Improving Perceived and Actual Text Difficulty for Health Information Consumers using Semi-Automated Methods

Gondy Leroy, PhD[1], James E. Endicott[1], Obay Mouradi[1], David Kauchak, PhD[2], Melissa L. Just, EdD[3],
[1]Claremont Graduate University, Claremont, CA; [2]Middlebury College, Middlebury, VT; [3]Rutgers University Libraries, New Brunswick, NJ

Abstract

We are developing algorithms for semi-automated simplification of medical text. Based on lexical and grammatical corpus analysis, we identified a new metric, term familiarity, to help estimate text difficulty. We developed an algorithm that uses term familiarity to identify difficult text and select easier alternatives from lexical resources such as WordNet, UMLS and Wiktionary. Twelve sentences were simplified to measure perceived difficulty using a 5-point Likert scale. Two documents were simplified to measure actual difficulty by posing questions with and without the text present (information understanding and retention). We conducted a user study by inviting participants (N=84) via Amazon Mechanical Turk. There was a significant effect of simplification on perceived difficulty (p<.001). We also saw slightly improved understanding with better question-answering for simplified documents but the effect was not significant (p=.097). Our results show how term familiarity is a valuable component in simplifying text in an efficient and scalable manner.

Introduction

As modern medicine has advanced, the number of tests, therapies and treatments has steadily increased. In addition, every day more people are diagnosed with chronic diseases requiring long-term management. More and more people are requiring accurate medical information, but few healthcare professionals are equipped or have the necessary time to meet the associated increased demand for up-to-date patient education. Making matters worse, patients have different health literacy levels, often unknown to healthcare professionals, making it difficult to provide appropriate and meaningful information at a level that is most beneficial to the patient. It is estimated that in the US, 89 million people have insufficient health literacy to understand treatments or preventive care[8]. Inadequate health literacy is defined as "limited ability to obtain, process, and understand basic health information and services needed to make appropriate health decisions and follow instructions for treatment"[28]. Studies show that medical care costs are higher for persons with lower health literacy[29]. Patients with low health literacy are less knowledgeable about their condition[14, 15], less inclined to search for information[27], less capable to learn from text[6, 10, 22], and more likely to be admitted to hospitals[4].

Different aspects influence a person's health literacy level. First and most important are personal characteristics. A person's intelligence, education, language skills and reasoning abilities play a role in the ability to understand, digest and act on information. Studies that evaluated understanding, readability grade levels, presentation and format, and education levels [7, 30, 19] showed that education level influenced understanding. Second, the type of material and the topic will affect the estimated health literacy. Although personal interaction with physicians is preferred, video, simulation and visualization provide interesting and effective alternatives. Unfortunately, they are still costly to develop and text remains the most time and cost efficient method to distribute information and educate patients. In addition, complicated procedures or the cause of a disease can be very difficult to explain and require substantial background knowledge of biology or medicine. Finally, the method of evaluating a person's health literacy will affect the estimated level. Current approaches include cloze-measures[25], a procedure where the n^{th} word is deleted and readers then fill in the blanks, multiple-choice or teach-back and open-ended questions. Each of these tests different attributes and can be expected to lead to different results.

Research on readability and increasing readability is important since two problems can be expected to develop when the available information is too difficult. The first is that consumers will not understand the information[6, 10, 22] and will be less informed. The second, a less obvious problem, is that consumers may resort to texts that seem easier to read. This easy-to-read information may be overly represented by less reputable sources, such as blogs or targeted advertising. This is troubling since in general, online health information affects decisions about health, healthcare, and visits to a healthcare provider for many readers[5].

Perceived and Actual Difficulty

We distinguish between perceived and actual difficulty, two variables that are often not separated in the literature. The distinction is necessary because the first, perceived difficulty, can be expected to influence one's willingness to read while the second, actual difficulty, can be expected to influence one's ability to understand the provided text. Perceived difficulty is also easy to measure, for example, by asking experts or consumers their opinion on the difficulty level of documents. Actual difficulty is more difficult to measure because it requires an estimate of information comprehension, retention of information or behaviors such as the ability to take appropriate actions.

