HHS Public Access

Author manuscript

IT Prof. Author manuscript; available in PMC 2017 May 25.

Published in final edited form as:

IT Prof. 2016 ; 18(3): 45–51. doi:10.1109/MITP.2016.50.

Moving Beyond Readability Metrics for Health-Related Text Simplification

David Kauchak, Ph.D.[1] and **Gondy Leroy, Ph.D.**[2]

[1]Computer Science Department, Pomona College, Claremont, CA

[2]Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ

Abstract

Limited health literacy is a barrier to understanding health information. Simplifying text can reduce this barrier and possibly other known disparities in health. Unfortunately, few tools exist to simplify text with demonstrated impact on comprehension. By leveraging modern data sources integrated with natural language processing algorithms, we are developing the first semi-automated text simplification tool.

We present two main contributions. First, we introduce our evidence-based development strategy for designing effective text simplification software and summarize initial, promising results. Second, we present a new study examining existing readability formulas, which are the most commonly used tools for text simplification in healthcare. We compare syllable count, the proxy for word difficulty used by most readability formulas, with our new metric 'term familiarity' and find that syllable count measures how difficult words 'appear' to be, but not their actual difficulty. In contrast, term familiarity can be used to measure actual difficulty.

Keywords

consumer health information; text readability; text simplification; health literacy; readability formulas

Introduction

As lifespans increase and medical knowledge improves, increasingly patients are expected to participate in managing their health. Participation necessitates that they understand health information presented to them, which requires that the information be presented in an understandable way. Current approaches to creating understandable text are expensive and time consuming. The resulting lack of optimized health information has become a widespread national public health problem with linkages to widening health disparities, less informed healthcare decision-making, and higher healthcare consumption. Limited

Corresponding Author: David Kauchak, Computer Science Department, 185 E 6th St, Pomona College, Claremont, CA 91711, Phone: 909-607-0473, david.kauchak@pomona.edu.

comprehension prevents an estimated 90 million Americans from obtaining, understanding and acting upon health information [1, 2].

Text can be informative and is easy to distribute broadly; the key challenge is writing easy-to-understand text efficiently. This is particularly important in healthcare where costs due to limited health literacy are estimated to be $238 billion annually [3]. Facilitating comprehension from text is one essential element in increasing health literacy and for decades governments and advocacy groups have encouraged writing texts that use ‘plain language’ and have high ‘readability’. These guidelines provide high-level guidance but are not backed with concrete guidance or effective tools. The most frequently used tool for ensuring that health-related texts are readable is the readability formula, e.g. Flesch-Kincaid grade level formula [4].

Even though formula use is excessive, few if any peer-reviewed studies show a positive impact of their application on reader comprehension. Readability formulas have not been successful for simplifying health-related texts because: 1) their outcomes are not associated with actual understanding, 2) they do not identify what aspects of a text are difficult, often only providing a single, numerical score for an entire text, 3) relatedly, alternatives for rewriting are not provided, 4) the features utilized by these formulas are rudimentary and do not capture the complexity of the concepts or the cohesiveness and organization of the text [5], ignore global text characteristics including the fluency, structure and content of the text, and do not incorporate current knowledge about the reading process [6].

Our goal is twofold: to understand the problem of text comprehension and readability, and to provide more intelligent and usable tools. We briefly address both in this paper. First, we introduce and argue for a data-driven approach to developing text simplification tools that utilizes large corpora, machine learning and concrete validation of the tool. Tool development is accomplished incrementally using an evidence-based development strategy: only algorithms with proven effect on user comprehension are included. We have already discovered several text features and have shown their relationship to comprehension in both English and Spanish. Our tools leverage a number of domain agnostic components, but the reliance on healthcare-related resources and testing on medically relevant text makes it most useful in the healthcare domain. Second, we provide a study comparing the common feature of syllable count currently used to guide simplification in medicine and find that it does not help differentiate between simple and difficult words.

