User Evaluation of the Effects of a Text Simplification Algorithm Using Term Familiarity on Perception, Understanding, Learning, and Information Retention

Gondy Leroy1,2*, MS, PhD; James E Endicott1*, David Kauchak3*, PhD; Obay Mouradi1*; Melissa Just4*  

1Information Systems and Technology, Claremont Graduate University, Claremont, CA, United States  
2Eller College of Management, Department of Management Information System, University of Arizona, Tucson, AZ, United States  
3Computer Science Department, Middlebury College, Middlebury, VT, United States  
4Rutgers University Libraries, Rutgers, The State University of New Jersey, New Brunswick, NJ, United States  
*all authors contributed equally

Abstract

Background: Adequate health literacy is important for people to maintain good health and manage diseases and injuries. Educational text, either retrieved from the Internet or provided by a doctor’s office, is a popular method to communicate health-related information. Unfortunately, it is difficult to write text that is easy to understand, and existing approaches, mostly the application of readability formulas, have not convincingly been shown to reduce the difficulty of text.

Objective: To develop an evidence-based writer support tool to improve perceived and actual text difficulty. To this end, we are developing and testing algorithms that automatically identify difficult sections in text and provide appropriate, easier alternatives; algorithms that effectively reduce text difficulty will be included in the support tool. This work describes the user evaluation with an independent writer of an automated simplification algorithm using term familiarity.

Methods: Term familiarity indicates how easy words are for readers and is estimated using term frequencies in the Google Web Corpus. Unfamiliar words are algorithmically identified and tagged for potential replacement. Easier alternatives consisting of synonyms, hypernyms, definitions, and semantic types are extracted from WordNet, the Unified Medical Language System (UMLS), and Wiktionary and ranked for a writer to choose from to simplify the text. We conducted a controlled user study with a representative writer who used our simplification algorithm to simplify texts. We tested the impact with representative consumers. The key independent variable of our study is lexical simplification, and we measured its effect on both perceived and actual text difficulty. Participants were recruited from Amazon’s Mechanical Turk website. Perceived difficulty was measured with 1 metric, a 5-point Likert scale, and actual difficulty was measured with 3 metrics: 5 multiple-choice questions alongside each text to measure understanding, 7 multiple-choice questions without the text for learning, and 2 free recall questions for information retention.

Results: Ninety-nine participants completed the study. We found strong beneficial effects on both perceived and actual difficulty. After simplification, the text was perceived as simpler ($P<.001$) with simplified text scoring 2.3 and original text 3.2 on the 5-point Likert scale (score 1: easiest). It also led to better understanding of the text ($P<.001$) with 11% more correct answers with simplified text (63% correct) compared to the original (52% correct). There was more learning with 18% more correct answers after reading simplified text compared to 9% more correct answers after reading the original text ($P=.003$). There was no significant effect on free recall.

Conclusions: Term familiarity is a valuable feature in simplifying text. Although the topic of the text influences the effect size, the results were convincing and consistent.
Introduction

Background and Significance

Text is an important source for health-related information. It is easy to create, maintain, and distribute, and medical practitioners often use it to provide instructions and details on treatments. Health-related text is becoming increasingly available with an estimated 80% of online users [1] from a wide array of backgrounds [2] using the Internet to obtain health-related information. The information itself is diverse and includes prevention, treatment, and management of diseases and comes from a variety of sources ranging from professionals to salespeople to patients.

Unfortunately, 90 million Americans have difficulty understanding and acting upon health information [3], and many find the text currently available difficult to read [4]. Some of this difficulty can be attributed to inherent complexity in understanding the diseases, their causes, and the associated treatments, which may require advanced knowledge of biology, chemistry, or physiology to understand in detail. Much of the difficulty, though, can be attributed to a mismatch between the content delivered and the consumers who often have limited health literacy, low general education, or inadequate language skills. Low health literacy reduces health statuses of individuals [3], is considered a “silent killer” [5], and is estimated to cost up to US$238 billion annually [6].