Evidence for the distinction between perceived and actual difficulty comes from two models. The Health Belief Model (HBM) has been used to explain and predict health-related behaviors. The model contains four dimensions: *perceived susceptibility*, *perceived severity*, *perceived benefits*, and *perceived barriers*. In a review study[13], *perceived barriers* was found to be the most significant in explaining health behavior. Similarly in psychology, the Theory of Planned Behavior (TPB), an extension of the Theory of Reasoned Action (TRA), has been brought forward to explain behaviors[2]. It contains components that reflect perceived difficulty. The TPB contains *perceived behavioral control* as a factor and several studies have shown support for the existence of two distinct components in this factor: *perceived difficulty* and *perceived control*. In an extensive set of studies, it was demonstrated that the two can be manipulated independently and that perceived difficulty was the stronger predictor of intentions and behavior[26]. Applied to health literacy, both models point to the importance of perceived text difficulty. When the difficulty level of text is perceived as too high, it may prevent many consumers from reading and learning from text. Naturally, perceived difficulty does not tell the entire story. It is insufficient for a text to be perceived as not difficult; actual difficulty plays a major role in the resulting comprehension of information.

In summary, reading and learning from text can be encouraged by providing texts that 1) pose low barriers to reading (perceived difficulty) and 2) are at an appropriate difficulty level for consumers (actual difficulty). A low barrier to reading is important to encourage consumers to self-educate. Anecdotal evidence suggests that when readers consider a text too difficult for them, they are not inclined to read or study it[11], a fact confirmed by many physicians who can list excuses, e.g., "I broke my glasses," made by patients who did not read the provided educational pamphlets. In addition, an appropriate actual difficulty level is important to facilitate actual understanding and learning.

Text Simplification using a Semi-automated Approach

Overview. Our goal is to develop tools that help writers produce text that is easier to understand. We impose two important constraints on our work. The first is that the tools should be useful to all writers and it should not require background knowledge in linguistics or education to improve text. This requires automated detection of difficult text. The second constraint is that use of the tools should be efficient and effective. This requires integration of the tools in the writing environment.

To accomplish this goal, we break the process into two phases. The first phase is the identification of difficult text. We use natural language processing tools to find features that differ between easy and difficult text. These features can be automatically found in text and shown to the writer without the need to rely on expert evaluators. The second phase is the presentation and evaluation of simpler alternatives for difficult text. This phase consists of the development of algorithms that retrieve and list easier alternatives from lexical resources. We evaluate the efficiency of such algorithms in user studies that mimic actual conditions. A writer is presented with alternatives and chooses from them to simplify the difficult section. The simplification is then evaluated for its effects on perceived and actual difficulty by asking questions about the content of the text. We opt for using question and answering since this method has been used for centuries in schools and universities to measure knowledge. In similar research, others have evaluated the impact of successful simplification using Cloze scores[16].

Corpus Analysis. To identify features that differentiate easy and difficult text, we compare corpora consisting of known easy and difficult text. We conduct three types of analyses: lexical, grammatical and compositional analysis. We collected documents known to contain, in general, easy or difficult text. We do not rely on expert analysis, since this may be a subjective judgment that reflects perceived difficulty. We also do not rely on existing readability formulas for making this distinction, since they were developed several decades ago, tend to be fairly simplistic and have not been validated for modern and medical/health text.

In earlier work, we conducted a systematic, lexical and grammatical comparison of difficult and easy text[17, 18] by analyzing corpora but without user evaluations. The grammatical analysis identified significant differences in the occurrence of different parts-of-speech. Easy texts contain a higher proportion of function words, adverbs and verbs. Difficult texts contain a higher proportion of nouns and adjectives. These findings were consistent across three different corpora[18].

For our lexical analysis, we constructed a metric, term familiarity, which represents the difficulty level of individual words in a document[17, 18]. Familiarity of terms using existing corpora has been used by others successfully to estimate perceived difficulty of terms[12]. We estimate familiarity by using the frequency counts of words in the Google Web Corpus. The Google Web Corpus contains word *n*-gram collected by Google and based on a corpus of a trillion words from their collection of public Web pages. The corpus contains 1,024,908,267,229 tokens and 95,119,665,584 sentences. It provides frequency counts for unigrams (single word), bigrams (word pairs), trigrams, 4-grams and 5-grams in their corpus. We assume that a term that is more familiar (i.e. more frequent) will also be easier for readers to understand. This is different from what is commonly done in readability formulas, e.g., using word length or the number of syllables. For example, the Google Web Corpus shows that "apnea" is less commonly found in text than "obesity". Our corpus analysis showed that term familiarity is higher in easy text than in difficult text, a finding that also was found consistently across different corpora[17, 18].

