Published in final edited form as:

J Health Commun. 2016 ; 21(Suppl): 18–26. doi:10.1080/10810730.2015.1131775.

Effects on Text Simplification: Evaluation of Splitting up Noun Phrases

Gondy Leroy¹, David Kauchak², and Alan Hogue³

¹Department of Management Information Systems, Eller College of Management, University of Arizona, Tucson, AZ, USA

²Computer Science Department, Pomona College, Claremont, CA, USA

³Department of Linguistics, University of Arizona, Tucson, AZ, USA

Abstract

To help increase health literacy, we are developing a text simplification tool that creates more accessible patient education materials. Tool development is guided by data-driven feature analysis comparing simple and difficult text. In the present study, we focus on the common advice to split long noun phrases. Our previous corpus analysis showed that easier texts contained shorter noun phrases. Subsequently, we conduct a user study to measure the difficulty of sentences containing noun phrases of different lengths (2-gram, 3-gram and 4-gram), conditions (split or not) and, to simulate unknown terms, use of pseudowords (present or not). We gathered 35 evaluations for 30 sentences in each condition (3×2×2 conditions) on Amazon's Mechanical Turk (N=12,600). We conducted a three-way ANOVA for perceived and actual difficulty. Splitting noun phrases had a positive effect on perceived difficulty but a negative effect on actual difficulty. The presence of pseudowords increased perceived and actual difficulty. Without pseudowords, longer noun phrase led to increased perceived and actual difficulty. A follow-up study using the phrases (N = 1,350) showed that measuring awkwardness may indicate when to split noun phrases. We conclude that splitting noun phrases benefits perceived difficulty, but hurts actual difficulty when the phrasing becomes less natural.

1. Introduction

Each year, chronic diseases afflict more people. For example, an estimated 50,000 people are infected with HIV yearly (Centers for Disease Control and Prevention, 2011. Published February 2013. Accessed July 2013.) and more than a third of adults were obese in 2010 (Ogden, Carroll, Kit, & Flegal, 2012), and both of these numbers are expected to grow. Additionally, treatments have become more complex and require lifestyle changes from the patient. These kinds of treatments benefit from participatory medicine (Keselman & Smith, 2012) where patients take an active role in their healthcare. Such active involvement requires that people understand their health problems and the possible solutions. Unfortunately,

Corresponding Author at: Gondy Leroy, McClelland Hall Room 430W, 1130 E. Helen Street, Tucson, Arizona 85721, Phone: 520-621-4106, Fax: 520-621-2433, gondyleroy@email.arizona.edu.

The content is solely our own responsibility and does not necessarily represent the official views of the National Institutes of Health.

limited comprehension of healthcare information (Weis, 2007) is leaving millions without sufficient health literacy (Committee on Health Literacy - Institute of Medicine of the National Academies, 2004), complicating care and increasing costs. The problem is not new but is becoming more critical. It has been argued that for the Patient Protection and Affordable Care Act to be successful, more effort is needed to increase the health literacy of millions of Americans. Similarly, the Healthy People 2020 statement by the Department of Health and Human Services identified improving health literacy (HC/HIT-1) as an important national goal.

While there are many options to measure health literacy, e.g., the eHealth literacy scale (eHEALS)(Norman & Skinner, 2006), Test of Functional Health Literacy in Adults (TOFHLA)(Nurss, Parker, Williams, & Baker, 1995) and Rapid Estimate of Adult Literacy in Medicine (REALM)(Davis et al., 1993), few approaches exist to improve existing levels of health literacy. A review by Pignone et al. (Pignone, DeWalt, Sheridan, Berkman, & Lohr, 2005) shows the difficulty of the problem with differential effects of interventions on people with different characteristics. Most approaches focus on improving how information is delivered. For example, teaching professionals writing skills to construct education materials for patients (Goto, Rudd, Lai, & Yoshida-Komiya, 2014). Historically, the most common advice is to simplify text and then use readability formulas to evaluate the text (McLaughlin, 1969; Mullan, Crookes, & Yeatman, 2003). These formulas generate a single number, often based only on word and sentence length, and are used as stand-ins for text complexity (DuBay, 2004). The Flesch-Kincaid grade level formula and the Simple Measure of Gobbledygook (SMOG) are among the most commonly used and recommended in healthcare literature (L.-W. Wang, Miller, Schmitt, & Wen, 2012) and have been used to evaluate a variety of texts, ranging from patient education materials (Kwak, Leroy, Martinez, & Harwell, 2013; Polishchuk, Hashem, & Sabharwal, 2012; Vallance, Taylor, & Lavallee, 2008), websites (Ahmed, Sullivan, Schneiders, & McCrory, 2012; Cameron, 2009; Lam, Roter, & Cohen, 2013) or drug labels (Didonet & Mengue, 2008), to specific information, e.g., abdominal aortic aneurysms (Bailey et al., 2012) or back pain (Hendrick et al., 2012), or for specific groups, e.g., Native Americans (Lease et al., 2013).