To increase health literacy, the method, medium, and language used play an important role. While one-on-one teaching may be the best solution, medical professionals do not have sufficient time or resources for this. Video and interactive methods can be very educative and are becoming increasingly available. The power of such methods to teach and demonstrate will likely play an important role in consumer health information. However, currently text remains the primary tool used to educate people.

Factors Influencing Text Difficulty and Its Measurement

Figure 1 provides an overview of three key factors representing the authors’ view on influences on understanding and learning from text: personal characteristics, text characteristics, and measurement characteristics. Personal characteristics describe attributes about the reader. Some are innate and cannot be changed, for example, native language and general intelligence. Others are acquired, for example, vocabulary size and domain knowledge. Many of these characteristics have a direct effect on text comprehension and indirectly on learning since comprehension has been shown to affect learning [7]. For example, stress, a personal characteristic, has been shown to affect reading behaviors. People with high stress rely more on visual summaries, even when incomplete, to answer text-based questions [8]. Moreover, increased stress has also been related to lower comprehension of medical terminology [9]. Other personal characteristics, such as the ability to form a good mental model, affects understanding since readers often rely on the mental model instead of the original text base [10]. In addition, past behaviors and acquired skills can have an impact. Exposure to print, for example, has been found to be related to understanding. Landi [11] found a positive relation with results for an author recognition test [12] and question-answering tasks, while in our own work, we found a positive relation between self-reported reading and results for a fill-in-the-blank Cloze test [13].

Text characteristics influence text difficulty and therefore understanding. These characteristics can be adjusted to improve the usefulness of text, but this has been shown to be challenging and very few studies have shown strong improvements in reader understanding. To further clarify the analysis of text characteristics and the text simplification problem in general, we distinguish between the perceived and actual text difficulty of a text. The distinction is based on evidence for the existence of perceived barriers from the Health Belief Model [14] and the importance of perceived difficulty of behavioral control from the Theory of Planned Behavior [15]. While actual difficulty is easily accepted as important, perceived difficulty cannot be ignored. At a minimum, it will impact whether or not a text will be read. However, it may affect health literacy in more ways; for example, Velayo [16] found that a higher perceived difficulty correlated with a decrease in the recall of information.

Text characteristics can include surface features, for example, spacing or font, and range from smaller units such as words, to larger units such as sentences or paragraphs. Using a Likert scale to measure perceived difficulty, it was found that texts with a higher ratio of function words, verbs, verb phrases, or containing more high-frequency words were seen as easier [17-19]. For actual difficulty, simple surface features such as font and line spacing were shown not to affect remembering [20]; however, using a fill-in-the-blanks test additive and causal connectors were shown to be easier than adversative or sequential connectors [21]. In addition to surface features, analysis can include broader features such as coherence, which is defined as good flow combined with a structured, logical argument [22,23]. We found that increasing coherence with proper spacing around subtopics and better logical connectors improved question-answering (actual difficulty) [13]. Not surprisingly, how a topic is presented in a text also influences learning; topics introduced as part of a refutation text, a text where misconceptions are explicitly addressed, led to increased learning and more valid inference but not increased quantity of information being recalled [7].

Measurement characteristics also play an important role in readability research, although they are often ignored. Historically, the most popular measurement has been readability formulas, which generate a single number often based only on relative word and sentence length and are used as stand-ins for text complexity [24]. These formulas have become popular even though they ignore current knowledge about the reading process,
have a shaky statistical basis, and are unhelpful as writing guidelines [25]. The Flesch-Kincaid Grade Level formula is the most common in health care literature [26]. Even though different tools using the formula sometimes return different levels for the same text [27], it has been used to evaluate patient education materials [28], general websites [29], and information on specific topics such as abdominal aortic aneurysms [30] and back pain [31]. Other readability formulas, such as the Simple Measure of Gobbledygook (SMOG) and Gunning Fog Index, have also been shown to be problematic for evaluating health-related materials for similar reasons [32]. Simplifying text based on these formulas sometimes results in more difficult text, that is, the simplicity paradox [5], because the simplification concentrates on writing style rather than content [2]. As a result, increasingly more concerns are raised about the effectiveness of these formulas for simplifying consumer health texts [33].