Semi-automated Text Simplification. We report here on an algorithm for text simplification based on term familiarity. It contains two modules followed by interaction with the writer:

1) automated identification of difficult words based on term familiarity
2) automated retrieval and rank-ordering of potential alternatives based on term familiarity
3) the writer chooses one of the alternatives to replace the difficult text segment or word

Automated identification of difficult words is based on data gathered from the Google Web Corpus unigram frequencies. In this corpus words are given a frequency based on the number of occurrences they have in Google's index of English language websites on the Internet. A frequency of 15,377,914 was chosen to be the threshold between easy and difficult words, which is the frequency of the 5,000th most common word in the corpus. Based on this cutoff, approximately 85% of all words in English are classified as easy with the remaining 15% difficult.

For each word labeled as difficult, a *list of potential replacements is generated and ranked*. The replacements can be single words or text snippets. They are retrieved from different sources: WordNet 2.1[20], the Unified Medical Language System (UMLS) Metathesaurus, simple English Wiktionary (http://simple.wiktionary.org) and regular English Wiktionary (http://en.wiktionary.org). Each source is searched for simpler alternatives. Alternatives are restricted based on the part-of-speech (POS) of the difficult word in the original text and by requiring a minimum term familiarity of the alternative. The POS of the original text is labeled using the General Architecture for Text Engineering (GATE)[9] and its implementation of the Hepple tagger. Only those entries in the replacement resource that match the original POS are considered. By identifying the POS of the word in the original text, the number of alternatives can be limited to those with the correct POS. However, when only one POS is present in the source, the designation is disregarded. Similarly, if multiple POS are present but none match, the designation is also disregarded. Term familiarity is used to ensure that no difficult word is replaced by a more difficult alternative. The minimum term familiarity of the alternative word (or of each word in a text snippet) needs to be greater than that of the difficult word to be replaced.

Below are each of the replacement resources used and how initial candidates are generated:

- *WordNet 2.1 synonyms and hypernyms*. 1) synonyms. One entry from each synset of the designated POS is selected. The selection is made based on term familiarity and the entry with highest term familiarity of all words is selected. All words in that entry that meet the term familiarity criteria are added to the list of potential replacements. 2) hypernyms. The selection process is repeated for the first hypernym for each synset of the designated POS. To keep the list manageable, only hypernyms one level removed are included. Hypernyms that meet the term familiarity criteria are added to the list of alternatives as follows: "<difficult word>, a kind of <hypernym>,"; "<difficult word>, kinds of <hypernym>,"; or "<difficult word>, a way to <hypernym>," depending on the POS.

- *UMLS Metathesaurus semantic types*. From this resource, we use the semantic types in a similar manner to the WordNet hypernyms. All semantic types of a word are retrieved first. If they meet the term familiarity requirement, they are appended to the list of potential alternatives as : "<difficult word>, a kind of <semantic type>."

- *WordNet and Wiktionary definitions*. Definitions for the matching POS are selected from WordNet, simple and regular English Wiktionary. The last two were included to provide more options since we do not limit our algorithm to replacing only medical terms. Alternatives are excluded if they are indicated to be "slang", "archaic", or "obsolete". Furthermore, definitions are truncated at the first occurrence of a semicolon, or the occurrence of "for example", "as in", "especially", or "such as". Only definitions where each word meets the term familiarity requirement are included as a potential alternative. Definitions are appended depending on their source. Simple English Wiktionary entries are appended to the list unmodified. Definitions from WordNet and regular English Wiktionary are modified to be in one of the following forms "<difficult word> is a <definition>"; "When something is <difficult word>, it is <definition>"; or "Something that is <difficult word>, is <definition>".

If the list of potential replacements does not contain at least seven options we also consider more distant hypernyms from WordNet that are one, two or three levels removed.

All alternatives are shown along with the original difficult word to the writer. The options are ordered with the original word first, then synonyms and finally hypernyms and UMLS semantic types. Within each category, words are ordered according to term familiarity. If additional hypernyms were added, they also are ordered by how far removed they are in the WordNet hierarchy.

The simplification process is completed by interacting *with the writer*. In our study, a medical librarian used the alternatives suggested by our algorithm above to simplify text. We ensured that a systematic approach was followed to facilitate evaluation of our algorithm. In particular, we did not encourage subjective changes in the text not offered by the algorithm. When the writer chose a different alternative, not in the list, it needed to have a higher familiarity than the original term. In the present study, two such words suggested by the librarian (not by our algorithm) for the document simplification were allowed and one was rejected because its term familiarity was too low. All other changes were based on our suggested alternatives. However, the librarian was responsible for choosing one of the alternatives (or keeping it the same) and making sure the selection was correct in the context of the sentence. When the best alternative, according to the librarian was a text snippet, this snippet was added as a separate sentence before or after the sentence containing the difficult word (see example below).