Even though text readability and the associated formulas have been the focus of much research, advice, and concern (Meade & Smith, 1991; Pichert & Elam, 1985), they continue to be the prevalent tool for text evaluation. As a result, two critical problems still persist. The first problem is that there is little evidence showing a relationship between these readability measures and user understanding. Specifically, few studies have shown evidence that text simplified based on these formulas results in increased user comprehension. Tanaka et al. (Tanaka, Jatowt, Kato, & Tanaka, 2013) found a weak relationship with perceived difficulty. Others reported a lack of a correlation with Cloze measure results (Friedman, Corwin, Dominick, & Rose, 2009), or discussed problems such as insensitivity to text cohesion (Graesser, McNamara, & Kulikowich, 2011; Wubben, van den Bosch, & Krahmer, 2012) and even an increase in difficulty, i.e. the simplicity paradox (Zarcadoolas, 2010), because the simplification concentrates on writing style rather than content (Y. Wang, 2006). The lack of strong evidence for increased comprehension after using readability formulas may indicate that it is perceived difficulty more than actual difficulty that is being manipulated: the text *looks* easier but may not necessarily *be* easier to understand. In previous work, we

found indirect evidence for this distinction: it is easier to improve the perceived text difficulty than the actual text difficulty (Mouradi, Leroy, Kauchak, & Endicott, 2013). The second problem is that there are few tools available to facilitate, support, and speed up the text simplification process. Providing one overall number indicating text difficulty is not helpful. Tools are needed that focus on specific text features for which there is clear evidence that simplification affects comprehension. The tools should pinpoint difficult sections and suggest easier alternatives.

In our work, we attempt to address both these problems. However, before committing to tool development, we evaluate the potential impact of each individual text feature in user studies. Using this approach, we have found two features that are indicative of difficult text that can be pinpointed algorithmically and show increased comprehension when simplified. They are term familiarity (Gondy Leroy & Endicott, 2011; G. Leroy, Endicott, Kauchak, Mouradi, & Just, 2013) and grammar familiarity. Both measure how frequently a term or grammar structure is encountered by laypersons.

In this work, we focus on noun phrase complexity. Splitting noun phrases into smaller chunks is commonly advised and there is some data-driven support for it. Noun phrase complexity is mentioned explicitly in the Plain Language initiative where it is advised to split noun phrases of more than three nouns by using prepositions (<http://www.plainlanguage.gov/howto/guidelines/FederalPLGuidelines/writeNoNounStrings.cfm>). For initial validation, we evaluated noun phrase complexity in three different corpora and found that difficult texts contain longer compound noun phrases (Gondy Leroy & Endicott, 2012b). While pinpointing long noun phrases can be automated and splitting can be semi-automated, we prefer to show first that splitting noun phrases leads to increased comprehension before developing tools for general use. We focus on reader comprehension and not formula-driven reading level. To our knowledge no studies have directly evaluated the impact of splitting noun phrases on perceived and actual text difficulty.

In this study, we measure the effect of splitting longer noun phrases into smaller constituents. To provide a detailed and systematic overview, we test increasingly longer noun phrases that contain phrases with two, three and four words. Furthermore, we test each sentence in two settings: with and without pseudowords. Adding pseudowords increases external validity since it simulates an often encountered situation by patients where the text contains medical terms that are not understood or where a text is not in the native language, e.g., in English for native Spanish speakers. In each condition, we evaluate the perceived and actual difficulty of the sentences.

2. Methods

2.1. Stimuli

We collected 8,247 articles from English Wikipedia's (<http://en.wikipedia.org/>) Disease category (now called "Diseases and Disorders") and parsed all the text using the Berkeley Parser. We utilized Wikipedia for the study since a majority of people obtain health-related texts on the web (Fox, 2011) and Wikipedia is a very common source of information on the web (Safran, 2012). We focus on single sentences in our study to tease out the effect of

changes in one noun phrase. Using longer text would require more simplification and might introduce confounding variables. We selected those sentences containing six or more nouns (needed for evaluation, see below) and at least one noun phrase containing two words (2-gram), three words (3-gram) or four words (4-gram).

We imposed several constraints to create a subset of sentences where we could test the effect of splitting while minimizing influences from other effects. First, we ensured that the noun phrase was not a technical term because such phrases require more than splitting to simplify. For example, sentences with phrases such as “monkeypox virus”, “DNS cross links” or “chronic obstructive lung disease” were discarded. Second, sentences with proper nouns, for example “Francis Xavier de Balmis” and “Boston pathologist Sidney Farber”, were also discarded. Third, we selected sentences where the noun phrase could be found in the Google Web Corpus so that we could control the phrase familiarity. In previous studies, term familiarity has been shown to be an important factor indicating that words with higher frequencies are easier to understand. To ensure a consistent dataset, we included only sentences where the frequency of occurrence of constituents after splitting the noun phrase increases. For example, the frequency of “motor nerve conduction velocities” is 791 (in the Google Web Corpus) and after splitting into “conduction velocities of motor nerves” the frequency increases to 10,498 (conduction velocities) and 12,804 (motor nerves) with an average of 11,650. After this selection process, we randomly selected sets of 30 example sentences for noun phrases of length two, three and four words for a total of 90 sentences. Each set was then used in the different experimental conditions.