Better measures should be developed and used to evaluate text and motivate algorithmic components. These must be evaluated on a representative sample and measure not just the perception of difficulty, but more importantly understanding and retention of information. By using different measures, we can better evaluate the impact of simplification tools. For example, question-answering tasks (eg, multiple-choice, open-ended, or free recall questions), fill-in-the-blanks tasks (eg, multiple-choice or open blanks) and teach-back methods (eg, explain a concept or summarize a topic) can be used to measure understanding. Measures that test retention of information can follow the same style of questions, while measures of learning from a text require a comparison between pre- and post-reading scores.

Figure 1. Factors that influence understanding and retention of information.

Interactive presentations can also exist between personal, text, and even measurement characteristics. For example, the impact of text coherence on the reader has been found to interact with user characteristics and with the type of measurement. Overall coherence did not affect recall (actual difficulty) but affected remembering and understanding when measured by question-answering (actual difficulty) for readers with high knowledge but low interest or low knowledge but high interest in a topic [10,22]. Personal interest in the topic has also repeatedly been shown to be relevant. A higher interest leads to increased learning [34] and recall [35], however, the coherence of text [34] and prior knowledge [35] influence this relationship.

Objective

Our objective is twofold. First, we address the need for an evidence-based algorithm that pinpoints difficult text. Second, we focus on providing appropriate, easier alternatives to a writer in an effective and efficient manner. We present here our first fully automated version of the lexical simplification algorithm, which identifies difficult terms and generates a list of easier alternatives based on information extracted from dictionaries and other databases. In a pilot study [36], we introduced the text simplification algorithm and presented an initial user study. This work builds upon the lessons learned in the pilot study and differs in a number of key dimensions: (1) the algorithm examined here is fully automated, (2) the simplification of text is done by an independent writer, not the developers, and (3) the evaluation is based on a new study with different participants, new stimuli, and new more comprehensive metrics.
Methods

Text Simplification Algorithm and Writing Process

The automated algorithm executes two steps. The first step is identification of difficult terms. We conducted corpus analyses and found that the term familiarity differed between easy and difficult texts [17,18]. Motivated by this, our algorithm uses the Google Web Corpus [37], which contains n-gram counts from a corpus of 1 trillion words from public webpages to identify difficult terms. Terms with a low frequency in this corpus are assumed to be less familiar and therefore more difficult since a reader would not encounter them often. We used unigrams and the 5000th most frequent word, which has a frequency of 15,377,914, as our threshold for distinguishing less familiar terms. Any term with a lower frequency is considered difficult and is a candidate for replacement.

We used the Google Web Corpus because its terms are representative of everyday readers without special medical knowledge. Other resources may provide additional value but may also introduce inconsistencies. For example, the Google Book Corpus contains many medical books resulting in higher frequencies for medical terms. The Unified Medical Language System (UMLS) contains both medical and general terms. Distinguishing between them algorithmically would be necessary, which is not an easy task, and may not improve upon the frequency-based approach by much.

The second step is the identification and presentation of easier alternatives for each difficult term. The list of candidate replacements is generated from synonyms and hypernyms from WordNet 2.0 [38,39]; definitions and semantic types from the UMLS; and definitions from both the English and Simple English Wiktionaries. Only alternatives that possess the same part of speech based on an automatic tagger are presented. In addition, only substitutions with a higher term frequency than the original word are suggested (ie, more familiar). The number of alternatives provided can be adjusted based on user preference or application; currently, we aim to provide a minimum of 7 alternatives. Candidate replacements are sorted both by source (for the convenience of the writer) and by their familiarity in the Google Web Corpus.

In contrast to the previous version of our simplification algorithm [36], which involved one of the authors manually looking up each word to generate the candidate suggestions, the current version is fully automated. To ensure that the algorithm is sufficiently efficient for later inclusion in a comprehensive tool, we tested its efficiency on Wikipedia articles. We selected 100 conditions randomly (see Multimedia Appendix 1) from a list of diseases provided by the Mayo Clinic. For each disease, we retrieved the corresponding Wikipedia article. The articles were on average 2573 words long. On average, 617 words were tagged as difficult per article, for which easier alternatives were produced by the algorithm where available. The average run time was 37 seconds per document.