User Study

Design. The goal of our user study is to measure the impact of the text simplification algorithm described above on perceived and actual difficulty of text. We include two study segments that allow us to independently measure perceived and actual difficulty. We also include demographic information questions and the Short Test of Functional Health Literacy in Adults (STOFHLA)[3, 21].

The first study segment focuses on measuring the *perceived difficulty* of text. We show two versions of a sentence to the participants and ask them to judge each version on a Likert-scale. The dependent variable is the score on a 5-point Likert-scale with a score of 1 representing a very easy sentence and a score of 5 representing a very difficult sentence. Our main independent variable (IV) is the sentence version. It has two conditions: original or simplified text. We also include a second IV, the Flesch-Kincaid grade level, for comparison with existing work. We include two general conditions: high and low readability grade level. We measured the grade level using Microsoft Word.

Twelve sentences were copied from patient materials available online (http://health.ucsd.edu/healthinfo/Pages/default.aspx?docid=/Library) from the University of California, San Diego (UCSD) Health System. To have enough diversity in our sentences and allow for better comparison with other work, we selected six sentences with a low and six with a high Flesch-Kincaid grade level (See Table 1). Each of the twelve sentences was processed by our simplification algorithm and a medical librarian chose from the alternatives provided by our algoritm to simplify the sentence. Table 1 provides an overview of the characteristics. Term familiarity is the weighted average of the Google frequencies of meaning bearing words (adjectives, nouns, adverbs, verbs).

- Original Sentence: "*Botulism is a neuromuscular disease caused by a bacterial toxin acting in the intestine and causing neuromuscular poisoning"*
- Simplified Sentence: "*Botulism is a neuromuscular disease caused by a bacterial toxin, which is a kind of poison. The poison acts in the intestine and causes neuromuscular poisoning. This means it affects both neural and muscular tissue."*

Each study participant judged all sentences. Sentences were presented in pairs and the pairs were shown in random order. We learned from previous studies that not pairing up the sentences makes the task very difficult and results in nonsensical scores. By presenting the two levels to each user we use a within-subjects design, which helps control for other variables such as prior knowledge of the topic or language skills.

Table 1. Sentence and Paragraph Descriptive Information.

		Text Version	
		Original	**Simplified**
	SENTENCES		
Avg. Word Count	Sentences w. Low Flesch-Kincaid Grade Level:	17.7	25.8
	Sentences w. High Flesch-Kincaid Grade Level:	27.0	45.7
	All Sentences:	22.3	35.8
Avg. Flesch-Kincaid Grade Level	Sentences w. Low Flesch-Kincaid Grade Level:	8.3	6.2
	Sentences w. High Flesch-Kincaid Grade Level:	20.0	11.2
	All Sentences:	14.1	8.7
Avg. Term Familiarity	Sentences w. Low Flesch-Kincaid Grade Level:	376,032,628	749,034,879
	Sentences w. High Flesch-Kincaid Grade Level:	347,969,408	911,144,848
	All Sentences:	362,001,018	830,089,864
	PARAGRAPHS		
Word Count	Pemphigus:	198	288
	Polycythemia Vera:	199	272
	Both Documents (Avg.):	199	280
Unique Word Count	Pemphigus:	107	134
	Polycythemia Vera:	113	135
	Both Documents (Avg.)	110	135
Flesch-Kincaid Grade Level	Pemphigus:	12.3	10.0
	Polycythemia Vera:	12.5	9.4
	Both Documents (Avg.):	12.4	9.7
Avg. Term Familiarity	Pemphigus:	1,205,490,261	2,069,728,988
	Polycythemia Vera:	552,669,521	1,259,835,071
	Both Documents:	879,079,891	1,664,782,029

The second study segment focuses on measuring the *actual difficulty* of text. As with the previous segment, the first IV is text version: original or simplified. To measure understanding, a text is presented and then the user is asked to answer multiple-choice and True/False questions. These questions types were chosen over open-ended questions, since they allow for objective scoring. We measure actual difficulty by asking questions with and without the text present, the former to test understanding and the latter to test retention. All questions focused on the content of the document and the same questions were used for both versions of a text. Questions and answers were carefully phrased so that no terminology was used that resembled or was biased toward one of the versions.