We include three independent variables. The first independent variable (IV1) measures the effect of split versus no split of the noun phrase, as is shown with examples 1–2 and examples 4–5 in Table 1. The splitting itself was conducted by the two native English speakers on the team (2nd and 3rd author). The second independent variable (IV2) measures the effect of pseudoword versus no pseudoword. The pseudowords were generated using Wuggy (Keuleers & Brysbaert, 2010). We generated a list of pseudowords (e.g., crumering, dutter, seducated, bimy, woft, jellage), randomized the order and used them to replace two nouns in each sentence, see example 3 in Table 1. The third independent variable (IV3) measures the effect of the size of the noun phrase to be split: 2-gram, 3-gram or 4-gram.

2.2. Metrics

We measure perceived and actual text difficulty. Perceived difficulty is measured with a 5-point Likert scale for each sentence. Participants are asked to rate the difficulty of a sentence by choosing from the following options: Very Easy, Easy, Neither, Difficult and Very Difficult. A lower number indicates an easier sentence (Scores used: Very Easy, 1 – Very Difficult, 5).

We measure actual difficulty using an adjusted Cloze test. While other metrics could be used, e.g., sentence completion, we chose the adjusted Cloze measure to facilitate comparison with our and others' previous work. Furthermore, it allows testing of individual sentences thereby reducing the chance of introducing confounding variables. Finally, it allows for automated testing thereby reducing the chances of introducing bias in scoring. The original Cloze measure requires that every n^{th} word of a text is deleted and participants

are asked to fill in the blanks. The measure was introduced and validated by Taylor (Taylor, 1953) to distinguish between texts with different readability levels. Later, it was adopted as a measure of user comprehension (Siddharthan, 2002). Different versions of the test, e.g., a different number or different word classes being blanked, lead to different absolute numbers but result in the same conclusions when comparing texts. For our test, we blank out four nouns in each sentence. The blanked nouns are then used to create five different multiple-choice options: one correct ordering and four random (incorrect) orderings of the words. Participants were asked to choose the option that resulted in a correct sentence when the words were inserted into the blanks in order. Figure 1 shows an example of one sentence task, as it was presented to participants online.

2.3. Study Procedure

We conducted the study using Amazon's Mechanical Turk (MTurk). MTurk is an online service that allows "requesters" to upload tasks (called HITS, human intelligence tasks) for participants (referred to as "workers") to do for a price. Workers search the list of possible HITS using the task description, keywords and payment and decide if they'd like to participate. MTurk has been used in a variety of settings including data collection and annotation and user studies (Kittur, Chi, & Suh, 2008). There are over 400,000 workers on MTurk with a broad range of demographics (Ross, Irani, Silberman, Zaldivar, & Tomlinson, 2010). When care is taken to validate worker submissions (e.g. by adding validation questions, see below) data collected via MTurk has been shown to be as good if not better than traditional sources (Zaidan & Callison-Burch, 2011).

For each sentence in each condition evaluations were gathered from 35 different participants resulting in 12,600 data points (35 evaluations x 30 sentences x 2 split/not split conditions x 2 pseudowords/no pseudowords condition x 3 2-gram/3-gram/4-gram conditions). We restricted our workers to those based in the US with a HIT approval rating of 95% or more.

Our main goal is to evaluate our text manipulation, i.e., the impact of an algorithm, not to study individuals. We will therefore analyze the results as a 2x2x3 between subjects design. The four conditions (pseudowords/no pseudowords x split/not split) were presented at four different dates. Each session contained 90 different sentences with 30 sentences for each of the three NP conditions (2-gram/3-gram/4-gram). This ensured no sentences were repeated to any participant within the test condition for that session. Participants could complete all 90 sentences or stop at any time. The order of the 90 sentences was randomized in each condition. There was minimally one week between the four conditions to allow workers to participate in each task, if they desired to. Since participants are not intentionally memorizing sentences and exposure to each sentence is limited to a few seconds, one week was estimated to be a sufficiently long time to avoid carry-over effects for those who did participate in multiple conditions. Each sentence was presented as one HIT and participants were paid 3 cents per HIT. All conditions are treated as between-subjects conditions because participant cannot be forced to complete all 90 sentences and because we allow sufficient time between the four different test conditions (a week) to ensure no carry over effects exist for those who participate in more than one condition.

When workers selected our task, they were first asked to complete five demographic questions. We asked participants about gender, race, ethnicity, education level, and English language skills. After completing these demographic questions, the HITS were presented one at the time and workers could complete as many as they liked (maximally 90 in each task).

