Health Information Text Characteristics

[a]Gondy Leroy, PhD, [a]Evren Eryilmaz, [b]Benjamin T. Laroya, MSN, RN
[a]Claremont Graduate University, Claremont, California
[b]City of Hope National Medical Center, Duarte, California

Abstract

Millions of people search online for medical text, but these texts are often too complicated to understand. Readability evaluations are mostly based on surface metrics such as character or words counts and sentence syntax, but content is ignored. We compared four types of documents, easy and difficult WebMD documents, patient blogs, and patient educational material, for surface and content-based metrics. The documents differed significantly in reading grade levels and vocabulary used. WebMD pages with high readability also used terminology that was more consumer-friendly. Moreover, difficult documents are harder to understand due to their grammar and word choice and because they discuss more difficult topics. This indicates that we can simplify many documents by focusing on word choice in addition to sentence structure, however, for difficult documents this may be insufficient.

Keywords: Text Readability, UMLS, Consumer-Friendly Display (CFD) Names, Blogs, WebMD

INTRODUCTION

The estimates differ, but all surveys show that millions of people search online for health information. A Pew survey estimates that 80% of adult Internet users, about 93 million Americans, searched online for at least one of 16 major health topics[1]. Baker et al.[2] estimate that 20% of the US population uses the Internet to find health information (40% of those with Internet access). About one third report that this information affects decisions about their health or health care. Even though the information available online is important to millions of Americans, the text is often too difficult to understand[3-5]. English sites require at least a 10th grade reading level and more than half present information at college level. This is perhaps a partial explanation for the fact that Internet usage for health information is strongly associated with higher education[2,6]. Moreover, reading level will be especially important to people with limited cognitive skills, incomplete command of the English language, or those under stress. For example, Doak et al[7] found that patients, who may be more stressed, read on average five grades lower than the last year completed in school. How well information is understood and remembered has consequences for the patient-doctor relationship, such as the treatments requested or the perceived patient value from a doctor's visit. There are also consequences for health care at large. Misunderstandings in health information will increase the risk of making unwise health decisions, leading to poorer health and higher health care costs[8]. In contrast, consumers will benefit if the information is easier to find and understand. The knowledge gained empowers them to ask more informed questions when seeing their caregiver and it lessens their fear of the unknown[6].

Rewriting all existing texts in simpler language is infeasible. Even ensuring that all new text, when written, is sufficiently simple will not be an easy matter. In answer to this problem, Soergel et al.[9] propose a framework that includes an interpretative layer. Our goal is to extend and specify their framework. Figure 1 shows our approach.

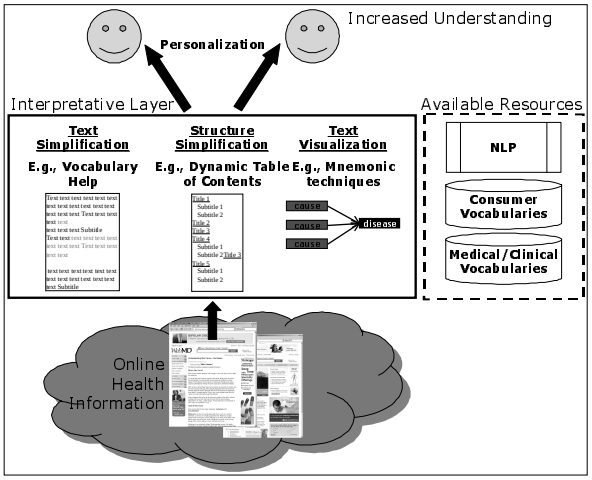

Figure 1: Detailed Framework for Health Information Simplification: Text Simplification, Structure Simplification, and Text Visualization

There are three different groups of techniques that need to be researched for their potential to contribute to text understanding without rewriting the text. Each has the potential to be automated. The first group focuses on the text itself: text simplification. The vocabulary used or the sentence syntax may contribute to increased difficulty. Some texts may lend themselves to being changed automatically to a simpler version, e.g., by changing from passive to active voice. The second is structure simplification to provide additional structure that augments the content. Adding additional labels, such as

'diagnosis', 'causes', or 'medication', may increase comprehension. The third group in our interpretative layer is text visualization. Some elements may be visualized in a way that contributes to understanding.