For this segment two documents were chosen from the same online resource, one on polycythemia vera and on one pemphigus. Both original documents are comparable in length and difficulty level (See Table 1). The topics were chosen because they are fairly rare diseases of which few people will have prior knowledge. Each was simplified by the medical librarian relying on the alternatives suggested by our algorithm.

By working with two different documents, we can present both an original and a simplified document to each participant. The order of the text and the version was counterbalanced, so that all possible orderings were presented an equal number of times. For example, a simplified polycythemia vera document was followed by an original pemphigus document and vice versa.

To measure understanding (actual difficulty measure), we ask the participants to answer multiple-choice questions with four options with the text present. Each question receives a score of 1 for a correct answer and 0 for an incorrect answer. Participants must answer all questions. Since the text was present, we opted not to use an "I don't know" option. Final scores are presented as percentage correct on a scale of 0 to 100.

To measure retention of information (actual difficulty measure), we ask the participants to answer 30 T/F questions without the texts present (15 for each topic). The T/F statements are presented in random order and for each statement participants need to indicate True, False or 'I don't know'. We assign a score of 1 to a correct answer, -1 to an incorrect answer and 0 for 'I don't know'. For consistency, final scores are also reformulated and presented on a scale of 0 to 100.

Study Participants. To allow us to conduct a study with a broad range of online consumers we used Amazon Mechanical Turk (AMT). AMT is a service where participants perform 'human intelligence tasks' (HITs) for a small payment (www.mturk.com). The service is widely used, for example in 1997 there were already more than 100,000 participants (called 'workers')[1] and in January 2010, there were more than 170,000 HITs available. Keyword descriptions and a short introduction allow participants to choose from available HITs. Although use of AMT is relatively new for healthcare studies, it has gained acceptance in other fields. Research suggests that data retrieved from AMT studies are approximately as reliable as traditional studies[23, 24]. And even though a higher rejection rate of data is often required due to users not completing surveys or not taking them seriously, the ease of conducting these studies makes up for this shortcoming.

Before conducting the present study, we conducted 2 pilot studies to verify our programming, database and scoring. We found that some AMT users use 'bots' to automatically provide answers (and get paid) while some other AMT users do not read the questions at all (random answers). To allow us to eliminate such participants, we include qualifying questions in different sections. In each multiple-choice section, we added one question that could easily be answered correctly *if* the participant read it. In the T/F section, we added two such qualifying questions. Participants who do not answer all of these questions correctly have all of their responses eliminated from the data set. For example, the qualifying question for the document discussing polycythemia is:

This text covers the following topic:

- *Post festum*
- *Popular demand*
- *Potato salad*
- *Polycythemia*

Procedure. Our study was hosted on a local server and all data was saved locally. A study description with a link to the study was posted on AMT. Participants were offered $1 if they completed the entire study. After following the link to the study, each participant was presented with the study sections ordered as follows:

1) Welcome page
2) Understanding - Text 1 with 4 multiple-choice questions and 1 qualifying question, in random order
3) Understanding - Text 2 with 4 multiple-choice questions and 1 qualifying question, in random order
4) Demographic questions
5) STOFHLA – Part A and B
6) Retention – 30 statements in random order, 15 for each topic
7) Final page with 'hit code' which is submitted on AMT to receive payment

Results.

After posting our study on AMT, 126 participants registered for the study. Of this group, 32 did not complete the study and 10 more were disqualified based on our qualifying questions. The remaining 84 participants completed the study. Two input errors resulted in demographic information being available for only 82 participants. Other results are based on all 84 participants.

Demographic Information. Table 2 shows an overview of the demographic information. Both genders were equally well represented with slightly more men (55%) than women (45%) in our sample. Most of the participants were white (80%) with lower representation of Asian (17%), Black / African American (4%). There were 5% who indicated Hispanic or Latino ethnicity.

Table 2. Demographic Information of Participants.

Category		% (N=82)
Gender	Female	45
	Male	55
Race	American Indian or Alaska Native	0
	Asian	17
	Black or African American	4
	Native Hawaiian or Other Pacific Islander	0
	White	80
Ethnicity	Hispanic or Latino	5
	Not Hispanic or Latino	95
STOFHLA	Average	32.9
	Minimum	16
	Maximum	36
Education	Less than high school	2
	High school diploma	35
	Associates	20
	Bachelor	24
	Master	18

Perceived Difficulty. On a scale of 1 (very easy) to 5 (very difficult), the original six sentences received an overall average score of 2.56 while the simplified six versions received an overall average score of 2.44, i.e. easier. The differences are consistent for each of the 12 sentence but they are larger for sentences that also have high Flesch-Kincaid readability grade levels as is shown in Figure 1. We conducted a repeated-measures ANOVA with sentence condition (original or simplified) as the within-subject variable since every subject evaluated both version of each sentence. We found a significant effect ($p<.001$) of text simplification.

