The effect of word familiarity on actual and perceived text difficulty

Gondy Leroy,† David Kauchak‡

ABSTRACT
There is little evidence that readability formula outcomes relate to text understanding. The potential cause may lie in their strong reliance on word and sentence length. We evaluated word familiarity rather than word length as a stand-in for word difficulty. Word familiarity represents how well known a word is, and is estimated using word frequency in a large text corpus, in this work the Google web corpus. We conducted a study with 239 people, who provided 50 evaluations for each of 275 words. Our study is the first study to focus on actual difficulty, measured with a multiple-choice task, in addition to perceived difficulty, measured with a Likert scale. Actual difficulty was correlated with word familiarity (r=0.219, p<0.001) but not with word length (r=−0.075, p=0.107). Perceived difficulty was correlated with both word familiarity (r=−0.397, p<0.001) and word length (r=0.254, p<0.001).

INTRODUCTION
There exist many different readability formulae, some of which were conceived years ago. Their continued popularity now is a testimony to the need for an efficient means to evaluate the difficulty of text for patients and consumers. Several formulae are provided in text editing software (eg, in Microsoft Word) or made available in online tools (eg, http://www.readabilityformulas.com). They are used in numerous research projects and are recommended to help simplify text for medical information.¹

Even though they are extremely popular, there is little evidence that simplifying text using these formulae is associated with increased understanding.² Exceptions are Swanson and Fox,³ who in 1953 found that simpler articles, according to the formulae, led to higher understanding but not higher retention, and Fred et al,⁴ who found higher recognition memory for experimental text with a decreased readability grade level but also more focused with crucial information in tables. However, neither study focused on using readability measures for text simplification. Work showing direct application of formulae leading to increased understanding, retention or learning is rare, possibly due to the difficulty of publishing non-significant results, especially for such a popular and accepted tool. In addition, studies often do not differentiate between actual and perceived difficulty, a distinction supported by both the health belief model (HBM)⁵ and the theory of planned behavior (TPB).⁶ In a review of 24 studies, the fourth dimension in the HBM, perceived barriers, was shown to be the most significant in explaining health behavior.⁵ Similarly, in TPB, perceived difficulty, a factor of perceived behavioral control, has been found to be the stronger predictor of intentions and behavior.⁶ We are working towards providing a modern approach to text simplification with demonstrated impact. In this paper we address one component used by most existing readability formulae: the difficulty of individual words. Currently, word length is used as a stand-in for word difficulty and is measured in characters or syllables (eg, in SMOG, Linsear, Lix, Coleman–Liau, Flesch grade level readability formulae).⁷ However, examples demonstrate that this is not always an accurate indicator of word difficulty: ‘disorientation’ or ‘diabetes’ would be considered more difficult than ‘apnea’ by most formulae, but in many cases people know the meaning of the first words but not the last.

We propose a new measure for evaluating word difficulty based on word familiarity. Familiarity can be practically estimated by the frequency with which a word occurs in a large corpus of English text. Words with a low occurrence frequency are assumed to be less familiar and therefore more difficult because a reader will not encounter them as often and is less likely to know their meaning. Similarly, text that uses more low frequency words can be expected to be more difficult. In earlier work we have seen indirect evidence of this relationship and found that easy texts used more words with higher word frequencies.⁸ ⁹ ¹⁰ We also saw a positive effect on understanding and learning when words with low familiarity (ie, low frequency) were replaced with high frequency equivalents.¹⁰ ¹¹ In contrast, Ryder and Hughes¹² did not find such an effect in their studies with high school students.

To our knowledge, no one has directly measured the difficulty for individual readers of a large set of words with different occurrence frequencies in English texts. Given the lack of clear data, we conducted a user study to evaluate the impact of word familiarity directly on word difficulty. The dataset will be made available to the community.

DATASET CONSTRUCTION

Word set
To generate a representative sample of words with different frequencies, we combined two resources: the Google web corpus¹³ and the Moby word list. The Google web corpus contains n-gram counts from a corpus of a trillion words from public webpages. It is made available by the Linguistic Data Consortium (http://www.ldc.upenn.edu/) for a small fee. There are 13 588 391 unigrams (single words) in the corpus, each with their frequency count. The Moby word list is a list of common English words and their definitions and is made available for free at Infochimps (http://

---

www.infochimps.com/collections/moby-project-word-lists). We used the set containing 64,000 common English dictionary words.