We focus here on the first component: the text itself. We look at characteristics that distinguish consumer/lay language from professional language and evaluate if existing, open source resources can help pinpoint what makes a text difficult.

METHODS

Most approaches use Flesch readability formulas to evaluate text. These formulas use syntax, word counts, and word length to assign readability levels. We chose not to use SMOG[10], a popular metric, since it is based entirely on syllable count. We will add comparisons later. Freda[11], however, found SMOG to assign reading levels 2 or 3 grades higher than Flesch. Our goal is to compare documents that are considered easy or difficult, according to these statistics, with patient educational materials and online blogs. We hypothesize that other factors besides syntax play a role in making a text accessible. If this is the case, then we will find variables, other than the syntax-based variables, that are significantly different between easy and difficult WebMD pages, patient educational material, and blogs. We evaluate both surface (word counts and syntax-related metrics) and content metrics (related to word choices).

A Health Information Documents

We collected four sets of documents. Three sets characterize texts provided for consumers and one set consists of text provided by consumers. The first two sets were documents we selected from WebMD (www.webmd.com) with high or low readability scores but similar lengths. We chose WebMD because it is one of the most popular consumer health web sites according to Alexa Web Search (www.alexa.com). The readability and grade metrics are based on the Flesch Reading Ease formula and the Flesch-Kincaid Grade Level assessment. We used Microsoft Word to calculate the readability scores and defined *easy* as pages that had a grade level of 7th grade or below (WebMD-E in tables) and *difficult* as those that had a grade level of 11th or above (WebMD-D in tables). It is noteworthy that it was difficult to find 50 easy WebMD documents but not difficult to find the 50 difficult WebMD documents.

Our third document set consists of 50 documents from patient educational materials provided by City of Hope (Patient Educ. in tables). City of Hope is a comprehensive cancer center in Duarte, California. The content of many documents is developed in house in collaboration with other healthcare professionals such as MDs, nurses, and pharmacists and based on many of the principles discussed in Doak et al. [7]. When working up the drafts, the target audience, age group, and the Flesch reading ease scores are kept in mind. Layout and design are also considered. Use of headings, bullet points, short sentences, and layman's language, as well as font style and size and use of white space is incorporated when possible. The goal is to provide materials at a 6th grade reading level: however, due to the sometimes technical nature and specialized vocabulary of the materials, the reading grade level appears artificially higher when using the Flesch measures. Additional documents are available online and patients are referred to specific sites at the National Institutes of Health. The materials are pilot tested with lay people and their feedback is used to revise content that may need clarification before final distribution.

The final set of documents consists of 50 blogs written by patients. We collected parts of the blogs that described diseases, conditions, or treatments. Such blogs are available from WebMD and other major blog sites such as www.blogger.com. We used keywords such as treatment, hospital, or several disease and cancer names to find blogs. One entry in a blog is taken as one 'blog' in our analyses. We did not include very short blog entries of only two or three sentences.

B Natural Language Resources

To evaluate the language used in the different types of documents we compare them with three existing vocabularies that are more or less medical in nature. The first vocabulary is the Metathesaurus included in the Unified Medical Language System (UMLS). We used the 2005AB version as one vocabulary to represent different specialties in medicine and included all vocabularies that do not need an extra license. Our set contains 1,570,372 terms mapped to 840,605 concepts. This resource is the most intense medical resource we use.

The second vocabulary is the list of consumer-friendly display (CFD) names, developed by Zeng et al.[12], that is available online at www.consumerhealthvocab.org. The version used here contains 41,274 terms mapped to 9,546 concepts. Each term has been assigned a score indicating its understandability to lay people, which is based on frequency counts of the terms in large text corpora. Concepts have a similar score. All scores range from 0 to 1; a higher score means an easier to understand term. In addition, this resource indicates if a term is a preferred description for a particular concept by the CFD names and/or by the UMLS.