3. Results

3.1. Participant Demographic Information

The data collection was completed in late Spring 2014 over a 4-week period. A total of 353 people participated. Since MTurk workers can fail to do the task appropriately, e.g., clicking an option without attempting to find the correct answer, we eliminated outliers from our sample. We calculated the average accuracy and standard deviation over all workers (mean: 86%; standard deviation: 20%). Since our task is fairly easy, we removed those workers who scored on average two standard deviations below average (46% accuracy or lower). As a result, nineteen workers were removed from the dataset as well as the 651 HITS they had completed. The remaining 334 workers accounted for a total of 11,949 HITS. In this group, workers completed on average 36 HITS. The minimum number of HITS completed by a worker was one and the maximum completed by one worker was 354.

Table 2 provides an overview of the demographic information of the 334 workers retained in our sample. Overall, there were almost an equal number of male (52%) and female (48%) participants. Most participants reported not to be Hispanic or Latino (93%). The largest racial group were those reporting as White (79%) and the smallest those reporting as Native Hawaiian/Pacific Islander (1%) and American Indian/Alaska Native (3%). There were almost an equal number of Asian (8%) and Black (9%) participants. Participants could indicate multiple races and 5% indicated two or more races.

Since we provide English language text and aim to measure comprehensions, we asked participants about language skills and education level. Most (95%) spoke exclusively English at home, with a much smaller group (4%) speaking mostly English or English only half of the time (1%). All educational levels were represented. Most participants had earned either a High School Diploma (33%) or Bachelor's Degree (37%) as their highest degree. The smallest groups were those with Less than a High School Diploma (1%) or a Doctoral Degree (2%). There was also a small group with a Master's Degree (9%).

3.2. Main Analyses: Actual and Perceived Difficulty

Since we conducted our study with single sentences and provide multiple-choice answers for the Cloze test, the average accuracy is high (low error) and the differences are small between conditions. However, our data set was sufficiently large to show systematic effects. For easier interpretation we present actual difficulty as an error percentage so that higher numbers consistently indicate greater difficulty in Figures 2–5.

We first conducted a three-way ANOVA for perceived difficulty. Perceived difficulty evaluation showed the expected results with three main effects and one interaction effect. There was a main effect of splitting the noun phrases. On average, sentences containing the

split noun phrases were perceived as being easier to understand, though the differences were small with an average score of 2.44 for sentences with the split phrase and 2.49 when not split. Figure 2 shows this difference to be more pronounced with longer noun phrases, but the interaction was not statistically significant. The interaction between splitting and the presence of pseudowords was significant ($F(1,11937) = 13.334, p < .001$) and Figure 2 shows that splitting noun phrases does not affect perceived difficulty when pseudowords are present.

There was a second main effect of the length of noun phrases ($F(2,11937) = 25.142, p < .001$) with sentences containing the 2-grams seen as the easiest and those containing the 4-grams as the most difficult. And, as expected, there was also a third main effect of pseudowords ($F(1,11937) = 148.478, p < .001$) with sentences containing pseudowords perceived to be more difficult.

We then conducted a three-way ANOVA for actual difficulty. This analysis clearly demonstrates the need for separating perceived and actual difficulty. We found a significant main effect of each independent variable and one significant interaction. Figure 3 shows the detailed results for all conditions. We found a main effect of splitting noun phrases ($F(1,11937) = 19,669, p < .001$), however, splitting decreased comprehension. On average, the multiple-choice tasks showed 8.5% error before splitting and the error increased to 10.9% after splitting noun phrases. For 4-grams, this difference was the largest with the error percentage increasing by 5% absolute after splitting: 7% error before splitting to 12% error after splitting. As seen in Figure 3, the patterns across different conditions vary and we found a significant 3-way interaction between our independent variables ($F(2, 11937) = 7.098, p = .001$). Larger errors were found with 3-grams and 4-grams when there are no pseudowords and for 2-grams and 3-grams when there are pseudowords.

The results also show a second main effect of the noun phrase length ($F(2,11937) = 9.413, p < .001$). However, there is not a clear relationship between accuracy and the length of the noun phrase. On average, the error was 8.6% for 2-gram sentences, 11.4% for 3-gram sentences and 9.3% for 4-gram sentences. The third main effect was for the use of pseudowords ($F(1,11937) = 6.136, p = .013$). As with perceived difficulty, the presence of pseudowords also lowered accuracy and the error was 9.1% for sentences without pseudowords and 10.4% for sentences with pseudowords.

3.3. Follow-Up Analysis

Given the contradiction in effects between perceived and actual difficulty, we conducted follow-up analyses to help understand these results. One possible explanation for the conflicting results is that the particular preposition used to split the noun phrase plays a role in the effectiveness of splitting a particular noun phrase. Table 3 shows the prepositions used when splitting longer noun phrases into shorter ones. The most commonly used preposition was “of” followed by “in”. These numbers help interpret Figure 4 and Figure 5.