To select a subset of words with sufficiently different word frequencies, we identified the top 1% most frequent words (1st percentile) in the Google web corpus, the 9–10% most frequent words (10th percentile), the 19–20% most frequent words (20th percentile) and so on until the 99–100% most frequent words (100th percentile). From each percentile we randomly selected 25 words for which there is also a definition available in the Moby word list. We excluded words that are formulae, html or internet-specific syntax, or number–letter combinations. We reviewed the word list to exclude proper names, for example, the word ‘Mendelsohn’ was excluded because it is the name of a German musician. Table 1 provides an overview of our word set characteristics. For example, ‘work’, ‘management’ and ‘power’ are included in the 1st percentile, ‘rancor’, ‘furin’ and ‘shorebird’ in the 50th percentile, and ‘shaggymane’, ‘drop-kicker’ and ‘hiplength’ in the 100th percentile.

For each word, we selected the most common meaning and then shortened the definition by removing usage examples. If the word itself appeared in its definition, the definition was rephrased based on WordNet, an online (http://wordnet.princeton.edu/) lexical dictionary.

For example, for ‘immorally’ (70th percentile), the original definition was

without regard for morality; “he acted immorally when his own interests were at stake”

which we shortened and rephrased to

without regard for traditionally held principles.

The final set consisted of 275 words, each with its correct definition.

Evaluating word difficulty
We examined two different aspects of word difficulty: the perceived and actual difficulty. To measure how difficult the words actually are, that is how many words are known to participants, we asked participants to choose the correct definition from four options. For each word, the four definition options were constructed by including the correct definition along with three randomly selected definitions from one of the other words in the set. To measure how difficult a word was perceived to be, we asked participants to judge its difficulty by indicating the difficulty level on a five-point Likert scale, with one indicating a very easy word and five a very difficult word. The specific question asked was: ‘How difficult would this word look in a text given to patients?’

Table 1 Word set characteristics

<table>
<thead>
<tr>
<th>Word frequency (familiarity)</th>
<th>Word length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word frequency (familiarity)</td>
<td>Minimum</td>
</tr>
<tr>
<td>1st Percentile</td>
<td>216,988,964</td>
</tr>
<tr>
<td>10th Percentile</td>
<td>6,008,690</td>
</tr>
<tr>
<td>20th Percentile</td>
<td>1,420,722</td>
</tr>
<tr>
<td>30th Percentile</td>
<td>523,999</td>
</tr>
<tr>
<td>40th Percentile</td>
<td>226,266</td>
</tr>
<tr>
<td>50th Percentile</td>
<td>101,654</td>
</tr>
<tr>
<td>60th Percentile</td>
<td>44,554</td>
</tr>
<tr>
<td>70th Percentile</td>
<td>18,051</td>
</tr>
<tr>
<td>80th Percentile</td>
<td>65,451</td>
</tr>
<tr>
<td>90th Percentile</td>
<td>1,902</td>
</tr>
<tr>
<td>100th Percentile</td>
<td>252</td>
</tr>
<tr>
<td>Overall</td>
<td>252</td>
</tr>
</tbody>
</table>

Data collection
We recruited participants using Amazon’s Mechanical Turk (MTurk, http://www.mturk.com/), an online crowd sourcing service where workers select small tasks to work on. Currently, there are over half a million workers and over 300,000 available tasks. MTurk has been used in a wide range of settings and attracts workers from all over the world with varied demographic characteristics. With precautions taken to filter out unproductive workers, the data quality is at least as good as that from more traditional approaches.

MTurk workers select human intelligence tasks (HIT) based on their title, description, payment and/or personal interest. Only workers located in the USA with a 95% or better HIT performance rating were invited to our HIT. We required workers first to answer common demographic questions as well as their use of English at home. Then, they were asked to evaluate different words. Workers were paid 2 cents for each word they evaluated.

To get a reasonable sample, we designed our study to collect evaluations from 50 participants for each word. As is customary on MTurk, workers were at liberty to evaluate as many or as few words as they chose. For each word, participants labeled the perceived difficulty on a five-point Likert scale and selected their best guess for the definition from the four options. The order of definitions was randomized per word and the order of the words was randomized. MTurk ensured that no worker attempted the same word twice and that 50 different workers evaluated each word.

EVALUATION
This large dataset allows us to calculate correlations analysis for individual data points (N=13,750), per word (N=275) and per percentile (N=11) for both actual and perceived difficulty. For brevity, we include only the analysis per word in which the 50 different evaluations are averaged per word.

Demographic information and metadata
A total of 239 workers participated in the data collection (see table 2). Slightly more than half were men (59%). The majority
identified themselves as white (82%), followed by Asian (14%) and black or African American (6%). Very few identified as Hispanic or Latino (6%). Many of them had a high school diploma (39%), and a slightly larger group had some college degree (16% associate’s degree and 32% bachelor’s degree). Only a small minority did not have a high school diploma (1%) or had a higher college degree (11% with a master’s degree and 2% with a doctorate).

On average, workers evaluated 75 words, with seven workers evaluating all 275 words and 11 workers evaluating only one word. The average time spent on a word was 12 s, with the minimum time spent 2 s and the maximum 58 s.

