Taylor & Francis
Taylor & Francis Group

Journal of Health Communication
International Perspectives

ISSN: 1081-0730 (Print) 1087-0415 (Online) Journal homepage: https://www.tandfonline.com/loi/uhcm20

Effects on Text Simplification: Evaluation of Splitting Up Noun Phrases

Gondy Leroy, David Kauchak & Alan Hogue

To cite this article: Gondy Leroy, David Kauchak & Alan Hogue (2016) Effects on Text Simplification: Evaluation of Splitting Up Noun Phrases, Journal of Health Communication, 21:sup1, 18-26, DOI: 10.1080/10810730.2015.1131775

To link to this article: https://doi.org/10.1080/10810730.2015.1131775

Journal of Health Communication, 21: 18–26, 2016

ISSN: 1081-0730 print/1087-0415 online
DOI: 10.1080/10810730.2015.1131775

Routledge
Taylor & Francis Group

Effects on Text Simplification: Evaluation of Splitting Up Noun Phrases

GONDY LEROY[1], DAVID KAUCHAK[2], and ALAN HOGUE[3]

[1]*Management Information Systems Department, University of Arizona, Tucson, Arizona, USA*
[2]*Computer Science Department, Pomona College, Claremont, California, USA*
[3]*Department of Linguistics, University of Arizona, Tucson, Arizona, USA*

To help increase health literacy, we are developing a text simplification tool that creates more accessible patient education materials. Tool development is guided by a 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 conducted a user study to measure the difficulty of sentences containing noun phrases of different lengths (2-gram, 3-gram, and 4-gram); noun phrases of different conditions (split or not); and, to simulate unknown terms, pseudowords (present or not). We gathered 35 evaluations for 30 sentences in each condition ($3 \times 2 \times 2$ conditions) on Amazon's Mechanical Turk ($N = 12{,}600$). We conducted a 3-way analysis of variance 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 phrases 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.

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), 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. In addition, treatments have become more complex and require lifestyle changes from the patient. These kinds of treatments benefit from participatory medicine (Keselman & Smith, 2012), in which patients take an active role in their health care. Such active involvement requires that people understand their health problems and the possible solutions. Unfortunately, limited comprehension of health care information (Weiss, 2007) is leaving millions without sufficient health literacy (Institute of Medicine, 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 U.S. Department of Health and Human Services identified improving health literacy (Health Communication and Health Information Technology [HC/HIT]-1) as an important national goal.

Although there are many options for measuring health literacy —for example, the eHealth Literacy Scale (Norman & Skinner, 2006), Test of Functional Health Literacy in Adults (Nurss, Parker, Williams, & Baker, 1995), and Rapid Estimate of Adult Literacy in Medicine (Davis et al., 1993)—few approaches exist to improve existing levels of health literacy. A review by Pignone, DeWalt, Sheridan, Berkman, and Lohr (2005) showed the difficulty of the problem, with differential effects of interventions on people with different characteristics. Most approaches focus on improving how information is delivered (e.g., teaching professionals writing skills to construct education materials for patients, Goto, Rudd, Lai, & Yoshida-Komiya, 2014). Historically, the most common advice has been 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 are among the most commonly used and recommended in the health care 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., on abdominal aortic aneurysms, Bailey et al., 2012; or back pain, Hendrick et al., 2012) or information for specific reader 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

Address correspondence to Gondy Leroy, Management Information Systems Department, University of Arizona, 1130 East Helen Street, McClelland Hall, Tucson, AZ 85721, USA. E-mail: gondyleroy@email.arizona.edu

(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, Jatowt, Kato, and Tanaka (2013) found a weak relationship with perceived difficulty. Others have 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, 2011) 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 of 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 (Leroy & Endicott, 2011; 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, in which 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 (Leroy & Endicott, 2012). Although pinpointing long noun phrases can be automated and splitting can be semiautomated, 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 measured the effect of splitting longer noun phrases into smaller constituents. To provide a detailed and systematic overview, we tested increasingly longer noun phrases that contained phrases with two, three, and four words. Furthermore, we tested each sentence in two settings: with and without pseudowords. Adding pseudowords increases external validity because it simulates a situation often encountered by patients in which the text contains medical terms that are not understood or in which the text is not in the native language (e.g., in English for native Spanish speakers). For each condition, we evaluated the perceived and actual difficulty of the sentences.