The last vocabulary resource is not medical in nature but consists of five word lists associated with reading levels at different grade levels[13]. Words that are repeated in later grades were deleted from our lists to ensure that each word is only represented once in the entire set. This leaves us with a unique set of words for each grade. There were respectively 339, 818, 801, 863, and 694 unique words in our lists for first, second, third, fourth, and fifth grade lists (G1, G2, G3, G4, and G5 in the tables). These lists can be used to create an indicator of the vocabulary level of a text.

C Natural Language Processing

We used the processing resources included with GATE[14], an open source toolkit for natural language processing (NLP), and adapted them where necessary. On each document we apply the tokenizer, sentence splitter, Hepple POS tagger [15], and a noun phraser. The tokenizer and sentence splitter were used without modification, but we adjusted the lexicon used by the POS Tagger. The original GATE lexicon contained 17,831 entries. We reformatted the UMLS Specialist Lexicon and combined it with the original GATE lexicon. For those words that appear in both lexicons, we used only GATE tags. This ensures that we use optimal tags for GATE's Hepple tagger. For example, "where" is tagged as possibly an adjective, conjunction, or noun in the UMLS Specialist Lexicon, but only as a wh-adverb in GATE. In addition, we tuned the lexicon in a few cases to increase tagging performance. For example, the tags for "cold" were swapped resulting in the entry cold NN JJ" instead of "cold JJ NN." For words such as "elderly"' we added a NN tag since it is often used as a noun in health information. The final lexicon contains 271,157 items. We developed our own noun phraser for this project with jape files, a GATE component.

We developed procedures to match noun phrases to the UMLS and CFD names. A noun phrase is first matched in its entirety. If there is no match and the noun phrase contains multiple words, we match sequentially smaller subphrases (head phrase matching) until a match is found or the final main noun has been tested. To match terms against the grade level vocabulary lists, we used word comparisons.

D Metrics

We distinguish between surface metrics based on character/word counts and syntax structure of the sentence and content metrics based on matches between the document and vocabularies. The surface metrics comprise the Flesch reading ease score and the Flesch grade level assessment. The content metrics comprise the percentage of noun phrases that can be found in the UMLS or in the CFD names, the average consumer score for terms and concepts found in the CFD names which indicates understandability, and the percentage of terms in the text that are preferred terms by the UMLS or CFD names.

RESULTS

We provide the means for all our metrics. We also performed an ANOVA for each metric with origin as the independent variable and the metric as the dependent variable. Due to space limitations, we report on a limited set of tested post-hoc contrasts. All contrasts are based on the Bonferroni test, which takes multiple comparisons into account when evaluating significance. In the tables, we include a * to indicate that the ANOVA showed a significant effect for origin at $p < .001$, ^ at $p < .05$. We include + when post-hoc contrasts are discussed

A Overview of Metrics

Table 1 shows the average length of the documents in the four groups. The patient educational materials are the longest documents, the blogs on average the shortest. Sentences are generally longer and contain more noun phrases in patient educational materials. The WebMD-D pages and patient educational materials have similar counts for words per sentence, but they differ in the number of noun phrases per sentence. In this case, the WebMD-D pages are more similar to blogs.

	Per Document			Per Sentence	
Origin	**S***	**W***	**NPs***	**W/S***	**NP/S***
WebMD-E	43	609	150	15	6
WebMD-D	31	611	168	21	4
Blogs	27	424	85	17	3
Patient Educ.	38	693	188	25	7

Table 1: Means for Basic Descriptors (S = Sentences, W = Words, NP = Noun Phrases)

Origin	**Readability Ease***	**Grade Level*+**
WebMD-E	71	7
WebMD-D	35	12
Blogs	72	7
Patient Educ.	51	9

Table 2: Means for Surface Descriptors

Table 2 show characteristics commonly used to distinguish between easy and difficult texts. As intended, the WebMD pages are either easy, on average 7th grade level, or difficult, on average 9th grade level. During collection of these pages, we found that is was especially difficult to find easy web pages that were longer than a paragraph. Blogs, on average, are written at a 7th grade level. The patient education material fell in between and was more difficult than a blog, but easier than the WebMD-D

pages. ANOVAs performed for both readability ease and grade level showed a significant difference for origin ($p < .001$). Post-hoc contrasts showed that all groups differed significantly for grade level ($p < .05$) with the exception of the blogs and WebMD-E pages.