In a second analysis, we then included the Flesch-Kincaid grade level (low or high) as a second independent variable. This two-way ANOVA showed a significant main effect on perceived difficulty of text simplification with simplified sentences considered easier ($p<.001$). A second main effect was found for the Flesch-Kincaid grade level, with sentences with a low grade level being seen as easier ($p<.001$). Finally, we found that the interaction between both variables was also significant ($p<.001$). Figure 1 shows how the difference in perceived difficulty is larger for sentences that have also a high Flesch-Kincaid readability grade level.

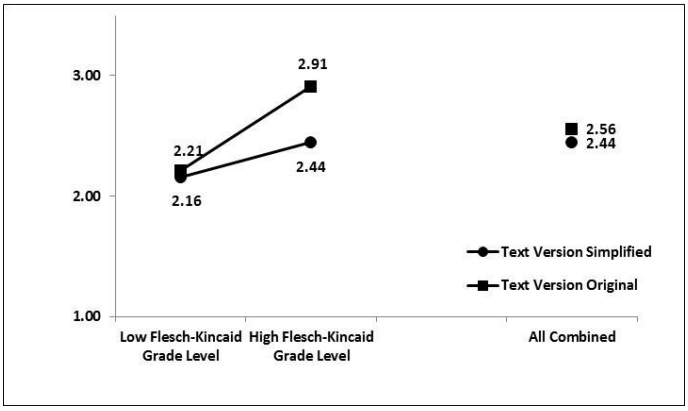

Figure 1. Perceived Difficulty of Sentences (0 = Very Easy, 5 = Very Difficult)

Actual Difficulty. Figure 2 shows an overview of the scores for the understanding of information. Overall, participants scored 58% correct with the simplified text and 51% correct with the original text. The difference

between the two text versions is much larger for the document on polycythemia vera. With this document, participants scored 71% correct with the simplified version and 58% correct with the original document. For the document on pemphigus, the simplified document led to a 2% decrease in scores.

We conducted a one-way ANOVA with text simplification (original vs. simplified) as the independent variable. We found that the difference in scores for the two conditions was not significant (p=.097). We then conducted a two-way ANOVA with text simplification (original vs. simplified) and document topic (pemphigus vs. polycythemia vera) as the independent variables. We found a significant main effect of topic (p<.001) with higher scores for the document on polycythemia vera. Figure 2 shows a small difference of simplification for the document on pemphigus, but a substantial (13%) increase in scores for the document on polycythemia vera. The interaction between the two variable was not significant (p=.055).

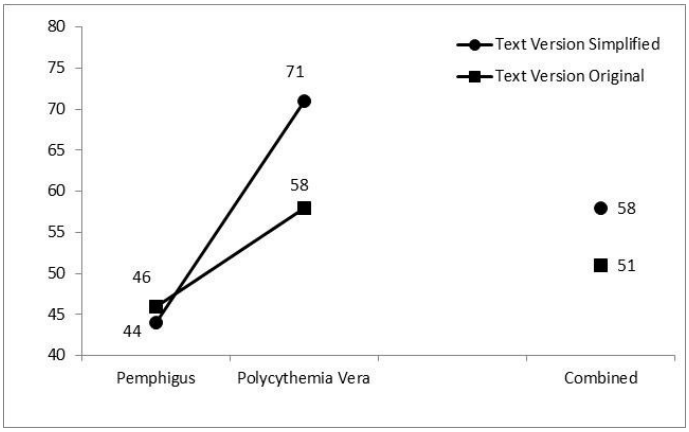

Figure 2. Actual Difficulty: Understanding of Information

Figure 3 shows an overview of the scores for retention of information. The results for retention look similar to those for understanding but the differences are smaller. Overall, participants scored 73% correct for statements about the simplified document and 71% for the original document. The difference is larger for the polycythemia vera document with 78% correct for the simplified document and 74% for the original document. The scores were 3% worse for the simplified pemphigus document compared to the original version.