**Actual word difficulty**

Figure 1 shows the average actual difficulty, that is, the percentage of words correctly defined, for the 25 words in each percentile. The results show that words with a lower frequency of occurrence (higher percentile) are more difficult and less often correctly defined by participants. We calculated a one-tailed Pearson correlation coefficient for the results. The average percentage correct correlated negatively with the percentile (N=275, r=-0.381, p<0.001) indicating that higher percentiles contain more difficult words. In addition, a complementary, more specific analysis using word frequencies instead of percentiles confirms this result, with a significant correlation showing that a lower frequency results in fewer words correctly defined (N=275, r=-0.219, p<0.001).

Because word length is a crucial factor in most readability formulae, we evaluated the relationship between word length and actual difficulty. We found no relationship between the word length and actual difficulty (N=275, r=-0.075, p=0.107): the percentage of words correctly defined did not relate to the word length.

To complete our analysis, we also evaluated the time the participants spent evaluating the word and found a significant negative correlation between actual difficulty and time spent (N=275, r=-0.674, p<0.001), indicating that more time was spent on words with lower scores.

**Perceived word difficulty**

Figure 2 shows the average perceived difficulty on a five-point Likert scale for the 25 words in each percentile. The results show that words with a higher frequency of occurrence (lower percentile) are consistently perceived as easier. We calculated a one-tailed Pearson correlation coefficient, which showed a significant positive correlation between perceived difficulty (lower is easier) and percentile (N=275, r=0.611, p<0.001). Similar to actual difficulty, the complementary analysis using word frequencies instead of percentiles confirms this relationship with a significant negative correlation between perceived difficulty and word frequency (N=275, r=-0.397, p<0.001). Words with higher word familiarity are seen as easier.

As with the analysis for actual difficulty, we also evaluated the relationship between perceived difficulty and word length. In contrast to actual difficulty, the length of the word does have an effect on perceived difficulty. There was a positive, significant correlation between perceived difficulty and word length (N=275, r=0.254, p<0.001): longer words are seen as more difficult.

For time spent evaluating the word, there was a significant correlation between time and word length (N=275, r=0.656,
p<0.001) with more time being spent on words seen as more difficult.

CONCLUSIONS
We evaluated the use of a reader’s familiarity with a word, estimated by the word’s frequency in common English text, as a stand-in for word difficulty. We conducted a user study to evaluate 275 words with different frequencies and gathered 50 evaluations for each word of actual difficulty (how well can people choose the correct definition of the word) and of perceived difficulty (how difficult does a word look). Our results show that word frequency is strongly associated with both actual difficulty and perceived difficulty. Words with higher frequency were more often defined correctly and were labeled as appearing less difficult. Word length, frequently used in readability formulae as a stand-in for difficulty, did not relate to actual difficulty and was only weakly related to perceived difficulty. Because of these results, we argue that word frequency is a better metric to estimate word difficulty than word length. Further studies with complete texts instead of single words and with different types of readers are needed to evaluate the metric for its relation to information understanding.

As with all practical studies, our evaluation has limitations. The first relates to workers on Amazon Mechanical Turk. It can be assumed that these workers are more computer literate than other Americans reading online text. Future work will focus on evaluating whether the relationship is as strong in younger readers and readers with lower reading skills. The second limitation relates to the general nature of the words. We did not limit our set to medical words. Furthermore, our procedure limits words to those found in the Moby word list, which may bias the word list towards common words. However, the list is large and our approach helps exclude technical terms or web-specific terms. While this makes the data less medically focused it is still useful for the medical domain because all words, not only medical words, need to be sufficiently simple in patient materials. Similarly, word familiarity was estimated using the Google web corpus, a general corpus. A corpus more specific to a particular patient population, for example, for different ethnicities or age groups, may lead to more fine-tuned formulae. A third limitation is the selection of alternative definitions. Because alternative definitions were assigned automatically and randomly, choosing the correct definition often allowed for easy elimination of one or more alternative definitions resulting in a high percentage of correct answers. We conducted our study in this manner to bring the most difficult case first. We expect that with alternative definitions more closely related to the word, the effect of word familiarity on difficulty will be even stronger. Finally, we point out that we worked with single words (unigrams) and future work will include multiword phrases.

REFERENCES
8 Leroy G, Endicott JE. Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty. Presented at the 2nd ACM SIGKDD International Health Informatics Symposium (ACM IHK 2012); Florida, Miami, 2012.

Acknowledgements The authors would like to thank their study participants.

Funding This work was supported by the US National Library of Medicine, NIH/ NLM 1R03LM010902.

Competing interests None.

Ethics approval The study was reviewed by the institutional review board of Claremont Graduate University.

Provenance and peer review Not commissioned; externally peer reviewed.