Given the difficulty of completely automated translation, especially in domains such as health where information may not be omitted, we require a writer to finalize the text. At present, a Microsoft Excel spreadsheet is generated containing each original sentence from a text, the same sentence with blanks for all difficult words, and alternatives for each difficult word. The alternatives are presented in a column and ordered according to source and term familiarity. The writer chooses the best alternative, replacing it in the original text. Ensuring grammatical correctness (eg, consistent pluralization) is currently the responsibility of the writer.

Original and Simplified Texts (Study Stimuli)

A subject expert (SE), a medical librarian, simplified the texts. To optimize external validity, we worked with one expert to rewrite the text since this is how the final tool will be used. To increase internal validity, we provided the SE with rules to ensure that we measured only the effects resulting from interaction with our algorithm. She was asked to “Try to replace as many words as possible” and when making a replacement “single words can just be replaced but longer fragments should be added before or after the sentence (with some adjustment for flow of text)”.

The SE served two main roles: (1) to determine if a difficult word flagged by the algorithm needs to be replaced, and (2) for those words requiring replacement, to select an appropriate substitution from the alternatives suggested by the algorithm. If the SE deemed that an appropriate synonym existed for a difficult word in the algorithmically generated options, then the difficult word was simply replaced by the synonym. If the simplification option selected by the SE was not a synonym, it needed to be added to the text so that no original information was deleted from the text. Simplifications containing longer phrases or sentences (eg, from definitions) were added by using parentheses or by adding a separate sentence before or after the target sentence. The text was adjusted by the SE as necessary to create grammatically correct sentences.

In previous work [36], we noticed that lexical simplifications by the authors reduced the flow of the text thereby increasing text difficulty. Therefore, the SE was asked to pay close attention to how alternatives were inserted and to choose the option that resulted in the best flow. If the SE preferred a term other than those suggested by the algorithm, she could add it to the text for familiarity verification. Once the text was rewritten, it was rerun through the simplification algorithm to ensure that newly added text was sufficiently simple. This included the verification of any synonyms by the SE.

To measure perceived difficulty, we selected 5 text snippets; these were individual sentences and in one case 2 short sentences combined. Such short snippets do not require much time to read, provide more data points than one long text, and ensure that study participants do not get overwhelmed. The sentences were taken from English Wikipedia articles, and each sentence was simplified by the SE using our algorithm. Our algorithm tagged an average of 11 words per sentence as difficult, of which 5.6 (53%) were replaced.

To measure actual difficulty, it was necessary to use longer texts to allow for questions about the content to be posed. We used two different texts so that each participant in the study worked with an original and simplified text for better (statistical) control of interpersonal differences. We chose a text on liver cirrhosis and one on asthma because most people are somewhat familiar
with them and both conditions have several commonly accepted myths associated with them. These myths were incorporated into our multiple-choice questions and provided an excellent opportunity to demonstrate learning. Each text was simplified using our approach described above. Texts were obtained from the initial summary paragraphs described from their Wikipedia Web pages and were similar in composition. Our algorithm tagged 210 words as difficult in the liver cirrhosis document, of which 66 (31%) were replaced by the writer during simplification. In the asthma document, 122 words were tagged as difficult and 53 (43%) were replaced during simplification.

Tables 1 and 2 show an overview of the text characteristics before and after simplification. We include the Flesch-Kincaid Grade Level for comparison with other work. Below are examples of an original and simplified snippet used as part of the study (perceived difficulty):

- original: “Gout is a disorder of purine metabolism, and occurs when its final metabolite, uric acid, crystallizes in the form of monosodium urate, precipitating in joints, on tendons, and in the surrounding tissues.”
- simplified: “Gout is a disease of the processing of the chemical substance called purine, and occurs when its last chemical product (uric acid) makes crystals (monosodium urate), which collect in joints, on tendons, and in the surrounding tissues.”

The texts, both original and simplified versions, are provided in Multimedia Appendix 2.