Figure 4 and Figure 5 show the perceived and actual difficulty separated based on the preposition used to split the noun phrase. By looking at the prepositions at this granularity, we see some patterns emerge and that splitting using certain prepositions does result in

easier looking text *and* text that is easier to understand. There were several prepositions for which we found improvements. For example, there was a consistent effect when splitting with the prepositions ‘using’ or ‘with’. Although they were only used a few times and were not used in any of the 2-grams, they were perceived as easier and also resulted in fewer errors. For 2-grams with ‘for’ and ‘in’ splits and 3-grams with ‘during’ and ‘involving’ a similar pattern emerges and the both perceived and actual difficulty improve after splitting. For other prepositions, the perceived difficulty is not as good an indicator of actual difficulty. For example, splitting a 2-gram using ‘by’ is seen as equally difficult but resulted in many more errors. This again highlights the need to separate evaluation between perceived and actual difficulty.

Another factor that may contribute to the results seen above is simply that not all noun phrases should be split. When splitting the noun phrases, we noticed that some of the split noun phrases tended to create more awkwardness. To understand this effect we evaluated the awkwardness of each phrase. If awkwardness, something experienced when reading some of our split phrases, influences difficulty, we may be able to capture this algorithmically for inclusion in writing tools. To this end, we conducted an additional small analysis of our phrases on MTurk. We presented each phrase used in our study (90 total) in its original and its split version. We asked MTurk workers to choose the version that was the most natural phrasing. We randomized the order of the two versions (split, not split) and added a third option to indicate when they were both equally natural sounding. Participants choose whether they preferred the split version, the original version or did not have a preference. We added four qualification tasks where the workers were instructed to choose a certain option. This allowed us to remove workers from our datasets who do not take the time to read each task.

For each phrase, we collected evaluations from fifteen different participants. The study was completed in the spring of 2015 by thirty participants for a total of 1,350 data points. We removed the HITS of five workers who failed one or more of the qualification tests. For the remaining 1,132 data points, we calculated the percentage of participants who preferred the noun phrase as split/not split/either for each preposition (e.g., ‘at’, ‘by’, ‘during’, ...) used to split the noun phrases (N=21). We found no relation between naturalness of the split phrases and perceived difficulty. However, we found a significant correlation between naturalness of split noun phrases and the actual difficulty when these split noun phrases were tested in a sentence (one-tailed Pearson Correlation, $r = .427$, $p = .027$): when the split version was preferred the accuracy was higher (fewer errors). Figure 6 shows the scatterplot for the significant correlation.

4. Discussion

Our overall goal is the development of a text simplification tool in support of providers of healthcare information. We start with large scale corpus analyses of known easy and difficult texts. We use natural language processing techniques to identify and compare differentiating features, e.g., length of noun phrases. By using algorithmic approaches from the start, we facilitate later translation into automated tools. Promising features are evaluated for two

important characteristics: impact on comprehension and potential for semi-automated simplification approaches.

This study demonstrated the necessity of addressing the effect on comprehension. We examined the general advice that long noun phrases increase the difficulty of text and contribute to reduced comprehension. In previous work (Gondy Leroy & Endicott, 2012a), we found evidence that easy texts, e.g., blogs by laypersons, contain shorter noun phrases. However, controlling as much as possible for confounding variables, we split target noun phrases in each sentence and, surprisingly, we found that splitting noun phrases did not increase the comprehension of the sentence. Although the sentences with the split noun phrases were perceived as easier, they were not easier to comprehend after splitting. We judge that splitting noun phrases requires more nuance than splitting every noun phrase. There is little motivation to split 2-grams. For 3-grams and 4-grams, only noun phrases that required ‘using’ and ‘with’ received a benefit from splitting the noun phrase. However, in some cases, a certain awkwardness was introduced that was clear to the authors. We quantified this awkwardness in a new, follow-up study by asking people which phrases, split or not split, were more natural. We found that phrases perceived as more natural after splitting were also those with fewer errors in the modified Cloze test.

To provide a sensitive test and increase external validity, we presented all sentences with and without pseudowords. We found clear effects on both perceived and actual difficulty of pseudowords: they increased difficulty. The effect of the pseudowords seemed to overshadow the effect of splitting noun phrases. When pseudowords were present, the perceived difficulty of the sentences remained the same regardless of whether the noun phrase was split or not.

Implications for practice and future research from the presented study:

- Common and longstanding advice needs to be verified empirically for its effects on comprehension. There exist many unsubstantiated rules for simplification that may be harmful not beneficial to simplifying text.
- Not all noun phrases should be split. The best advice resulting from this study is to split phrases only when the split phrases feel more natural. Additional studies are needed to tease out different effects for different prepositional or compound noun phrases, to quantify ‘natural’ and automate discovering such phrases. ‘Natural’ was defined here by native English speakers.
- Terms that are not understood, e.g., they may be too technical or because the reader is not a native speaker, have a great impact on overall comprehension. Future tools should help writers replace difficult terms or allow readers look up the meanings of words in a convenient manner.
- The preposition required when splitting the noun phrase may be another source of information for determining whether or not to split the noun phrase. We saw some initial patterns in our sample, but a study over a larger set of noun phrases may help make the role of the preposition clearer.