Methods

Stimuli

We collected 8,247 articles from English Wikipedia's (http://en.wikipedia.org/) "Disease" category (now called "Diseases and Disorders") and parsed all of the text using the Berkeley Parser. We utilized Wikipedia for the study because the 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 focused on single sentences in our study to tease out the effect of changes in one noun phrase. Using longer text would have required more simplification and might have introduced confounding variables. We selected those sentences containing six or more nouns (needed for evaluation, as described subsequently) 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 on which 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 in which 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 data set, we included only sentences for which the frequency of occurrence of constituents increased after splitting the noun phrase. For example, the frequency of "motor nerve conduction velocities" was 791 (in the Google Web Corpus), and after splitting into "conduction velocities of motor nerves" the frequency increased 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 two, three, and four words in length, for a total of 90 sentences. Each set was then used in the different experimental conditions.

We included three independent variables. The first independent variable measured the effect of split versus no split of the noun phrase, as is shown with Examples 1 and 2 and Examples 4 and 5 in Table 1. The splitting itself was conducted by the two

native English speakers on the team (the second and third authors). The second independent variable measured 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 measured the effect of the size of the noun phrase to be split: 2-gram, 3-gram, or 4-gram.

Metrics

We measured perceived and actual text difficulty. Perceived difficulty was measured with a 5-point Likert scale for each sentence. Participants were asked to rate the difficulty of a sentence by choosing from the following options: very easy, easy, neither, difficult, and very difficult, where 1 = *very easy* and 5 = *very difficult*. A lower number indicates an easier sentence.

We measured actual difficulty using an adjusted Cloze test. Although other metrics (e.g., sentence completion) could have been used, we chose the adjusted Cloze measure to facilitate comparison with our and others' previous work. Furthermore, it allowed for the testing of individual sentences, thereby reducing the chance of introducing confounding variables, and finally also allowed 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 be deleted and participants be asked to fill in the blanks. The measure was introduced and validated by 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., different numbers or different word classes being blanked out) lead to different absolute numbers but result in the same conclusions when comparing texts are compared. For our test, we blanked out four nouns in each sentence. The blanked-out nouns were 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.

Study Procedure

We conducted the study using Amazon's Mechanical Turk (MTurk). MTurk is an online service that allows requesters to upload tasks (called human intelligence tasks [HITs]) 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 whether they would 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 more than 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, as described subsequently), data collected via MTurk have been shown to be as good if not better than data collected from 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 × 30 sentences × 2 split/not split conditions × 2 pseudowords/no pseudowords conditions × 3 2-gram/3-gram/4-gram conditions). We restricted our workers to those based in the United States with a HIT approval rating of 95% or more.

Table 1. Example study sentences (one feature highlighted)

No.	IV 1: Noun phrase split or not split	IV 2: Pseudoword or no pseudoword	IV 3: 2-gram, 3-gram, or 4-gram	Example
1	Not split	No pseudoword	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 pseudoword	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	Not split	Pseudoword	3-gram	The polysomnogram involves continuous *encming* of **sleep brain waves** and a number of nerve and muscle functions during *ebunt* sleep.
4	Not split	No pseudoword	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 pseudoword	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	Not split	No pseudoword	4-gram	**Lung volume reduction surgery** (LVRS) can improve the quality of life for certain carefully selected patients.

Note. IV = independent variable.

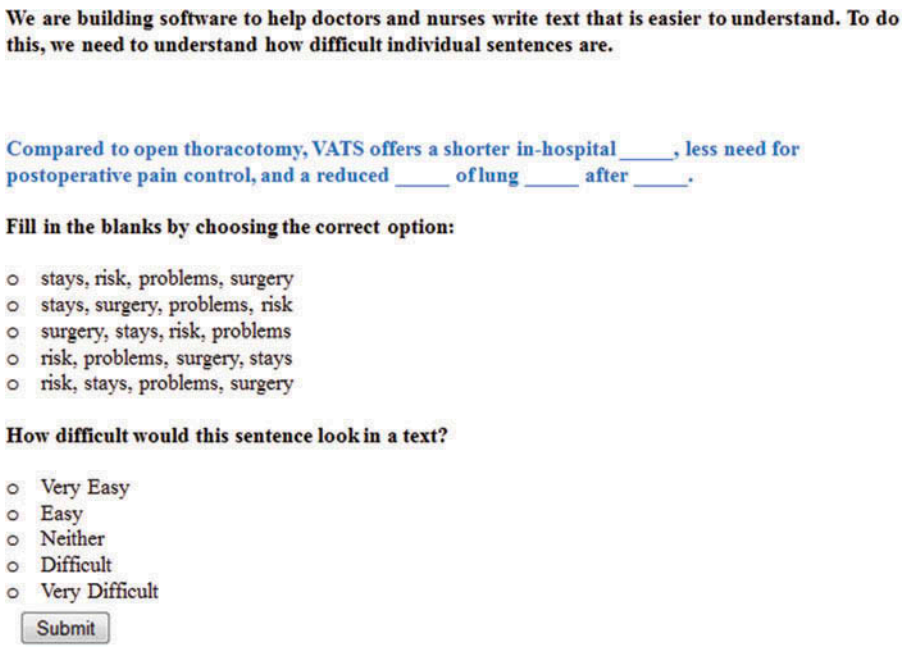

Fig. 1. Example human intelligence task on Amazon's Mechanical Turk.

