Published in final edited form as:

Digit Libraries Cult Herit Knowl Dissem Future Creat (2011). 2011 October ; 7008: 307–310. doi: 10.1007/978-3-642-24826-9_38.

Term Familiarity to indicate Perceived and Actual Difficulty of Text in Medical Digital Libraries

Gondy Leroy and **James E. Endicott**

School of Information Systems and Technology, Claremont Graduate University

Gondy Leroy: Gondy.Leroy@cgu.edu; James E. Endicott: EndicotJ@cgu.edu

Abstract

With increasing text digitization, digital libraries can personalize materials for individuals with different education levels and language skills. To this end, documents need meta-information describing their difficulty level. Previous attempts at such labeling used readability formulas but the formulas have not been validated with modern texts and their outcome is seldom associated with actual difficulty. We focus on medical texts and are developing new, evidence-based meta-tags that are associated with perceived and actual text difficulty. This work describes a first tag, term familiarity , which is based on term frequency in the Google corpus. We evaluated its feasibility to serve as a tag by looking at a document corpus (N=1,073) and found that terms in blogs or journal articles displayed unexpected but significantly different scores. Term familiarity was then applied to texts and results from a previous user study (N=86) and could better explain differences for perceived and actual difficulty.

Keywords

Natural Language Processing; Health Informatics; Perceived Difficulty; Actual Difficulty; Meta Information; Lexical Tags

1 Introduction

In healthcare, pamphlets and brochures are often used to explain medical procedures, effects and guidelines. The text needs to be written in a manner that consumers, who may have low literacy, can understand. Today s readability formulas could form an excellent starting point to facilitate writing easier text. Unfortunately, they suffer from several shortcomings. They have not been validated for modern texts, topics or sources. Moreover, they have not been clearly demonstrated to be related to text understanding. Many studies rely on expert opinions instead of measured user understanding. We are working on an evidence-based difficulty checker that indicates where text may be too difficult for consumers to understand.

Difficulty levels of text are often measured with readability formulas. While many different formulas exist, the Flesch-Kincaid readability grade level is among the most popular in the U.S. It is based on syllable and word counts and its outcome represents the education needed to understand the text (e.g., 12.0 = finished 11^{th} grade and started 12^{th} grade). Most writing guidelines, e.g., by the American Medical Association [1], advise writing at 6–8th grade level. However, based on these formulas most English-language Internet sites require a 10th

grade or even college level education [2]. There are readability formulas that leverage lists of common words (e.g., Dale-Chall) but they used fixed word lists that are evenly weighted.

We distinguish between perceived and actual difficulty because both are useful, but for different reasons. It is important for text to have an easy perceived difficulty so it is not intimidating. Actual difficulty level is important to facilitate understanding of the material. The distinction between perceived and actual difficulty is important but seldom acknowledged. However, evidence for the distinction can be found in different models. For example, the Health Belief Model contains 4 dimensions: perceived susceptibility, perceived severity, perceived benefits, and perceived barriers. In a large review study [3], perceived barriers was found to be the most significant dimension in explaining health behaviour.

2 Term Familiarity

We are working on defining text features, to label text difficulty levels, in an evidence-based and computationally efficient manner. We focus on 3 groups: grammatical, lexical, and composition features. In earlier work [4], we found several grammatical metrics (noun phrase complexity, function word density, and grammatical structures) that are strongly related to perceived text difficulty. However, these were only slightly related to actual difficulty as measured by understanding and retention of information.

This work focuses on a lexical measure: term familiarity. We use the frequency of terms in the Google corpus as an approximation of term familiarity. The Google web corpus contains n-gram models calculated based on a corpus of a trillion words from the Google collection of public Web pages. It contains 1,024,908,267,229 tokens and provides frequency counts for unigrams, bigrams, trigrams, 4-grams and 5-grams. For example, the corpus shows that the word “apnea” is less common than “obesity”. We believe such frequencies may help explain difficulty where readability formulas would fail, e.g., most formulas would judge “obesity” as the more difficult term based on syllable or letter counts.

2.1 Corpus Analysis - Feature Detection

Study Design—A corpus was developed containing patient blogs from Blogspot (up to 6 posts per blogger) representing easy text and full-text journal articles from Wiley Online Library (top 100 matches) representing difficult text. For each, we selected documents on 7 topics: alcoholism, autism, cancer, depression, diabetes, obesity and post-traumatic stress disorder.

We calculated the ratio of terms in each document that were nouns, verbs, adjectives, and adverbs as these are the main content bearing terms. Each term was submitted to the Google corpus and its frequency was noted. We then calculated our term familiarity measure for each document as the weighted average of term frequencies adjusted for their relative proportion in the document.