A one-way ANOVA with text version as the independent variable showed no significant effect. A two-way ANOVA with text simplification (original vs. simplified) and document topic (pemphigus vs. polycythemia vera) as the independent variables also showed no significant main effects and no significant interaction effect (p=.058).

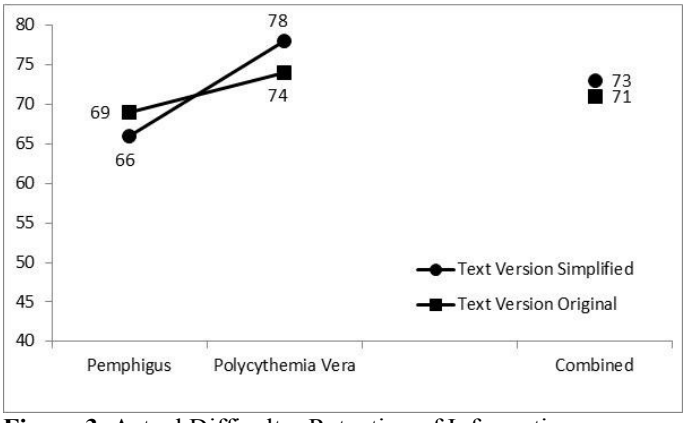

Figure 3. Actual Difficulty: Retention of Information

Discussion

Even though not all differences were significant, there are several interesting findings in our study. First, even though the documents have similar Flesch-Kincaid grade levels, the results for question and answering indicate different difficulty levels. This suggests that our metric is more sensitive to differences in difficulty and a better measure of current reading levels.

The results also suggest that either our changes to the text were not drastic enough to make a significant difference or that the effect is so small that a much larger sample is needed (e.g., the observed power for understanding was .299 with alpha .05). In addition to doing larger studies, we will evaluate whether more words could have been replaced in the text. Lowering the threshold, which was arbitrarily chosen, of what is 'difficult' would enable this, though this would increase the number of words the writer would have to examine.

Finally, measuring actual difficulty of a document is difficult to do objectively. The phrasing of questions and the type of questions asked will affect the scores. We have taken care to avoid phrasing our questions with text present in any of the document versions, however, the difference between topics may indicate that one document was generally more difficult, or that we asked more difficult questions.

Conclusion

Overall, our goal is the development of text simplification algorithms that can support writers in simplifying text. We conduct corpus analyses to identify features that differ between easy and difficult text. Candidate features are used in our algorithms to support writers. In this paper, we reported on one such feature, term familiarity, and how it can be incorporated in a semi-automated text simplification algorithm.

For our study, we were true to our self-imposed constraints. Difficult sections in text were automatically identified so that writers do not need expertise in linguistics or education (first constraint) and alternatives were suggested automatically for each such difficult section (second constraint). After conducting our user study, we found that the simplified text was perceived as significantly easier. In addition to this perceived difficulty, we also tested actual difficulty by using a question answering approach. We found that our algorithm can help simplify text, but the effect also depends on other features of the original document.

Finally, we consider term familiarity a valuable metric in evaluating text difficulty and for supporting text simplification. It is, however, only one feature and in the future we intend to test and combine additional features for semi-automated text simplification.

Acknowledgements

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

References

1 Artificial Intelligence, With Help From the Humans. The New York Times. 25 March 2007.

2 Ajzen I. The Theory of Planned Behavior. Organizational Behavior and Human Decision Processes. 1988;50:179-211.

3 Baker D, Williams M, Parker R, Gazmararian J. Development of a brief test to measure functional health literacy. Patient Educ Couns 1999;38(1):33-42.

4 Baker DW, Gazmararian JA, Williams MV, Scott T, Parker RM, Green D, et al. Functional Health Literacy and the Risk of Hospital Admission Among Medicare Managed Care Enrollees. American Journal of Public Health. 2002;92(8):1278-83.

5 Baker L, Wagner TH, Signer S, Bundorf MK. Use of the Internet and E-mail for Health Care Information: Results from a National Survey. Journal of the American Medical Association. 2003 May 14;289(18):2400-6.

6 Berland GK, Elliott MN, Morales LS, Algazy JI, Kravitz RL, Broder MS, et al. Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish. JAMA. 2001;285:2612-21.

7 C. A. Weaver I, Renken A. Applied Pscyhology of Readability. International Encyclopedia of the Social & Behavioral Sciences. 2004:12789-91.

8 Committee on Health Literacy - Institute of Medicine of the National Academies, ed. Health Literacy: A Prescription to End Confusion. Washington, DC: The National Academies Press 2004.