5. Conclusion

Overall, text difficulty affects how well readers can learn the content of a text. Two aspects of text difficulty are the perceived and actual difficulty. We measure both separately in all our work and often discover different aspects affecting them. In this project, we evaluated the effect of long noun phrases on the difficulty of text. We found that overall splitting noun phrases is beneficial for improving perceived difficulty. The effect on actual difficulty is more complex and we only found evidence that splitting the noun phrases increased comprehension in a few cases, frequently where the split noun phrase was identified as sounding more natural than the original.

Acknowledgments

The authors would like to thank their study participants and also the creators of Wuggy (<http://crr.ugent.be/programs-data/wuggy>) for making their work available for free to the research community.

7. Funding

Research reported in this article was supported by the National Library of Medicine of the National Institutes of Health under Award Nos. R03LM010902 and R01LM011975.

References

- Ahmed OH, Sullivan SJ, Schneiders AG, McCrory PR. Concussion Information Online: Evaluation of Information Quality, Content and Readability of Concussion-related Websites. *Br J Sports Med.* 2012; 46(9):675–683.10.1136/bjsm.2010.081620 [PubMed: 21504964]
- Bailey MA, Coughlin PA, Sohrabi S, Griffin KJ, Rashid ST, Troxler MA, Scott DJ. Quality and Readability of Online Patient Information for Abdominal Aortic Aneurysms. *J Vasc Surg.* 2012; 56(1):21–26.10.1016/j.jvs.2011.12.063 [PubMed: 22521801]
- Cameron KA. A Practitioner's Guide to Persuasion: An Overview of 15 selected Persuasion Theories, Models and Frameworks. *Patient Education and Counseling.* 2009; 74:309–317. [PubMed: 19136229]
- Centers for Disease Control and Prevention. HIV Surveillance Report. 2011. Published February 2013 [Accessed July 2013]
- Committee on Health Literacy - Institute of Medicine of the National Academies. *Health Literacy: A Prescription to End Confusion.* Washington, DC: The National Academies Press; 2004.
- Davis T, Long S, Jackson R, Mayeaux E, George R, Murphy P, Crouch M. Rapid Estimate of Adult Literacy in Medicine: a Shortened Screening Instrument. *Family Medicine.* 1993; 25(6):391–395. [PubMed: 8349060]
- Didonet J, Mengue SS. Drug labels: Are they a readable material? *Patient Education and Counseling.* 2008; 73:141–145. [PubMed: 18572350]
- DuBay, WH. The principles of readability. Impact information. 2004. <http://www.impact-information.com/impactinfo/readability02.pdf>
- Fox S. Health topics. Pew Internet & American Life Project. 2011
- Friedman DB, Corwin SJ, Dominick GM, Rose ID. African American Men's Understanding and Perceptions about Prostate Cancer: Why Multiple Dimensions of Health Literacy are Important in Cancer Communication. *Journal of Community Health.* 2009; 34:449–460. [PubMed: 19517223]
- Goto A, Rudd RE, Lai AY, Yoshida-Komiya H. Health Literacy Training for Public Health Nurses in Fukushima: A Case-Study of Program Adaptation, Implementation and Evaluation. *Japan Med Assoc J.* 2014; 57(3):146–153.
- Graesser AC, McNamara DS, Kulikowich JM. Coh-Metrix: Providing Multilevel Analyses of Text Characteristics. *EDUCATIONAL RESEARCHER.* 2011; 40:223–224.10.3102/0013189X11413260