### Table 1. Text snippet characteristics.

<table>
<thead>
<tr>
<th>Sentences (N=5)</th>
<th>Lexical simplification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
</tr>
<tr>
<td>Word count (avg)</td>
<td>28.4</td>
</tr>
<tr>
<td>Flesch-Kincaid grade level (avg)</td>
<td>18.6</td>
</tr>
</tbody>
</table>

### Table 2. Document characteristics.

<table>
<thead>
<tr>
<th>Documents</th>
<th>Lexical simplification</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Original</td>
</tr>
<tr>
<td>Topic</td>
<td>Asthma</td>
</tr>
<tr>
<td>Word count</td>
<td>623</td>
</tr>
<tr>
<td>Sentence count</td>
<td>31</td>
</tr>
<tr>
<td>Flesch-Kincaid grade level</td>
<td>13.9</td>
</tr>
</tbody>
</table>

### Metrics

To measure perceived difficulty, participants judged a sentence using a 5-point Likert scale with the following labels: Very Easy, Easy, Neither, Hard, Very Hard. Perceived difficulty is the score on this scale with 1 representing Very Easy and 5 Very Hard.

To measure actual difficulty, we used metrics covering understanding, learning, and retention of information. For understanding of the text, we used 5 multiple-choice questions posed alongside the text. The questions targeted different sections of the text. Understanding was measured as the percentage of questions answered correctly.

To measure learning, we compared scores on 7 multiple-choice questions shown both before and after reading the text. The text itself was not visible when the questions were presented. By asking the same questions before and after, we were able to use participants as their own controls. For each text, we created the multiple-choice questions based on commonly accepted myths. The myths were gathered by searching the Internet for “common myths about…”.” Learning was measured as the increase in the percentage of questions answered correctly after versus before reading the text.

To measure retention, we asked participants after all sections have been completed to list all facts (one per line) that they remembered from the texts. Retention can be simply measured as the number of facts listed, however, since these facts may contain errors, they were also graded by the authors. Two authors per topic independently graded all facts. Even though participants were asked to list 1 fact per line, many lines included multiple facts per line. Each fact was considered and awarded points separately: +1 for a correct fact and -1 for an incorrect. To grade the answers in an objective manner, the order of answers was randomized per grader and the experimental condition unknown. In cases with a large disparity between grades (scores diverged by more than 100%), a third grader (the SE) judged the results and provided the final score (similar to original manual GRE scoring [40]). Retention was then measured with 2 metrics: the number of listed facts and the sum of the grades assigned to those facts.

In addition to study questions, we also included qualifying questions. These were simple questions for which the answer was obvious. They helped filter results of participants who were not serious about the study. We included a qualifying question with each set of multiple-choice questions and filtered any participant who did not get all qualifying questions right.
Participants

Participants were recruited using Amazon’s Mechanical Turk. MTurk is an online crowdsourcing service that allows for small tasks to be accomplished by human workers. Currently, Amazon has over 300,000 requested tasks and over half a million workers. Workers are paid a small sum for each task accomplished. MTurk has been used in a wide range of settings ranging from user studies to data annotation to subjective rating generation [41]. The workers are a diverse group from all over the world with varied demographic characteristics [42,43]. When precautions are taken to filter out ineffective workers, the quality of the data obtained has been shown to be at least as good as data obtained from more traditional approaches [43,44].

Procedure

Participants were directed to our study website from MTurk, and the sections were presented in the following order:

- The first page contained the welcome note and instructions to complete the study sections in order and without use of external sources. From this point, the browser back button was disabled.
- The first study section showed the myth-based questions for a topic. Then, the text was shown together with new questions, followed by a repetition of the myth-based questions without the text. For each participant, the order of the questions and answers for each question were randomized. The topic was either liver cirrhosis or asthma, and the version was either original or simplified.
- The second study section was identical to the first, but with a different text in a different version. Each participant received one original and one simplified version. The order and topics were balanced over the study so that all combinations of topic and difficulty level were presented.
- The third study section contained the individual sentences that participants judged for perceived difficulty. The original and simplified version of a sentence were paired because showing all sentences in one list made it very difficult for participants to notice differences and provide a rational judgment. The order within each pair and the order of the 5 pairs were randomized per participant.
- The fourth study section contained demographic questions.
- The fifth and sixth study sections contained the PSS-10 [45], a standardized stress survey, and the STOFHLA [46,47], a standardized health literacy measure.
- The seventh and eighth sections contained the request for free recall of information for the first and second text.
- The final page showed a Thank You note and the code to be submitted for payment at MTurk.