Existing Approach: Readability Formulas

The most commonly used formulas in the health domain are the Flesch-Kincaid readability formulas [7], which measure text difficulty using two components: average number of syllables per word and average number of words per sentence. For example, the Flesch-Kincaid grade level formula is:

$$0.39^{*}\text{Word\_per\_sentence}+11.8^{*}\text{syllables\_per\_word}-15.59$$

The result is then interpreted as the U.S. school grade level needed to understand the text: 1-12 corresponding to grades 1 through 12, 13 representing the first year of college, etc. Other prevalent readability formulas also rely heavily on word and sentence length, e.g. the Simple Measure of Gobbledygook (SMOG) [8], Gunning-Fog index, DISCERN [9] and HON code (http://www.hon.ch/).

To apply a readability formula, content creators use a tool that calculates the formula value of a text (e.g. Microsoft Word has a calculator). If the number from the formula indicates that the text is too difficult, the author must adjust the text to try and reduce the difficulty. Two critical problems exist with this approach. First, there is no guidance on what to change. Even for users who understand the formulas, the only guidance that can be inferred is to use shorter sentences and words with fewer syllables. Second, improving the readability metric score does not guarantee better comprehension by readers. Even newer readability measures that use additional text characteristics still suffer from these two problems [10, 11]: text simplification requires information and guidance in addition to assessment.

A Computational, Evidence-Based Approach

We argue for a text simplification tool that guides simplification with concrete suggestions and is designed using evidence from large-scale data sets and evaluated through interaction with representative readers. Figure 1 shows an overview of our tool design approach.

First, we identify candidate “interesting features” for discriminating between simple and difficult texts. Table 1 shows an overview of candidate features. The features are derived from existing theory, common advice used in practice, and from data analysis. They span different levels of text, ranging from single words and phrases to grammatical structure to document-level phenomena.

From this initial set of features, we examine their occurrence in “parallel” corpora consisting of simple and difficult texts on the same topics where the difficulty can be implied, e.g. patient blogs (simple) vs. medical journal abstracts (difficult). For those features that “differentiate” between text difficulty levels, we verify that they can be accomplished algorithmically and can also be used for simplification, e.g. by suggesting simpler alternatives. Those that meet all these criteria are “feasible features” (the process outlined in the top third of Figure 1).

Besides corpus analyses, it is critical to identify features that actually have an impact on reader understanding. To verify this, for each feasible feature, we conduct large-scale user studies to verify their effectiveness. We utilize Amazon's Mechanical Turk (MTurk), which allows for studies to be accomplished efficiently on a large, demographically diverse group [15]. When precautions are taken, data quality from MTurk has been shown to be at least as good as with traditional approaches [15]. All studies are conducted using existing healthcare text, which is simplified using the feature being tested. Those features that can be shown to impact understanding positively result in the “verified features”. Those features shown in **bold** in Table 1 represent features that have been vetted at some level, either as “feasible” features (corpus verified) or “verified” features (user study verified).

Finally, those features that are shown to be effective at improving user understanding will be combined into the final tool. To thoroughly test the final tool, we will validate it in a real-world setting through user studies in clinical environments. The tool is designed to be used *by* health educators to produce text *for* patients (e.g. medical instructions, prescription directions, clinical trial materials, etc.).

Available Data Sources and Tools

Critical to this type of data-driven approach are resources, in particular, corpora, datasets and text analysis tools.