- Hendrick PA, Ahmed OH, Bankier SS, Chan TJ, Crawford SA, Ryder CR, ... Schneiders AG. Acute low back pain information online: an evaluation of quality, content accuracy and readability of related websites. *Man Ther.* 2012; 17(4):318–324.10.1016/j.math.2012.02.019 [PubMed: 22464886]
- Keselman A, Smith CA. A Classification of Errors in Lay Comprehension of Medical Documents. *Journal of Biomedical Informatics.* 2012; 45:1151–1163. [PubMed: 22925723]
- Keuleers E, Brysbaert M. Wuggy: A Multilingual Pseudoword Generator. *Behavior Research Methods.* 2010; 42(3):627–633. [PubMed: 20805584]
- Kittur, A.; Chi, EH.; Suh, B. Crowdsourcing user studies with Mechanical Turk. Paper presented at the Proceedings of the SIGCHI conference on human factors in computing systems; 2008.
- Kwak M, Leroy G, Martinez JD, Harwell J. Development and Evaluation of a Biomedical Search Engine using a Predicate-based Vector Space Model. *Journal of Biomedical Informatics.* 2013; 15(7):e144.
- Lam CG, Roter DL, Cohen KJ. Survey of quality, readability, and social reach of website on osteosarcoma in adolescents. *Patient Education and Counseling.* 2013; 90:82–87. [PubMed: 22959332]
- Lease, M.; Hullman, J.; Bigham, JP.; Bernstein, MS.; Kim, J.; Lasecki, W.; ... Miller, RC. Mechanical Turk is Not Anonymous. *Social Science Research Network.* 2013. (<http://dx.doi.org/10.2139/ssrn.2228728>). Retrieved from doi:<http://dx.doi.org/10.2139/ssrn.2228728>
- Leroy, G.; Endicott, JE. Term Familiarity to Indicate Perceived and Actual Difficulty of Text in Medical Digital Libraries. Paper presented at the International Conference on Asia-Pacific Digital Libraries (ICADL 2011) - Digital Libraries -- for Culture Heritage, Knowledge Dissemination, and Future Creation; Beijing, China. 2011 Oct 24–27.
- Leroy, G.; Endicott, JE. Combining NLP with Evidence-based Methods to Find Text Metrics related to Perceived and Actual Text Difficulty. Paper presented at the 2nd ACM SIGHIT International Health Informatics Symposium (ACM IHI 2012); Florida, Miami. 2012a.
- Leroy, G.; Endicott, JE. Combining NLP with Evidence-based Methods to Find Text Metrics related to Perceived and Actual Text Difficulty. Paper presented at the 2nd ACM SIGHIT International Health Informatics Symposium (ACM IHI 2012); Florida, Miami. 2012b Jan 28–30.
- Leroy G, Endicott JE, Kauchak D, Mouradi O, Just M. 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 (JMIR).* 2013; 15(7):e144. doi: 110.2196/jmir.2569. [PubMed: 23903235]
- McLaughlin GH. SMOG Grading: a New Readability Formula. *Journal of Reading.* 1969; 12:636–646.
- Meade CD, Smith CF. Readability Formulas: Cautions and Criteria. *Patient Education and Counseling.* 1991; 17:153–158.
- Mouradi, O.; Leroy, G.; Kauchak, D.; Endicott, JE. Influence of Text and Participant Characteristics on Perceived and Actual Text Difficulty. Paper presented at the Hawaii International Conference on System Sciences; Maui. HI. 2013.
- Mullan J, Crookes PA, Yeatman H. Rain, fog, smog and printed educational material. *Journal of Pharmacy Practice and Research.* 2003; 33(4):284–286.
- Norman C, Skinner H. eHEALS: The eHealth Literacy Scale. *Journal of Medical Internet Research (JMIR).* 2006; 8(4):e27. [PubMed: 17213046]
- Nurss, JR.; Parker, RM.; Williams, MV.; Baker, DW. Test of Functional Health Literacy in Adults. Hartford, MI: Peppercorn Books & Press; 1995.
- Ogden CL, Carroll MD, Kit BK, Flegal KM. Prevalence of Obesity in the United States, 2009–2010. *NCHS Data Brief.* 2012:82.
- Pichert JW, Elam P. Readability Formulas May Mislead You. *Patient Education and Counseling.* 1985; 7(1985):181–191.
- Pignone M, DeWalt DA, Sheridan S, Berkman N, Lohr JN. Interventions to Improve Health Outcomes for Patients with Low Literacy: A Systematic Review. *J Gen Intern Med.* 2005; 20(2):185–193. [PubMed: 15836553]