Results

Participant Characteristics

We invited MTurk workers located in the United States with a 95% approval rate on tasks previously performed for other requesters. They were paid US$1.50 for completing the survey. Upon start, 134 participants signed up and 105 completed the study. Of those who completed, 6 did not pass our filtering criteria resulting in a total of 99 valid participants. Completing the survey took on average 33 minutes. The shortest time spent was 13 minutes and the longest was 45 minutes.

Table 3 provides the demographic information. Most participants (80%) were between 21 and 50 years old, with only a small group younger than 20 (3%) or older than 60 years (4%). The majority were female (63%), white (89%), and not Hispanic or Latino (93%). Most had moderate education: 48% had a high school diploma, 16% an associate’s degree, and 25% a bachelor’s degree. The majority (89%) spoke exclusively English at home.

Perceived Difficulty

We found a significant beneficial effect of simplification on perceived difficulty with simplified sentences being judged as simpler. Figure 2 shows an overview of the average score and standard error bars for each sentence and for all sentences combined. A paired-samples t-test showed the difference to be significant for all pairs ($P<.001$) and for all pairs combined ($P<.001$).
Table 3. Participant demographic information (n=99).

<table>
<thead>
<tr>
<th>Characteristics</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td></td>
</tr>
<tr>
<td>20 or younger</td>
<td>3</td>
</tr>
<tr>
<td>21-30</td>
<td>35</td>
</tr>
<tr>
<td>31-40</td>
<td>24</td>
</tr>
<tr>
<td>41-50</td>
<td>21</td>
</tr>
<tr>
<td>51-60</td>
<td>12</td>
</tr>
<tr>
<td>61-70</td>
<td>4</td>
</tr>
<tr>
<td>71 or older</td>
<td>-</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
</tr>
<tr>
<td>Female</td>
<td>62</td>
</tr>
<tr>
<td>Male</td>
<td>37</td>
</tr>
<tr>
<td>Race (multiple choices allowed)</td>
<td></td>
</tr>
<tr>
<td>American Indian / Native Alaskan</td>
<td>2</td>
</tr>
<tr>
<td>Asian</td>
<td>7</td>
</tr>
<tr>
<td>Black or African American</td>
<td>5</td>
</tr>
<tr>
<td>Native Hawaiian or Other Pacific Islander</td>
<td>-</td>
</tr>
<tr>
<td>White</td>
<td>88</td>
</tr>
<tr>
<td>Ethnicity</td>
<td></td>
</tr>
<tr>
<td>Hispanic or Latino</td>
<td>7</td>
</tr>
<tr>
<td>Not Hispanic or Latino</td>
<td>92</td>
</tr>
<tr>
<td>Education (highest completed)</td>
<td></td>
</tr>
<tr>
<td>Less than High School</td>
<td>1</td>
</tr>
<tr>
<td>High School Diploma</td>
<td>48</td>
</tr>
<tr>
<td>Associate’s Degree</td>
<td>16</td>
</tr>
<tr>
<td>Bachelor’s Degree</td>
<td>25</td>
</tr>
<tr>
<td>Master’s Degree</td>
<td>6</td>
</tr>
<tr>
<td>Doctorate</td>
<td>3</td>
</tr>
<tr>
<td>Language skills (frequency of speaking English at home)</td>
<td></td>
</tr>
<tr>
<td>Never English</td>
<td>-</td>
</tr>
<tr>
<td>Rarely English</td>
<td>1</td>
</tr>
<tr>
<td>Half English</td>
<td>3</td>
</tr>
<tr>
<td>Mostly English</td>
<td>6</td>
</tr>
<tr>
<td>Only English</td>
<td>89</td>
</tr>
</tbody>
</table>
Actual Difficulty: Understanding, Learning, and Retention

Figure 3 shows the mean scores and standard error bars for understanding. We conducted a two-way analysis of variance (ANOVA) with simplification and topic as independent variables. Topic was included to provide a more nuanced view. For understanding, we found two main effects. The first is for simplification with higher scores for simplified text. There were on average 52% correct answers with an original document and 63% with a simplified document, ($F_{1,198} = 13.869, P < .001$). There was also a main effect for topic ($F_{1,198} = 13.869, P < .001$) with higher scores achieved for the asthma document. Since the increases in understanding after simplification were comparable for both topics, the interaction effect was not significant.