- *Corpora* are necessary for initial feature validation, particularly parallel data that include both simple and difficult texts since they allow for concrete comparison of features across different difficulty settings. Simplifying health-related text requires simplifying both medical and non-medical terms, so both types of text are useful. Many corpora exist out there, though ones we have frequently used include general domain corpora like the Google Web Corpus [12], Corpus del Español, English Wikipedia and Simple English Wikipedia, along with medical specific data sources like PubMed, Cochrane and patient health blogs.
- *Structured data sources:* Once candidate features have been identified from corpus studies, algorithms must be developed to suggest simplifications. General dictionaries and thesauri can be useful, e.g. WordNet and English Wiktionary and medical-specific resources are also available, such as the Unified Medical Language System (UMLS) and Medical Subject Heading (MeSH) hierarchy.
- *Tools: S*oftware tools are also required to process the text and build algorithms for suggesting alternatives. Tools available include tokenizers, sentence splitters, parsers (Stanford Parser, Berkeley Parser, Freeling), part-of-speech taggers and tool aggregators like GATE, Odin and NLTK.

An Example of our Workflow: Term Familiarity

As an example of this development process, we highlight *term familiarity*, one feature that we believe will be useful for text comprehension. In our study below, we compare and contrast it with syllable count, the metric imbedded in most readability formulas. We quantify familiarity by measuring the frequency of a word/term in a large corpus of text. In English, we use the Google Web Corpus, an n-gram corpus containing counts from a trillion words from public web pages. In Spanish, we use the Corpus del Español, a 20,000-lemma list with frequencies. To first validate this as a “feasible feature”, we employed corpus studies. In both English and Spanish, average word frequency is higher in easy texts [16, 17].

For English, we have already “verified” this feature in an individual user study. We simplified health-related texts by automatically identifying unfamiliar (low frequency) words and then suggesting candidate, higher frequency simplifications on existing text resources (UMLS, WordNet and Wiktionary). A medical librarian then chose from this list

of candidate simplifications to generate the final output. Using this text simplification technique we demonstrated with several user studies that this algorithmic tool for increasing average term frequency produces text that participants view as easier and that is easier to understand [14, 16].

Additional features go through this same procedure. The benefit of this framework is that it can easily accommodate additional features. Since each individual feature represents a single method for simplifying text, the final tool, will combine all verified features. For example, the tool could include lexical simplification based on term familiarity and grammatical simplification guided by verified structure changes, like utilizing connectors and spatial coherence. We will validate this tool with user studies with health educators, i.e., to design an interface with optimal presentation of simplification alternatives, and with patients, i.e., to measure impact (bottom third of Figure 1).

The Problem With Readability Formulas: An Example

One of the main drivers for examining a data-driven approach to simplifying health-related materials is that the current approaches (readability formulas) are, at best, ineffective and can be counterproductive in some instances. To illustrate this problem, we examined how well syllable count and term familiarity correlate with a person's imagined (perceived) and actual understanding of a word.

Data Set Creation

We used all 13.6 million unique words in the Google Web Corpus and created 11 different frequency bins based on their occurrence on the web: top 1% most frequent, 1-10% based on frequency, 10-20%, …, 10% least frequent words. For the experiment, we randomly selected 25 words from each bin for a total of 275 words with a range of frequencies.

For each word, we measured the difficulty with readers using two metrics: perceived difficulty and actual difficulty. Perceived difficulty quantifies how difficult a word is perceived to be and was measured on a 5-point Likert scale (1: ‘Very Easy’ to 5: ‘Very Difficult’). Actual difficulty quantifies whether readers knew the meaning of the word and was measured using a multiple-choice test. For each word participants were presented with five definitions, one correct and four randomly chosen from one of the other words in the data set. Word definitions were obtained from the Moby Word List, part of the Moby Project (http://icon.shef.ac.uk/Moby/).

We collected evaluations from 50 participants *per word* using Amazon's Mechanical Turk resulting in 13,750 data points. We averaged the 50 scores per word, which resulted in a data set of 275 words with both perceived difficulty (score between 1 and 5) and actual difficulty (percentage correct). Separating out both perceived and actual difficulty allows us to analyze two different aspects of text difficulty, a distinction often ignored by others.

Syllables as a Measure of Word Difficulty

For each of the 275 words, we calculated different metrics for quantifying word difficulty: term familiarity, as described above, the number of syllables and the number of characters.