9 Cunningham H. GATE, a General Architecture for Text Engineering. Computers and the Humanities. 2002 May;36(2):223-54.

10 D'Alessandro D, Kingsley P, Johnson-West J. The Readability of Pediatric Patient Education Materials on the World Wide Web. Arch Pediatr Adolesc Med. 2001;155:807-12.
11 Davis TC, Dolan NC, Ferreira MR, Tomori C, Green KW, Sipler AM, et al. The Role of Inadequate Health Literacy Skills in Colorectal Cancer Screening. Cancer Investigation. 2001;19(2):193-200.
12 Elhadad N. Comprehending Technical Texts: Predicting and Defining Unfamiliar Terms. AMIA Annu Symp Proc; 2006; 2006. p. 239–43.
13 Janz NK, Becker MH. The Health Belief Model: A Decade Later. Health Education Quarterly. 1984;11(1):1-47.
14 Kalichman SC, Benotsch E, Suarez T, Catz S, Miller H, Rompa D. Health Literacy and Health Related Knowledge Among Persons Living with HIV/AIDS. American Journal of Preventive Medicine. 2000;18(4):325-31.
15 Kandula NR, Nsiah-Kumi PA, Makoul r, Sager J, Zei CP, Glass S, et al. The Relationship between Health Literacy and Knowledge Improvement after a Multimedia Type 2 Diabetes Education Program. Patient Education and Counseling. 2009:(In Press).
16 Kandula S, Curtis D, Zeng-Treitler Q. A Semantic and Syntactic Text Simplification Tool for Health Content. AMIA Annu Symp Proc; 2010; 2010. p. 366–70.
17 Leroy G, Endicott JE. Term Familiarity to Indicate Perceived and Actual Difficulty of Text in Medical Digital Libraries. International Conference on Asia-Pacific Digital Libraries (ICADL 2011) - Digital Libraries -- for Culture Heritage, Knowledge Dissemination, and Future Creation; 2011 October 24-27; Beijing, China; 2011.
18 Leroy G, Endicott JE. Combining NLP with Evidence-based Methods to Find Text Metrics related to Perceived and Actual Text Difficulty. 2nd ACM SIGHIT International Health Informatics Symposium (ACM IHI 2012); 2012 January 28-30; Florida, Miami; 2012.
19 Mazor K, Dodd K, Kunches L. Communicating Hospital Infection Data to the Public: A Study of Consumer Responses and Preferences. Am J Med Qual. 2009 24(2):108-15.
20 Miller GA, Beckwidth R, Fellbaum C, Gross D, Miller K. Introduction to WordNet: An On-line Lexical Database. 1998 [cited; Available from: http://www.cogsci.princeton.edu/~wn
21 Nurss JR, Parker RM, Williams MV, Baker DW. Test of Functional Health Literacy in Adults. Hartford, MI: Peppercorn Books & Press 1995.
22 Root J, Stableford S. Easy-to-Read Consumer Communications: A Missing Link in Medicaid Managed Care. Journal of Health Politics, Policy, and Law. 1999;24:1-26.
23 Schnoebelen T, Kuperman V. Using Amazon Mechanical Turk for linguistic research. Research Methods and Techniques. 2010;43(4):441-64.
24 Sprouse J. A Validation of Amazon Mechanical Turk for the Collection of Acceptability Judgments in Linguistic Theory. Behavioral Research Methods. 2011;43(1).
25 Taylor WL. Cloze procedure: A new tool for measuring readability. Journalism Quarterly. 1953; 30:415-33.
26 Trafimow D, Sheeran P, Conner M, Finlay KA. Evidence that Perceived Behavioral Control is a Multidimensional Construct: Perceived Control and Perceived Difficulty. British Journal of Social Psychology. 2002;41:101-21.
27 von Wagner C, Semmler C, Good A, Wardle J. Health Literacy and Self-efficacy for Participating in Colorectal Cancer Screening: The Role of Information pprocessing. 2009.
28 Weis BD. Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians. Second Edition ed: AMA and AMA Foundation 2007.
29 Weiss BD, Palmer R. Relationship Between Health Care Costs and Very Low Literacy Skills in a Medically Needy and Indigent Medicaid Population. JABFP. 2004;17(1):44-7.
30 Yan X, Song D, Li X. Concept-based document readability in domain specific information retrieval. *Proceedings of the 15th ACM international conference on Information and knowledge management*. Arlington, Virginia, USA: ACM 2006.