Figure 4 shows the mean scores and standard error bars for the learning of information. We conducted a comparable two-way ANOVA with the simplification and topic as independent variables. We found a significant main effect of simplification of text with more learning from simplified documents (18%) than from the original documents (9%) ($F_{1,198} = 9.238, P = .003$). A second main effect was found for topic ($F_{1,198} = 22.301, P < .001$) with more learning with the liver cirrhosis document (20%) than with the asthma document (6%). The interaction between both independent variables was also significant ($F_{1,198} = 4.071, P = .045$) with the learning being more pronounced with the liver cirrhosis than with the asthma document.

Table 4 provides an overview of the retention of information using both raw and graded scores. With simplified documents, slightly more facts were listed (5.04) than with original documents (4.70). There were also slightly more words (43.60) and unique words (32.36) used after reading simplified documents compared to original documents (40.07 words and 30.79 unique words). These differences were not statistically significant. The graded scores show similar small differences. There were slightly more correct facts after reading simplified documents (5.04 facts) than after reading the original documents (4.70 facts). However, the difference is not statistically significant.
Figure 3. Average understanding scores.

Understanding

- Original
- Simplified

Percentage Correct

30 40 50 60 70 80

Asthma  Liver Cirrhosis  All

69 57 47

57 52

63

Figure 4. Average learning scores.

Learning

- Original
- Simplified

Increase in Percentage Correct

0 5 10 15 20 25 30

Asthma  Liver Cirrhosis  All

5 13

8 18

9
They also show the importance of using different metrics. We need to complete reading, among other factors. how perceived difficulty affects motivation to read and ability

perceived difficulty shows that or small sample sizes. Even so, this effect is important and

and strong effects are found even when using short text snippets

improve perceived difficulty is more straightforward. Consistent

of text.

Principal Findings

This work reported on a lexical simplification algorithm that automatically detects difficult terms and suggests easier alternatives. The writing process is semiautomated since the final replacements are made by the writer. A controlled user study showed how simplifying text in this manner led to significant improvements in both perceived and actual difficulty of text.

The results on perceived difficulty corroborate earlier work on manual lexical simplification. In general, changing the text to improve perceived difficulty is more straightforward. Consistent and strong effects are found even when using short text snippets or small sample sizes. Even so, this effect is important and shows that lexical simplification has a beneficial impact on perceived difficulty. Future studies will look more closely at how perceived difficulty affects motivation to read and ability to complete reading, among other factors.

The results on actual difficulty are strong and very encouraging. They also show the importance of using different metrics. We found a strong effect on understanding with simplified text being better understood. However, this effect also depended on the topic being studied. Learning showed a similar strong effect: there was more learning with simplified documents. These effects lead to our conclusion that lexical simplification is beneficial and has an immediate impact on understanding and learning. However, we did not find an effect of simplification on retention of information. This may be due to a lack of sustained learning or it may be due to the study design. In previous work on search engines [48], we found that many study participants stop finding information at some given point, regardless of how easy or difficult a task is. We may be witnessing a similar effect with participants submitting “enough” facts regardless of how many they remember. In future work, we aim to provide better incentives to encourage participants to submit more facts.

Limitations

There are several limitations we would like to point out. First, we evaluated our approach with short texts taken from Wikipedia. Different effects may be found for longer or more difficult texts. However, working with short texts allows for a controlled experiment, thereby avoiding potentially confounding variables. Future work will look for repeat effects in longer documents. Second, we worked with general topics. Automatically recognizing which different texts, either distinguished by difficulty level or other factors, would benefit from simplification would be an important addition to our work. In addition, working with personally relevant topics may increase effects, since motivation has been shown to be important to the reading process. Third, we worked with only one subject expert who rewrote text. Comparing different writers may show further strengths and weaknesses of our approach. Working with a team of writers may provide a more balanced gold standard; however, this approach has also been shown to introduce noise when experts disagree [49]. Further research is needed to understand the impact of each of these limitations.