We used the Knuth-Liang algorithm to determine word syllable counts [13], which has been widely employed, including in Latex. We calculated a 2-tailed Pearson's correlation between the human difficulty measures and the different metrics for quantifying word difficulty.

Figure 2 shows the results, aggregated by the number of syllables (top), by the number of characters (middle) and, term familiarity bin (bottom). Groups with only one word in it were combined with the adjacent group (e.g. there was only on word with 7 syllables). For perceived difficulty, longer words were seen as more difficulty, though some very long words (5 and 6 syllable words and 13 character words) were seen as slightly easier. Over the 275 words, both the number of syllables and the number of characters correlated with user judgments of how difficult the words looked ($r = 0.177$, $p < 0.01$ and $r = 0.254$, $p < 0.001$, respectively). *Words that have more syllables and longer words are perceived as being more difficult*.

For actual difficulty, there is no such trend. The difference in understanding between words with a small number of syllables is small, particularly 2 and 3 syllable words and words containing 5-10 characters, which make up a majority of the words. Additionally, for very long words (5 and 6 syllables), users performed better than on any of the short syllable words. Because of this, over the 275 words there is no significant correlation between the percentage of participants who knew the correct definition of a word and the syllable count or the length of that word. *Words that have more syllables and longer words are not more difficult than shorter words: word length does not impact understanding*.

For comparison, the bottom of Figure 2 shows the results for the same data with word difficulty quantified with term familiarity (in this case, frequency in the Google Web Corpus). Unlike syllable count and word length, term familiarity shows a strong, consistent trend for both perceived difficulty ($r = 0.219$, $p < 0.001$) and actual difficulty ($r = -0.397$, $p < 0.001$). *Words that are less frequent both look more difficult and are less likely to be known*.

This study highlights two critical failings with readability formulas that use length as a proxy for difficulty. First, word length is not indicative of user understanding. Second, word length does indicate whether people *think* a word is easier to understand. Guided by these formulas, medical writers will select words with fewer syllables and may also perceive the word as being simpler. Unfortunately, this does not mean that the words are actually easier to understand. As a result, the use of formulas may be counter-productive to helping text consumers.

Future Work

The final tool we are creating will be available for free online. The main users are intended to be providers of health and medical information, with applications including patient information materials, online medical sources (e.g. Cochrane database) and clinical trial materials. They will type or upload text to be simplified via our publicly available website. On the server, the text will be preprocessed, difficult text components identified, and candidate simplifications generated. A marked up version of the text document will be provided to the user with difficult components flagged and candidate simplifications shown.

The user can modify the text based on the suggestions. Although online software has limitations, here it provides many benefits. First, many of the resources required for the simplification algorithm are large and have licensing restrictions. An online application allows these resources to be stored and accessed with the application. Second, it allows for a broad range of users to be reached both from a location standpoint and from a system standpoint (i.e. different platforms, operating systems, etc.). Finally, it allows for rapid deployment of new updates/development.

Conclusions

Readability measures, particularly those currently recommended (and enforced) in the medical field, are not an effective text simplification tool for improving understanding of health-related texts. We are using a data-driven framework to develop new tools that improve user comprehension and show, using term familiarity as an example, how this approach can be used to create an effective and efficient alternative. As long as text resources and study participants are available, our approach is language agnostic and we have begun initial investigations in English and Spanish, the two most frequently used languages in the U.S. We have discovered several useful features and we intend to provide the results of our efforts as a free online service.