Table 3 shows that both WebMD-E pages and blogs have the highest number of noun phrases in the UMLS. The patient educational material and WebMD-D pages had the lowest percentage. Post-hoc contrasts verified that blogs and WebMD-E pages did not differ from each other; neither did difficult WebMD-D pages and patient educational materials. However, the other contrasts were significant ($p < .05$). Similar differences existed for noun phrases found in the CFD names. Blogs and WebMD-E pages were similar; WebMD-D pages and patient education material were similar, but the other differences were significant ($p < .05$).

Origin	% NPs in UMLS*+	% NPs in CFD*+	CFD term score*+	CFD concept score*
WebMD-E	51	58	.70	.92
WebMD-D	40	43	.70	.87
Blogs	53	57	.73	.90
Patient Educ.	41	46	.68	.87

Table 3: Means for Content Descriptors

Term understandability scores (found in the CFD names) were the same for both WebMD-E and WebMD-D pages. However, blogs generally used words with higher understandability scores; patient education material generally used words with lower understandability scores. Only this contrast between the blogs and the patient educational material was significant ($p < .05$). The associated concept scores differed for the WebMD-E and WebMD-D pages. They were the same for WebMD-D pages and the patient educational material. Post-hoc contrasts showed WebMD-D pages to be significantly different from blogs and WebMD-E pages ($p < .05$). The patient education material was also different from blogs and WebMD-E pages ($p < .05$).

Table 4 shows the percentage of terms that are considered preferred terminology in the CFD names and in the UMLS. More than half of the terms in the WebMD-E pages and of the blogs are preferred terms in the CFD names. These numbers are lower for WebMD-D pages and patient educational materials. Post-hoc contrasts showed that WebMD-E and blogs did not significantly differ, nor did WebMD-D and patient educational materials. All other contrasts were significant. The percentage of terms considered preferred in the UMLS are surprisingly similar. Overall, fewer terms in each document set are preferred UMLS terms. The post-hoc contrasts showed WebMD-E pages and Blogs to be similar and so are WebMD-D pages and patient educational material. The other contrasts are significant.

Origin	% terms preferred by CFD*+	% terms preferred by UMLS *+
WebMD-E	51	33
WebMD-D	37	24
Blogs	54	32
Patient Educ.	40	26

Table 4: Means for Content Descriptors (cont.)

Table 5 shows the percentage of words in different grade levels. Almost one third of the words were found in the first grade level list, almost one fifth belonged to the second grade level. Less than ten percent belonged in the third or fourth grade level and less than one percent in the fifth grade level. The WebMD easy pages and the blogs showed the highest percentage of grade one words. The percentage grade four words was almost identical for WebMD easy pages, WebMD difficult pages, and patient education material but higher than that of blogs.

Origin	% words in G1*	% words in G2*	% words in G3*	% words in G4*	% words in G5^
WebMD-E	30	16	6	7	<1
WebMD-D	24	18	6	6	<1
Blogs	39	15	8	4	<1
Patient Educ.	25	13	6	7	<1

Table 5: Means for Content Descriptors (cont.)

B Correlations between Metrics

We calculated both the Pearson's correlation (linear) coefficient and Spearman's rho (also non-linear) for six variables: Flesch reading ease, Flesch grade level, percentage of noun phrases in the UMLS, percentage of noun phrases in the consumer-friendly terms, term scores, and concept scores. We calculated these correlations for each dataset per origin. Reading ease and grade level are strongly correlated in each dataset. Since they are based on the same principles we will not further mention them. The percentage of terms found in the UMLS and in the CFD names is also strongly correlated in each dataset and requires no further comment. In the blogs, an additional linear correlation of interest is that between the reading ease and the concept score in the CFD names ($r = .324$, $p < .05$). Text that is easier to read is also text about concepts that are easier to understand. This correlation was more strongly present in the patient educational material ($r = .713$, $p < .01$). Two non-linear correlations of interest in the blog dataset are between the average concept score and percentage of noun phrases found in the UMLS ($r = -.376$, $p < .01$) or in the CFD names ($r = -.329$, $p < .05$). Both relationships show that

increases in the percentage of terms found in the source vocabularies were associated with smaller increases but lower average concept scores.