- Polishchuk DL, Hashem J, Sabharwal S. Readability of Online Patient Education Materials on Adult Reconstruction Web sites. *J Arthroplasty*. 2012; 27(5):716–719.10.1016/j.arth.2011.08.020 [PubMed: 22000573]
- Ross, J.; Irani, L.; Silberman, M.; Zaldivar, A.; Tomlinson, B. Who are the crowdworkers?: shifting demographics in mechanical turk. Paper presented at the CHI'10 Extended Abstracts on Human Factors in Computing Systems; 2010.
- Safran, N. Wikipedia in the SERPs. 2012. from <http://www.conductor.com/blog/2012/03/wikipedia-in-the-serps-appears-on-page-1-for-60-of-informational-34-transactional-queries/>
- Siddharthan, A. Resolving Attachment and Clause Boundary Ambiguities for Simplifying Relative Clause Constructs. Paper presented at the 40th Meeting of the Association for Computational Linguistics (ACL); 2002.
- Tanaka, S.; Jatowt, A.; Kato, MP.; Tanaka, K. Estimating Content Concreteness for Finding Comprehensible Documents. Paper presented at the Sixth ACM International Conference on Web Search and Data Mining (WSDM'13); Rome, Italy. 2013.
- Taylor WL. Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*. 1953; 30:415–433.
- Vallance JK, Taylor LM, Lavalley C. Suitability and readability assessment of educational print resources related to physical activity: Implications and recommendations for practice. *Patient Education and Counseling*. 2008; 72:342–349. [PubMed: 18450409]
- Wang L-W, Miller MJ, Schmitt MR, Wen FK. Assessing Readability Formula Differences with Written Health Information Materials: Application, Results, and Recommendations. *Research in Social & Administrative Pharmacy*. 2012 In Press.
- Wang, Y. Automatic Recognition of Text Difficulty from Consumers Health Information. Paper presented at the 19th IEEE International Symposium on Computer-Based Medical Systems; 2006.
- Weis, BD. Health Literacy and Patient Safety: Help Patients Understand. Manual for Clinicians. 2. AMA and AMA Foundation; 2007.
- Wubben, S.; van den Bosch, A.; Krahmer, E. Sentence Simplification by Monolingual Machine Translation. Paper presented at the Proceedings of the Annual Meeting of the Association for Computational Linguistics; Jeju Island, Korea. 2012.
- Zaidan, OF.; Callison-Burch, C. Crowdsourcing translation: Professional quality from non-professionals. Paper presented at the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1; 2011.
- Zarcadoolas C. The Simplicity Complex: Exploring Simplified Health Messages in a Complex World. *Health Promotion International*. 2010; 26(3):338–350.

We are building software to help doctors and nurses write text that is easier to understand. To do this, we need to understand how difficult individual sentence are.

Compared to open thoracotomy, VATS offers a shorter in-hospital ____, less need for postoperative pain control, and a reduced ____ of lung ____ after ____.

Fill in the blanks by choosing the correct option:

- ☐ stays, risk, problems, surgery
- ☐ stays, surgery, problems, risk
- ☐ surgery, stays, risk, problems
- ☐ risk, problems, surgery, stays
- ☐ risk, stays, problems, surgery

How difficulty would this sentence look in a text?

- ☐ Very Easy
- ☐ Easy
- ☐ Neither
- ☐ Difficult
- ☐ Very Difficult

Figure 1.
Example HIT on Amazon Mechanical Turk

Figure 2.
Perceived Difficulty of Sentences (Higher Score = More Difficult)

Figure 3.
Actual Difficulty of Sentences (Higher Score = More Errors)

Figure 4.
Perceived Difficulty by Split Word (Higher Score = More Difficult)

Figure 5.
Actual Difficulty by Split Word (Higher Score = More Errors)

Figure 6.
Scatter Plot of Accuracy x Preference for Splitting Phrase

Table 1

Example Study Sentences (one feature highlighted)

Nr.	IV1: Split or No Split	IV2: Pseudo or No Pseudo	IV3: 2-gram, 3-gram or 4-gram	Example
1	No split	No Pseudo	2-gram	Parenting style seems to have no major effect, although people with supportive parents do better than those with critical or hostile parents.
2	Split	No Pseudo	2-gram	Style of parenting seems to have no major effect, although people with supportive parents do better than those with critical or hostile parents.
3	No Split	Pseudo	3-gram	The polysomnogram involves continuous <i>encming</i> of sleep brain waves and a number of nerve and muscle functions during <i>ebunt</i> sleep.
4	No split	No Pseudo	3-gram	Gene replacement studies in mice suggest that autistic symptoms are closely related to later developmental steps that depend on activity in synapses and on activity-dependent changes.
5	Split	No Pseudo	3-gram	Studies involving gene replacement in mice suggest that autistic symptoms are closely related to later developmental steps that depend on activity in synapses and on activity-dependent changes.
6	No split	No Pseudo	4-gram	Lung volume reduction surgery (LVRS) can improve the quality of life for certain carefully selected patients.

Table 2**Participant Demographic Information**

	N	%
Total	334	
Gender		
Male	175	52
Female	159	48
Ethnicity		
Hispanic of Latino	23	7
Not Hispanic or Latino	311	93
Race (Multiple Choices Allowed)	(356)	
American Indian/Alaska Native	9	3
Asian	30	8
Black	33	9
Native Hawaiian/Pacific Islander	4	1
White	280	79
More than One Race	18	5
Education Level		
Less than High School Degree	2	1
High School Diploma	111	33
Associate Degree	58	17
Bachelor's Degree	125	37
Master's Degree	31	9
Doctoral Degree	7	2
Language Spoken at Home		
Never English	0	-
Rarely English	0	-
Half English	2	1
Mostly English	14	4
Only English	318	95

Table 3

Prepositions Used to Split the Noun Phrases

Words used	Noun Phrase Length			
	2-gram N	3-gram N	4-gram N	Combined N
at	1	-	-	1
by	1	-	-	1
during	-	2	-	2
for	1	5	7	13
from	1	2	7	10
in	8	7	2	17
involving	-	1	-	1
of	16	10	12	38
that	2	1	-	3
using	-	-	1	1
with	-	2	1	3