Evaluating Online Health Information: Beyond Readability Formulas

Gondy Leroy, PhD[1], Stephen Helmreich, PhD[2], James R. Cowie, PhD[2], Trudi Miller[1], Wei Zheng[1]
[1]Claremont Graduate University, Claremont, CA; [2]New Mexico State University, Las Cruces, NM

Abstract

Although understanding health information is important, the texts provided are often difficult to understand. There are formulas to measure readability levels, but there is little understanding of how linguistic structures contribute to these difficulties. We are developing a toolkit of linguistic metrics that are validated with representative users and can be measured automatically. In this study, we provide an overview of our corpus and how readability differs by topic and source. We compare two documents for three groups of linguistic metrics. We report on a user study evaluating one of the differentiating metrics: the percentage of function words in a sentence. Our results show that this percentage correlates significantly with ease of understanding as indicated by users but not with the readability formula levels commonly used. Our study is the first to propose a user validated metric, different from readability formulas.

Introduction

The Internet provides hundreds, if not thousands, of sites dedicated to the provision of medical information to laypersons. In fact, some medical providers worry that the information may cause confusion and result in patients or caregivers ignoring some of the direct advice or instructions they have been given. Given that most patients or caregivers using the Internet are attempting to supplement the information given by a care provider (whose time is often limited), the enormous variety of available sources can be seen as truly useful. It is almost a reflex action to carry out a search or go to a trusted website to look for information. In fact, for anyone of college age, or younger, the Internet is the normal source of all "research" information.

Health information provides both a prologue and an epilogue to interactions with medical personnel. If sites are to be identified as "trusted" this must mean more than just having an official sanction from the care provider or her organization. It must mean that the information on the site is clear, free from errors and sources of confusion (a problem we do not address here), and, importantly, it must be easy to read and comprehend. The guidelines offered by Rudd[1] contain more than a dozen sources of information on how to develop materials for consumers with low literacy levels.

That people are able and willing to be their own health information providers is something to be encouraged. To do this the information must be available in forms that suit the readers and it must be clearly labeled with its provenance and its intended users. Professionals base their evaluation of text suitability often on the outcome of readability formulas. Such formulas provide a single approach to difficulty estimation and as a result they may under- or overestimate text difficulty. Our goal in the research project described here is to develop new, relevant linguistic metrics which can be measured automatically.

Evaluation of Online Health Information

Standard measures of readability based on syllable and word counts per sentence are insufficient to rate the difficulty of understanding health information texts[2]. These readability measures are used to assign grade reading levels to documents. However, these scores appear to be insufficient to categorize the difficulty of health related texts[3].

Studies such as those carried out by Rosemblat[4] explore other aspects that influence the difficulty of health information. The key factors related to difficulty seem to be the main point of the document and the difficulty of the vocabulary used. This research is based on the opinions of health literacy experts – "*All annotators held doctorates in mass or health communication*". The approach has its merits for the insights it provides, but in terms of evaluating the actual difficulty for the end consumers of the information it leaves much to be desired. The ideal measures of readability should be based on the assessments and performance of the actual users of the information. In earlier work, we found that experts and consumers provide different estimates for the same documents: our expert readily judged documents as too difficult, while the consumers considered them to be at a suitable level of difficulty[5].

Methodology

The metrics we test are complementary to the readability formulas commonly used and recommended. We report here on the first set of analyses in our effort to develop a comprehensive and balanced toolkit of metrics that will indicate how difficult a document will be to read and understand. One constraint we enforce is that it should be possible to measure the metric automatically. Metrics that require human expert evaluations are too time-consuming and too difficult to apply systematically across experts and documents.

This study contains three sections. We first calculated readability scores for all documents in our corpus. We then chose two documents with different readability levels and performed a detailed analysis of grammatical and semantic characteristics. Finally, we selected one metric and evaluated this with a user study.

Corpus Collection and Overview. We collected web pages discussing five common diseases: cancer, depression, diabetes, heart disease, and obesity. For each topic, we collected documents from six different sources. Four of these sources are composed of sites that provide information useful for lay people: websites with information on clinical trials such as ClinicalTrials.gov, consumer websites such as WebMD, government sponsored websites such as those from the National Cancer Institute, and hospital and doctor sponsored websites, such as the Jefferson University Hospitals website. We selected two additional sources to compare and contrast: Medline represents the professional literature while patient blogs represent consumer language.

