of-Speech Tagger is based on the Hepple tagger [27], a modified version of the Brill tagger [28], which classifies each word looking for adjectives, nouns, proper nouns, plural nouns, adverbs, and verbs. The Noun Phrase Splitter is based on the transformation noun phrase chunking by Lance Ramshaw and Mitchell Marcus [29] which takes the previously classified nouns and looks at the adjacent words to identify the complete noun phrase boundaries. With the main text annotation complete, the UMLS Concept Finder class connects to a local MySQL database containing a 2004AB release of the UMLS vocabulary [18] and iterates over the annotated noun phrases identifying UMLS concepts within the annotation. It adds mapped CUI annotations to the original document annotation and adds all semantic types that the CUI is mapped to within the database. The Text Statistics Generator component is then executed to calculate the readability scores which are then displayed in the bottom right of the application’s GUI.

Once the natural language processing and annotation steps executed by the GATE library [20] are complete, the tool then iterates over each identified medical concept creating a ConceptTO object for each concept which stores the concept’s CUI, text representation, and start and stop locations within the document. For each concept a readability score is obtained by querying the Consumer Health Vocabulary [19] stored locally in a MySQL database. This score is used by the application view when determining which words and phrases to highlight green, red, and blue to identify the easy, difficult, and unscored concepts.

With the document loaded and parsed, the view then responds to right-click actions that the user executes on highlighted text within the JTextPane. For each right-click event the view will then pass the cursor location into the readability model to identify which medical concept, if any, the user has clicked and all replacement phrases available that may be easier for a reader to understand. The list of replacement words or phrases is generated by identifying synonyms or explanations from three sources: preferred names for the CUI within the Consumer Health Vocabulary [19], explanatory phrases based on parent-child relationships within the UMLS (i.e. “CHILD CONCEPT (a type of PARENT),” or “CHILD CONCEPT (e.g. PARENT CONCEPT)”); or explanatory phrases based on relationships from this concept to related terms (i.e. “CONCEPT (a condition affecting RELATED CONCEPT),” “CONCEPT (a procedure performed on RELATED CONCEPT),” “CONCEPT (a part of RELATED CONCEPT),” “CONCEPT (can have a trade name of RELATED CONCEPT),” or “CONCEPT (a device/instrument used in a procedure called RELATED CONCEPT)”.

For each explanatory phrase the text for the parent concept or related concept is determined by checking the Consumer Health Vocabulary [19] for a preferred text for that concept before defaulting to the text found within the UMLS vocabulary [18].

Once the replacement strings are identified the view generates a floating popup menu that lists all UMLS concepts identified for the concept in that
location followed by a “Replace with” submenu that contains zero or more replacements strings along with the readability score, if known, for the replacement string. Figure 3 is an example of a user right-clicking on the term “exertion” and being presented with two potential replacement strings “effort” and “exertion (e.g., exercise).”

Figure 3. Sample replacement strings

3.3. Testing

To assess the potential of using the ReDE tool in health content development, a convenience sample of ten full-length medical texts (clinical notes, scientific journal articles, and clinical trial descriptions) were obtained from the Internet from PubMed, ClinicalTrials.gov, and other online sources. Clinical notes, journal articles, and clinical trial summaries are generally recognized as difficult for patients to understand as they are primarily written for health professionals as the target audience. While these difficult documents may not be intended for patient consumption and provide a real challenge to simplify, these documents become more relevant as patients start to access these resources to learn more about a recent diagnosis they have been given or research alternative treatment options that may be available for their condition.

One clinician (A) was asked to use the ReDE tool to revise the ten professional documents for patients’ consumption. To estimate inter-rater agreement, a second clinician (B) was recruited to perform the same task on four of the ten documents. While the ReDE tool highlighted the difficult terms and provided suggestions, the clinicians were free to change the threshold for difficult terms, ignore the suggestions, or revise the text manually. The edits made by the clinicians were recorded.

To assess ReDE’s ability to diagnose term-related readability issues, we calculated the percentage of difficult concepts identified by ReDE that were revised by the clinicians and the percentage of clinician revised terms that were identified by ReDE as difficult (using 0.7 as the threshold for difficulty). To assess ReDE’s ability to enhance readability, we measured the percentage of concepts with ReDE suggested replacements that were replaced by the clinicians with one of the suggested replacements.

4. Results

The number of edits made by the clinicians were analyzed. A majority of all edits (78% by clinician A and 68% by clinician B) were made to medical concepts that were marked as difficult by ReDE (see Figures 4 and 5 and Table 1). The rest of the edits were made to easy concepts (10% by clinician A and 15% by clinician B) or unscored concepts (12% by clinician A and 17% by clinician B) as some of these concepts are in fact difficult for non-clinicians to understand. A majority of the concepts (61%) recognized by ReDE as difficult were revised by the clinicians as they simplified the documents.