CONCLUSION

We compared syntax- and content-based characteristics of four types of documents currently available on the Internet. Two groups of WebMD pages were selected based on reading grade levels. Easy WebMD pages of sufficient length were difficult to find. The easy WebMD pages are the most similar to patient blogs. As we expected, both the syntax and the vocabulary used differed between different groups. The terms and medical concepts that are discussed in blogs have higher understandability scores. This may be an indication that blogs and easy WebMD pages do not address the more difficult information presented in the other documents. All documents used more CFD names preferred terms than UMLS preferred terms. We would like to point out that using blogs might be a limitation. We may overestimate the reading and writing skills of average consumers. Bloggers enjoy writing; they may be more proficient at both reading and writing. We also used only two readability formulas. Several other formulas exist and a comprehensive comparison test with these would complete the picture.

In the future, we will try to automatically translate difficult documents in an easier format by optimizing word choice. In addition, it would be very helpful to have additional metrics that address the underlying content of a document and the amount of information being conveyed. Such an information metric could be based on entropy measures or expert opinion. Moreover, a metric based on consumer's opinions would complete the set. A collection of such metrics could form a nice indicator of documents: how difficult are they compared to the amount of information conveyed.

ACKNOWLEDGEMENT

The authors thank Annette Mercurio for her help in providing the patient educational materials. This work was funded by a grant from the National Library of Medicine, R21-LM008860-01.

REFERENCES

1. Fox S, Fallows D. Internet Health Resources. Washington D.C.: Pew Internet & American Life Project; 2003 July 16.
2. Baker L, Wagner TH, Signer S, Bundorf MK. Use of the Internet and E-mail for Health Care Information: Results from a National Survey. JAMA 2003;289(18):2400-2406.
3. 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;285:2612-2621.
4. 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;155:807-812.
5. 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.
6. Fox S, Fallows D. Internet Health Resources - Health searches and email have become more commonplace, but there is room for improvement in searches and overall Internet access. Washington D.C.: Pew Internet & American Life Project; 2003 July 16.
7. Doak CC, Doak LG, Root JH. Teaching Patients With Low Literacy Skills: Lippincott Williams & Wilkins; 1996.
8. 'Ad Hoc Committee on Health Literacy for the Council on Scientific Affairs AMA. Health Literacy: Report of the Council on Scientific Affairs,. JAMA 1999;281,: 552-557.
9. Soergel D, Tse T, Slaughter L. Helping Healthcare Consumers Understand: An "Interpretative Layer" for Finding and making Sense of Medical Information. In: Press I, editor. MEDINFO;San Francisco, USA; 2004. p. 931-935.
10. McLaughlin GH. SMOG Grading: a New Readability Formula. Journal of Reading 1969;12:636-646.
11. Freda MC. The Readability of American Academy of Pediatrics Patient Education Brochures. Journal of Pediatric Health Care 2005;19(3):151-156.
12. Zeng QT, Tse T, Crowell J, Divita G, Roth L, Browne AC. Identifying Consumer-Friendly Display (CFD) Names for Health Concepts. In: AMIA Fall Symposium; Washington DC, USA; 2005. p. 859-863.
13. TampaReads. National Reading Vocabulary Lists for Grades - 1 - 2 - 3 - 4 (and now) Grade 5. In; http://www.tampareads.com/trial/vocabulary/index-vocab.htm, 2006.
14. Cunningham H, Maynard D, Bontcheva K, Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Applications. In: ACL, editor. 40th Anniversary Meeting of the Association for Computational Linguistics; Philadelphia; 2002.
15. Hepple. M. Independence and commitment.' Assumptions for rapid training and execution of rule-based POS taggers. In: 38th Annual Meeting of the Association for Computational Linguistics; Hong Kong; 2000. p. 278-285.