Each document was scored using the Readability Analyzer developed at the National Library of Medicine[6]. The analyzer provides 5 numerical readability evaluations per document and also the average per document of all 5 numbers. We report here the Flesch-Kincaid Grade Level (from here on, referred to as "grade level") since this is the most common and well-known metric. The intuition behind this formula is that it represents the minimum schooling (grade) the reader should have completed to understand a document. The formula is based on counts such as the number of syllables per word and the number of words per sentence. As a note to readers: many published papers report Flesch-Kincaid readability scores calculated using Microsoft Word office software. However, that embedded algorithm had an upper limit of 12th grade. This has been corrected in the latest version of the software (MS Word 2007) and higher grade levels are now also reported.

Detailed Document Analysis. Two documents, one from a clinical trials site and one patient blog, were chosen for detailed grammatical and semantic analysis. These two documents represent the most difficult language and the easiest language that consumers will encounter and are expected to understand, that is, documents meant for them.

For our grammatical analysis, we evaluated the use of function words, negation, and noun phrases. We use a broad definition for function words and include all pronouns, modals, auxiliaries, prepositions, and determiners, almost all words that do not add direct medical content to the document. Our hypothesis is that a higher percentage of function words spaces out the content words, making them easier to assimilate. To gain a first indication of writing style, i.e. the type of language used, we also matched noun phrases to two existing, controlled vocabularies: the Unified Medical Language (UMLS) Metathesaurus and the Consumer Health Vocabulary (CHV)[7]. The UMLS Metathesaurus is made up of several contributing vocabularies. We used the 2007AA version, which has over 4.3 million phrases grouped by related medical concepts. The CHV[8] maps medical terms commonly used by consumers to the same concepts of the UMLS Metathesaurus. We used the November 2006 version, which has over 156,826 phrases. Both vocabularies can serve as indicators of writing style; the UMLS represents more formal, clinical phrases while the CHV provides more informal, non-professional phrases.

For our semantic analysis, we looked at how the content matches to UMLS Semantic Types. Each concept in the UMLS has semantic types associated with it. These semantic types can be seen as a higher level description of the content. As such, the semantic types can provide an indication of the diversity of topics discussed in each document. To accomplish this mapping, we mapped each noun phrase to the UMLS concepts and then the concepts to the UMLS semantic types. The matching algorithm[9] first tries to find the entire phrase in the UMLS Metathesaurus. If this is unsuccessful, it proceeds with matching the head phrase to concepts. If multiple matching semantic types are assigned to a concept, there are all retained.

Flesch-Kincaid Grade Level (Number of Documents)	Topic					
Source	**Cancer**	**Depression**	**Diabetes**	**Heart Disease**	**Obesity**	**Average**
Clinical Trials	13.9 (10)	17.3 (10)	16.1 (10)	17.7 10)	17.5 (10)	**16.5**
Consumer Sites	10.5 (20)	9.8 (20)	13.9 (10)	13.4 (10)	10.6 (10)	**11.2**
Government	12.9 (20)	14.3 (20)	15.1 (10)	10.9 (10)	11.3 (10)	**13.1**
Hospital, Doctor	14.7 (10)	15.5 (13)	15.7 (10)	12.3 (9)	11.8 (10)	**14.1**
Medline	17.6 (10)	17.6 (10)	18.3 (10)	18.2 (10)	18.0 (10)	**17.9**
Patient Blogs	10.5 (10)	9.3 (10)	9.7 (10)	11.6 (10)	7.5 (11)	**9.7**
Average	**12.9**	**13.5**	**14.8**	**14.0**	**12.7**	

Table 1. Flesch-Kincaid Readability Grade level. (N: number of documents)

User Readability Evaluation. We tested the influence of the amount of function words, as a metric, with representative users. To evaluate its potential impact, we selected the first sentence in a clinical trials document and constructed the following five versions:

- We ask patients eligible to participate in this study to consider a research consent form which includes the following information: ... (40% function words)
- Patients eligible to participate in this study will be asked to consider a research consent form which includes the following information: ... (43% function words, original sentence)
- We ask patients who are eligible to participate in this study to consider a research consent form which includes the following information: ... (45% function words)
- We ask those patients who are eligible to participate in this study to consider a research consent form which includes the following information: ... (48% function words)
- We ask those patients who are eligible to participate in this study to consider a research consent form with the following information: ... (50% functions words)

We randomized the sentence order, deleted the information on the percentage of function words, and explained to users that these were multiple versions of the same sentence from a clinical trial website. The user's task is to rate each of these 5 sentences using the following 4-point scale:

- Score 1 = very easy
- Score 2 = easy
- Score 3 = difficult
- Score 4 = very difficult

Results

Corpus Collection and Overview. Table 1 shows an overview of the corpus and grade levels. Our goal was to have about 10 documents for each case but more documents were added when this resulted in a better representation of the available information for that category. Documents, including blogs, had to be about one page long (to avoid documents that were extremely short or long).

We found a clear difference in the grade levels for different types of documents and for different topics. On average, an 18^{th} grade level is required to understand Medline documents. This is expected since these documents are meant for medical professionals. However, with the exception of the patients' own writings (blogs), all others require a grade level that is above 11^{th} grade. Clinical trials text was written at a 16.5^{th} grade level. This overview also shows that different topics require different grade levels. Text on diabetes required the highest grade levels (almost 15) while obesity required the lowest (almost 13).

Detailed Document Analysis. A clinical trials document and patient blog discussing cancer were chosen. They had approximately the same length: 383 words in the clinical trials document and 403 words in the patient blog. The readability grade levels differed enormously: the clinical trials document scored almost twice as high (14 grade level) compared to the patient blog (7.5 grade level). Both documents contained similar numbers of negation: 5 negations in the clinical trials document and 4 in the patient blog.

The grammatical analysis showed an unexpected difference in the use of function words. In the patient blog, 64% of the words were function words. In the clinical trials document, 45% of the words were function words, which is almost 20% lower. Table 2 shows additional results from the writing style analysis. The percentage of complete noun phrases matched to the CHV or UMLS is very similar for both types of documents. However, matching levels differ for partial, head phrases: more matches to the CHV were found for the patient blog than for the clinical trials document. The trend is reversed for partial matching to the UMLS.

Table 3 shows results for the semantic analysis. More topics are discussed in the clinical trials document. From the Clinical Trials document, we could match the noun phrases to 27 unique semantic types, from the Patient Blog we could match to 19 unique semantic types.

	Clinical Trials		Patient Blog	
	Total NPs	Unique NPs	Total NPs	Unique NPs
Count (100%)	87	61	49	41
Percentage (%) of Phrases Found in CHV				
Complete phrase	55	51	57	49
Head phrase	25	31	31	37
None	20	18	12	15
Percentage (%) of Phrases Found in UMLS				
Complete phrase	57	59	61	56
Head phrase	33	38	22	26
None	9	3	16	17

Table 2. Matching to controlled vocabularies.

Clinical Trials		Patient Blog	
Semantic Types	**Freq.**	**Semantic Types**	**Freq**
Functional Concept	12	Intellectual Product	6
Research Activity	9	Biomedical Occupation or Discipline	5
Biomedical Occupation or Discipline	7	Disease or Syndrome	5
Daily or Recreational Activity	7	Temporal Concept	5
Qualitative Concept	5	Patient or Disabled Group	3
Neoplastic Process	4	Body Part, Organ, or Organ Component	2
Organic Chemical	4	Finding	2
Substance	4	Functional Concept	2
Finding	3	Idea or Concept	2
Idea or Concept	3	Professional or Occupational Group	2
Professional or Occupational Group	3	Daily or Recreational Activity	1
Amino Acid, Peptide, or Protein	2	Neoplastic Process;	1
Biomedical or Dental Material	2	Family Group; Mental or Behavioral	1
Human	2	Dysfunction; Organism Attribute; Physiologic	
Patient or Disabled Group	2	Function; Population Group; Spatial Concept;	
Pharmacologic Substance	2	Therapeutic or Preventive Procedure	
Quantitative Concept	2		
Animal; Cell; Chemical Viewed	1		
Functionally; Clinical Attribute; Conceptual			
Entity; Eicosanoid; Gene or Genome; Health			
Care Related Organization; Indicator,			
Reagent, or Diagnostic Aid; Manufactured			
Object; Occupational Activity			
Temporal Concept	1		

Table 3. Matching to UMLS Semantic Types (overlap is shown by shading).