Analysis and Results—In total, 372 blogs and 697 journal articles were found. Difficulty in finding patient blogs without over-representing a single author is the reason

that fewer blogs were selected for each topic. Naturally, journal articles are much longer than blogs and therefore all measures are normalized by word length.

Table 1 provides an overview of the results. It is interesting that the proportion of nouns is significantly higher in the articles, while the proportion of verbs is significantly higher in the patient blogs. The term frequencies also differ, with patient blogs using nouns and adjectives that have significantly higher frequencies. The frequency of the verbs did not differ between the two groups. Our familiarity measure, which takes the ratio of each word group into account when combining frequencies per document, is significantly higher for patient blogs than for journal articles.

2.2 User Study - Feature Application

Study Design—In a previous user study [4], we evaluated 8 sentences (Part 1) and 2 paragraphs (Part 2) for 3 features: noun phrase complexity (the simple version had fewer compound nouns), sentence structure (the simple version had shorter and active sentences), and function word density (the simpler version had more function words).

- Part 1: For each sentence, all 8 versions were shown and participants chose the sentence that looked the easiest.
- Part 2: We worked with 1 paragraph on depression and 1 on heart disease and constructed an easy and difficult version for each. Each participant received an easy and a difficult paragraph with the topics and order counterbalanced. To measure perceived difficulty we asked the participants to indicate from which paragraph they thought they remembered most. To measure actual difficulty, we posed 3 multiple-choice questions for each paragraph.

Analysis and Results (N=86)—As reported earlier [4], we found very clear effects of our grammatical metrics on perceived difficulty for the individual sentences (Part 1). The results for the paragraphs (Part 2) could not all be explained by the differences in grammatical features. However, Table 2 shows how these results also are in line with term familiarity. The first unexplained result was that 62% of all participants chose the depression document as the easiest (perceived difficulty) of the two documents they read, even though half of the participants received its difficult version. However, term familiarity was higher in both version of the depression document. The second unexplained result was that the percentage correct answers (actual difficulty) was only slightly influenced by surface metrics. Again, these results are in line with term familiarity for the different documents. For both results, higher familiarity is in line with documents perceived as easier and more correct answers.

Follow-up studies will be conducted to collect more data to verify or refute these relationships statistically.

3 Conclusion

We are working toward the development of meta-information that indicates text difficulty. This work introduced a composite feature, term familiarity, which can be assigned to text by computational means. From a corpus analysis, we can conclude it distinguishes between easy and difficult text. It also served to clarified hereto unexplained results of a user study for perceived and actual difficulty of text.

Later work will use larger samples sizes to statistically verify or refute these findings. Term familiarity represents only one aspect of text and we plan to combine it with others for a more comprehensive measurement of text difficulty.

Acknowledgments

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

References

1. Weis, BD. Manual for Clinicians. 2. AMA and AMA Foundation; 2007. Health Literacy and Patient Safety: Help Patients Understand.
2. Berland GK, Elliott MN, Morales LS, Algazy JI, Kravitz RL, Broder MS, Kanouse DE, Muñoz JA, Puyol JA, Lara M, Watkins KE, Yang H, McGlynn EA. Health Information on the Internet: Accessibility, Quality, and Readability in English and Spanish. JAMA. 2001; 285:2612–2621. [PubMed: 11368735]
3. Janz NK, Becker MH. The Health Belief Model: A Decade Later. Health Education Quarterly. 1984; 11:1–47. [PubMed: 6392204]
4. Leroy G, Helmreich S, Cowie J. The Influence of Text Characteristics on Perceived and Actual Difficulty of Health Information. International Journal of Medical Informatics. 2010; 79:438–449. [PubMed: 20202895]

Table 1

Corpus Descriptive Statistics

Feature	Patient blog	Journal Article
Average Total Word Count:	726	5158
Average Percentage Nouns*+:	25	39
Average Percentage Verbs*+:	19	14
Average Percentage Adjectives*+:	7	10
Average Percentage Adverbs*+:	6	4
Average Noun Frequency*+:	131,114,039	77,111,994
Average Verb Frequency:	718,591,781	702,892,877
Average Adjective Frequency*+:	203,763,085	104,070,055
Average Adverb Frequency*+:	508,593,450	588,474,771
Term Familiarity*:	215,181,482	159,667,621

*significant at $p < .001$, independent samples two-tailed t-test,

+with Bonferroni correction $\alpha/4$.

Table 2

Actual difficulty in relation to grammar and lexical features.

Topic:	Depression		Heart Disease	
Difficulty Level:	**Easy**	**Difficult**	**Easy**	**Difficult**
Actual Difficulty (% correct):	67	57	57	54
Term Familiarity:	303,670,073	296,258,700	229,500,975	158,856,104