Acknowledgments

Research reported in this publication was supported by the National Library Of Medicine of the National Institutes of Health under Award Number R01LM011975. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Nielsen-Bohlman, L. L. Institute of Medicine/Committee on Health. Health literacy: a prescription to end confusion. USA: National Academies Press; 2004.
2. Yan X, Song D, Li X. Concept-based Document Readability in Domain Specific Information Retrieval. Proceedings of Information and Knowledge Management. 2006
3. Vernon, J., et al. Low Health Literacy: Implications for National Health Policy. National Bureau of Economic Research; Storrs, CT: 2007.
4. Wang LW, et al. Assessing Readability Formula Differences with Written Health Information Materials: Application, Results, and Recommendations. Research in Social & Administrative Pharmacy. 2012
5. Forsyth RA, et al. The Iowa Tests of Educational Development (ITED): Guide to Research and Development. 2003
6. Bruce B, Rubin A, Starr K. Why Readability Formulas Fail. IEEE Transactions on Professional Communication. 1981
7. Kincaid JP, et al. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. 1975 DTIC Document.
8. McLaughlin GH. SMOG grading: A new readability formula. Journal of reading. 1969
9. Chharnock D, et al. DISCERN: an instrument for judging the quality of written consumer health information on treatment choices. Epidemiology and Community Health. 1999
10. Collins-Thompson K. Computational assessment of text readability: A survey of current and future research. International Journal of Applied Linguistics. 2014
11. Kim, H., et al. AMIA Annual Symposium Proceedings. American Medical Informatics Association; 2007. Beyond surface characteristics: a new health text-specific readability measurement.

12. Brants T, Franz A. The Google Web 1T 5-gram corpus version 1.1. 2006 LDC2006T13.
13. Liang, FM. Word Hy-phen-a-tion by Com-put-er. Citeseer; 1983.
14. Leroy G, Kauchak D. The effect of word familiarity on actual and perceived text difficulty. Journal of the American Medical Informatics Association. 2013
15. Buhrmester M, Kwang T, Gosling SD. Amazon's Mechanical Turk A New Source of Inexpensive, Yet High-Quality, Data? Perspectives on Psychological Science. 2011
16. Leroy G, et al. User evaluation of the effects of a text simplification algorithm using term familiarity on perception, understanding, learning, and information retention. Journal of medical Internet research. 2013
17. Rodriguez E, Armenta B, Leroy G. Report: Determining Text Difficulty by Word Frequency and Parts-of-Speech Analysis in Health Information. Latin American Summer Research Program. 2014

Biographies

David Kauchak is an assistant professor at Pomona College. His main research interests are natural language processing, particularly with applications in text simplification and the health domain. He received his PhD in computer science from UC San Diego.

Gondy Leroy is an associate professor at University of Arizona. She received her PhD in Management Information Sciences and does work in number of application areas including natural language processing, text mining, medical informatics and information retrieval. She is an IEEE senior member. gondyleroy@email.arizona.edu