Conclusions

In addition to these study limitations, there is also much room for future development of our algorithm. We aim to more precisely target difficult words so that fewer words are tagged for replacement while still impacting the overall difficulty of text. We aim to provide a shorter and more precise list of potential replacements by working with resources such as the Consumer Health Vocabulary [50-52]. This will make the process more efficient for the writer while requiring less time to generate alternatives. For example, we plan to test phrases in addition to individual words to estimate difficulty and work

Table 4. Retention of information: mean and standard deviation.

<table>
<thead>
<tr>
<th>Average counts</th>
<th>Original text</th>
<th>Simplified text</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Asthma</td>
<td>Liver cirrhosis</td>
</tr>
<tr>
<td>Facts</td>
<td>4.76</td>
<td>4.55</td>
</tr>
<tr>
<td>Words</td>
<td>40.50</td>
<td>39.63</td>
</tr>
<tr>
<td>Unique words</td>
<td>31.38</td>
<td>30.20</td>
</tr>
<tr>
<td>Average score (graded facts)</td>
<td>4.50</td>
<td>4.91</td>
</tr>
</tbody>
</table>

Relationships With Participant Characteristics

To complete our analysis, we conducted a correlation analysis using a 2-tailed Pearson product-moment correlation coefficient ($r$). We evaluated the personal characteristics and the scores for perceived and actual difficulty over experimental conditions. We assigned a code to the education level and language skills with a higher score indicating a higher level or skill. We also included the PSS scores and STOHLA scores.

Overall, there were few significant correlations. There were no significant correlations between the perceived difficulty of sentences and the personal characteristics. For actual difficulty, only education mattered. There was a positive correlation between education and understanding ($r=.244$, $P=.015$), facts listed ($r=.296$, $P=.003$), graded facts ($r=.411$, $P\leq.001$), and both the word count ($r=.316$, $P=.001$) and unique word count ($r=.329$, $P=.001$). Among the personal characteristics themselves, two correlations were significant. There was a negative correlation between language skills and stress levels, indicating higher stress related to lower language skills ($r=-.210$, $P=.037$) and also a negative correlation between language skills and education level ($r=-.260$, $P=.009$). Upon closer inspection, this last negative correlation was due to a few individuals with higher degrees who speak a different language at home, that is, Chinese, Tamil, or Farsi.
with different thresholds. We also are working toward simplification of relevant text features, combining lexical simplification with other forms of

Acknowledgments
The authors would like to thank their study participants. The study was reviewed by the Institutional Review Board (IRB) of Claremont Graduate University.

This work was supported by the US National Library of Medicine, NIH/NLM 1R03LM010902-01.

Conflicts of Interest
None declared.

Multimedia Appendix 1
List of conditions.
[PDF File (Adobe PDF File), 7KB - jmir_v15i7e144_app1.pdf ]

Multimedia Appendix 2
Stimuli: Original and simplified texts.
[PDF File (Adobe PDF File), 62KB - jmir_v15i7e144_app2.pdf ]

References


40. How the GRE tests are scored. URL: http://www.ets.org/gre/institutions/scores/how [accessed 2013-02-06] [WebCite Cache ID 56Fh5ILJE]


Abbreviations

ANOVA: analysis of variance
GRE: Graduate Record Examinations
SE: subject expert
UMLS: Unified Medical Language System

©Gondy Leroy, James E Endicott, David Kauchak, Obay Mouradi, Melissa Just. Originally published in the Journal of Medical Internet Research (http://www.jmir.org), 31.07.2013. This is an open-access article distributed under the terms of the Creative Commons Attribution License (http://creativecommons.org/licenses/by/2.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in the Journal of Medical Internet Research, is properly cited. The complete bibliographic information, a link to the original publication on http://www.jmir.org/, as well as this copyright and license information must be included.