User Readability Evaluation. Ten users participated in the study. They were adults between 21 and 55 years old. Each user evaluated each sentence. Table 4 shows the average scores for each sentence and the Flesch-Kincaid grade level. We calculated the Pearson Correlation coefficient and found a significant, negative correlation between user ratings and percentage function words: -0.960 ($p < .01$)

indicating that a higher number of function words leads to easier sentences. The user ratings also correlated with the Flesch-Kincaid readability grade level: 0.892 ($p < .05$) indicating that a lower grade levels is associated with easier to read sentences. It stands out, however, that there is no significant correlation ($p = .065$) between the Flesch-Kincaid grade levels and the percentage of function words.

	Percentage function words				
	40%	43%	45%	48%	50%
Average user readability score (1-4 scale, 1 is easiest)	2.6	2.2	1.9	1.9	1.5
Flesch-Kincaid grade level	14.6	14.5	14.4	14.4	13.9

Table 4. User and grade levels scores

Conclusion

Our goal is to develop a toolkit of metrics that indicates how difficult documents are for average laypersons. We evaluate metrics that are more precise than readability scores, can be automatically calculated, and are validated with user studies.

We compared documents from different sources discussing different topics. Some documents such as clinical trial information, intended to be read by laypersons, were written at a very high grade level. Additionally, different topics also led to different readability scores: documents on obesity were the easiest, those on diabetes the most difficult. Patient blogs differed both in content and language use. The patient blogs contained fewer topics and especially fewer clinical topics. Patients also used less formal vocabulary. Moreover, patient blogs displayed a much higher use of function words. In a subsequent user study, we manipulated the amount of function words in a sentence and found that this amount correlated with perceived ease of understanding but not with the readability grade levels.

In the future, we aim to combine our metrics with existing ones and provide laypersons and professionals an easy-to-use software toolkit that shows the difficulty levels of text according to specific indicators.

Acknowledgements

The authors would like to thank Graciela Rosemblat for running the NLM Readability Analyzer on our document collection.

This work was made possible by a grant from the National Science Foundation (NSF), 0742223, "U3 – Understanding User Understanding".

References

1. RE R. How to create and assess print materials. *Harvard School of Public Health: Health Literacy Website. 2005. Available at: http://www.hsph.harvard.edu/ healthliteracy/materials.html.* Accessed [2007].
2. Kim H, Goryachev S, Rosemblat G, Browne A, Keselman A, Zeng-Treitler Q. Beyond Surface Characteristics: A New Health Text-Specific Readability Measurement. *AMIA*. Washington DC; 2007.
3. Zeng-Treitler Q, Kim H, Goryachev S, Keselman A, Slaughter L, Smith C. Text Characteristics of Clinical Reports and Their Implications on the Readability of Personal Health Records. *Medinfo*; 2007.
4. Rosemblat G, Logan R, Tse T, Graham L. Text Features and Readability: Expert Evaluation of Consumer Health Text. *Mednet 2006: 11th World Congress on Internet in Medicine the Society for Internet in Medicine*. Toronto, Canada.; 2006.
5. Leroy G, Miller T, Rosemblat G, Browne A. A Balanced Approach to Health Information Evaluation: A Vocabulary-based Naïve Bayes Classifier and Readability Formulas. *Journal of the American Society for Information Science and Technology.* 2008, 59(9): 1409-1419.
6. Gemoets D, Rosemblat G, Tse T, Logan R. Assessing readability of consumer health information: an exploratory study. Paper presented at: Medinfo, 2004.
7. Zeng QT, Tse T. Exploring and Developing Consumer Health Vocabularies. *Journal of the American Medical Informatics Association.* February 2006;13(1):24-29.
8. *Open-Access and Collaborative Consumer Health Vocabulary (OAC CHV)* [computer program]. Version; 2007.
9. Miller T, Leroy G. Dynamic Generation of a Health Topics Overview from Consumer Health Information Documents. *International Journal of Biomedical Engineering and Technology.* Forthcoming 2008.