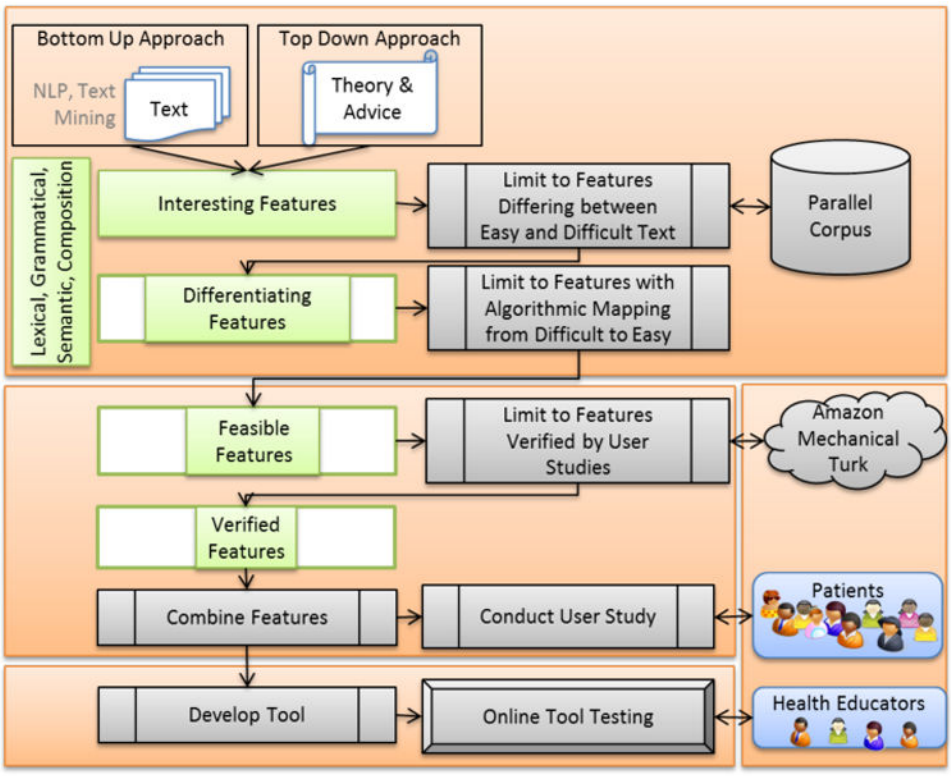

Figure 1. Development Flow diagram

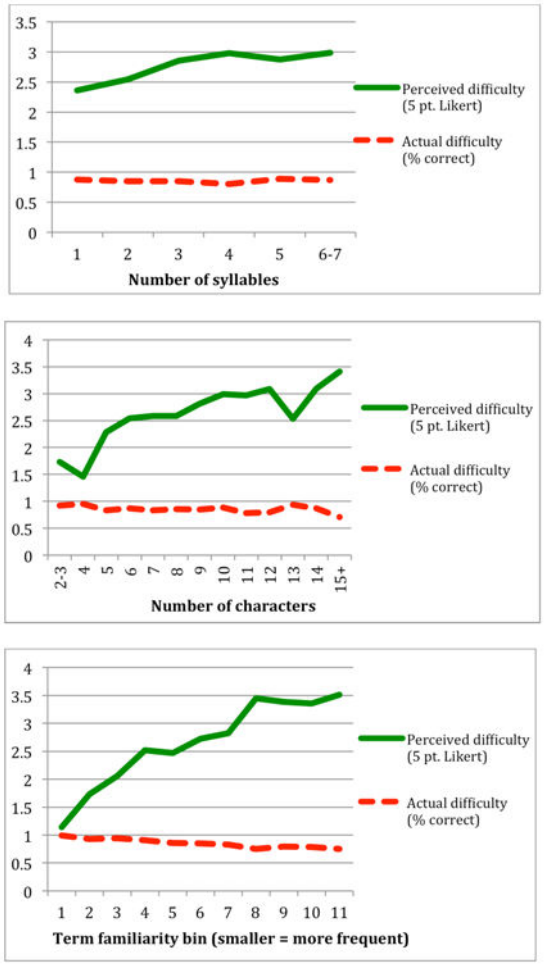

Figure 2.

Perceived and Actual difficulty aggregated by syllable count (top), number of characters (middle) and term familiarity bin (bottom) for 275 words randomly sampled from words from the Google Web Corpus.

Table 1

Feature Set Under Consideration (those already experimentally shown to be useful are highlighted in bold)

	Unit of Analysis		
Feature Group	**Word/Phrase/Constituent**	**Sentence**	**Paragraph/Page**
Lexical	**Term Familiarity**, Double Negatives	Lexical Density, Numerical Expressions Density	
Grammatical	**Compound vs. Periphrastic/ Prepositional Phrases**, Appositives	**Grammar Familiarity, Sentence Type**, Subject Placement, Relative Clauses; Extent & Recursivity of Subordination	Proportion telic/atelic verbs
Semantic	Tier 2 & Tier 3 Words, Concreteness	Semantic Patterns	Semantic Patterns, Tier 2 & Tier 3 Word Cues
Composition & Discourse			Structural Cues, Placement, Concept Hierarchy, **Connectors, Spatial Coherence**