The edits made by clinician B to the four documents in common showed a large overlap between the medical concepts edited by both clinicians in order

Table 1. Results

<table>
<thead>
<tr>
<th></th>
<th>Clinician A</th>
<th>Clinician B</th>
<th>Combined</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Clinical</td>
<td>Scientific</td>
<td>Clinical</td>
</tr>
<tr>
<td>Total Words Reviewed</td>
<td>1143</td>
<td>8430</td>
<td>450</td>
</tr>
<tr>
<td>Total Edits</td>
<td>160</td>
<td>1059</td>
<td>48</td>
</tr>
<tr>
<td>Edits to Difficult Concepts</td>
<td>131</td>
<td>818</td>
<td>26</td>
</tr>
<tr>
<td>Suggestions Used</td>
<td>31</td>
<td>205</td>
<td>6</td>
</tr>
<tr>
<td>Difficult Suggestions Used</td>
<td>30</td>
<td>170</td>
<td>6</td>
</tr>
</tbody>
</table>
Table 2. Example suggested replacements and edits by clinicians

<table>
<thead>
<tr>
<th>Original Text</th>
<th>Suggested Replacements</th>
<th>Clinician A</th>
<th>Clinician B</th>
</tr>
</thead>
</table>
| He has had **persistent symptoms**, although he describes some improvement on atenolol. He denies **radiation of the pain**, nausea, vomiting, diaphoresis, palpitations, syncope, or near syncope. | **atenolol**: 1) atenolol (a type of cardiac drugs)  
**radiation of the pain**: 1) radiation of the pain (a type of pain)  
**diaphoresis**: 1) excessive sweating 2) diaphoresis (a type of sweating affecting skin)  
**palpitations**: 1) palpitations (e.g. heart pounding affecting heart)  
**syncope**: 1) syncope (e.g. black outs) 2) fainting                                                                 | He has had **ongoing symptoms**, although he describes some improvement on atenolol. He denies pain spreading from the left chest, nausea, vomiting, excessive sweating, heart pounding, fainting, or near fainting. | He continues to experience symptoms, although he describes some improvement from taking atenolol. He says he has not experienced any spreading of the pain, nausea, vomiting, sweating, unusual heartbeats, fainting, or almost fainting. |

Concept Readability Coloring  
**Green** = Easy Concept  
**Red** = Difficult Concept  
**Blue** = Unscored Concept  

Replacement/Edit Coloring  
**Yellow** = Suggested Replacement, **Brown** = Manual Edit
5. Discussion

Despite advances in the development of new readability assessment tools in the health domain there are few automated diagnostic and enhancement tools for health content. Content developers typically rely on their own judgment and experience. As traditional readability formulae are still being used, a writing style (typically using short sentences and shorter words) is commonly used that reduces “grade level” but does not necessarily make the content more consumer friendly. Our tool takes factors other than word and sentence length into consideration when generating the overall readability score. It also highlights difficult terms and provides suggestions.

The fact that a majority of edits made by clinicians are also identified by ReDE as difficult suggests that ReDE has the potential for use as a diagnostic tool. On the other hand, the clinicians did not accept a majority of the suggested replacements.

Worth noting is that the two clinicians did not always agree with each other on which term to replace and how to replace them. This is not a big surprise since content development is still in some ways an “art.” It is also not a surprise that clinicians often rejected the machine-generated suggestions. Even after decades of research, machine translation cannot compete with human-translation. Nevertheless, since human labor is very expensive, even a small amount of automation is desirable. While it may be many more years before machine language tools like ReDE reach the level of accuracy that would be needed to automatically process the medical documents written and consumed across clinicians and researchers, one of the focuses in the near term for ReDE will be on the documents that are intended for patient consumption like discharge instructions given to patients leaving a clinical setting and consumer health documents available online. Due to the limited sample size of documents simplified by the clinicians a comparison of the difference between ReDE’s suggested replacements by document type was not done but is being considered in the future.

One of the limitations of the existing software is the ability to recognize when a particular generated replacement suggestion does not make logical sense or does not provide enough of an explanation. One example includes the word “his” which was correctly scored as an easy term but internally had a generated suggested replacement of “his (a type of amino acids)” which is a technically correct suggested replacement for “his” being used as a noun as an abbreviation for the histidine amino acid but was not a valid suggested replacement given the word “his” being used as a possessive adjective in the sentence. The tool likewise has difficulty recognizing redundant suggested replacements that do not provide much additional explanation. When encountering the concept “hematocrit” the tool correctly identifies the term as a difficult concept and offers a reasonable replacement of “hematocrit (a type of red blood cell test)” that provides additional insight into the term. However, after identifying the difficult concept “red blood cells” a suggested replacement of “red blood cells (a type of cell)” is offered which does not provide additional benefit to a reader given that the clarifying text “cell” was already found in the original text. This type of explanation is generated because the “red blood cell” concept exists within the UMLS vocabulary as a child of the concept “cell” which is correct but not very useful in simplifying the term. Future improvements are planned to the tool to eliminate suggested replacements that contain redundancy in the explanation such as the red blood cell example.

Additional work is planned to incorporate the Google Corpus as an supplementary reference to use when calculating the readability score for concepts that were identified but did not have sufficient data within the UMLS vocabulary to calculate a readability score. With this additional data point, the frequency for which the term was found on the Internet as recorded within the Google Corpus will be used to help to determine a readability score for the term. This will reduce the number of unscored concepts highlighted after the application processes a document.

During the evaluation of the tool, feedback from the clinicians included the need for Search &
Replace/Replace-All feature. These enhancements are planned for a future iteration along with the ability to store manual edits as potential suggested replacements for the suggestion engine.

6. Conclusion

The evaluation of ReDE demonstrates that the application effectively identifies the difficult medical concepts within medical documents. Improvements are needed to create better replacement suggestions that users will select more often and to capture user-entered revisions to increase the tool’s knowledge.

7. Acknowledgements

This work is funded by grants from the National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK075837) and the National Institute of Health (R01 LM07222).

8. References