Our main goal was to evaluate our text manipulation (i.e., the impact of the algorithm), not to study individuals. We therefore analyzed the results as a 2 × 2 × 3 between-subjects design. The four conditions (pseudowords/no pseudowords × split/not split) were presented on four different dates. Each session contained 90 different sentences with 30 sentences for each of the three noun phrase conditions (2-gram/3-gram/4-gram). This ensured that 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 a minimum of 1 week between the four conditions to allow workers to participate in each task if they desired. Because participants were not intentionally memorizing sentences, and exposure to each sentence was limited to a few seconds, 1 week was estimated to be a sufficiently long time to avoid carryover effects for those who did participate in multiple conditions. Each sentence was presented as one HIT, and participants were paid $0.03 per HIT. All conditions were treated as between-subjects conditions because participants could not be forced to complete all 90 sentences and because we allowed sufficient time between the four different test conditions (a week) to ensure no carryover effects existed for those who participated in more than one condition.

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

Results

Participant Demographic Information

The data collection was completed in late Spring 2014 over a 4-week period. A total of 353 people participated. Because MTurk workers can fail to do a 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 (M = 86%, SD = 20%). Because our task was fairly easy, we removed those workers who scored on average 2 SD below average (46% accuracy or lower). As a result, 19 workers as well as the 651 HITs they had completed were removed from the data set. The remaining 334 workers accounted for a total of 11,949 HITs. Workers in this group 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 demographic information for the 334 workers retained in our sample. Overall, there were almost equal numbers of male (52%) and female (48%) participants. Most participants reported not being Hispanic or Latino (93%). The largest racial group was those reporting as White (79%), and the smallest were those reporting as Native Hawaiian/Pacific Islander (1%) and American Indian/Alaska Native (3%). There were almost equal numbers of Asian (8%) and Black (9%) participants. Participants could indicate multiple races, and 5% indicated two or more races.

Because we provided English-language text and aimed to measure comprehension, we asked participants about their 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

Table 2. Participant demographic information (N = 334)

Characteristic	*n*	%
Gender		
Male	175	52
Female	159	48
Ethnicity		
Hispanic or Latino	23	7
Not Hispanic or Latino	311	93
Race (multiple choices allowed; N = 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 a high school diploma	2	1
High school diploma	111	33
Associate's degree	58	17
Bachelor's degree	125	37
Master's degree	31	9
Doctoral degree	7	2
Language spoken at home		
Never English	0	0
Rarely English	0	0
Half English	2	1
Mostly English	14	4
Only English	318	95

education levels were represented. Most participants had earned either a high school diploma (33%) or a 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%).

Main Analyses: Actual and Perceived Difficulty

Because we conducted our study with single sentences and provided multiple-choice answers for the Cloze test, the average accuracy was high (low error) and the differences were 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 analysis of variance for perceived difficulty. The 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 for sentences with phrases that were 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, $F(1, 11937) = 13.334$, $p < .001$.

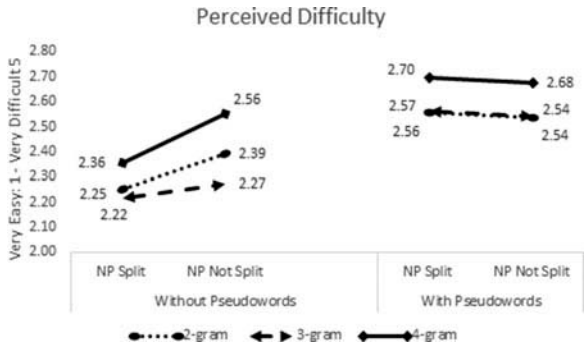

Fig. 2. Perceived difficulty of sentences (higher score = more difficult). NP = noun phrase.

Figure 2 shows that splitting noun phrases did not affect perceived difficulty when pseudowords were 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 as being more difficult.

We then conducted a three-way analysis of variance for actual difficulty. This analysis clearly demonstrated the need to separate 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 the noun phrases were split. This difference was the largest for 4-grams, with the error percentage increasing by an absolute 5% after splitting: 7% error before splitting to 12% error after splitting. As seen in Figure 3, the patterns across different conditions varied, and we found a significant three-way interaction between our independent variables, $F(2, 11937) = 7.098$, $p = .001$. Larger errors were found with 3-grams and 4-grams

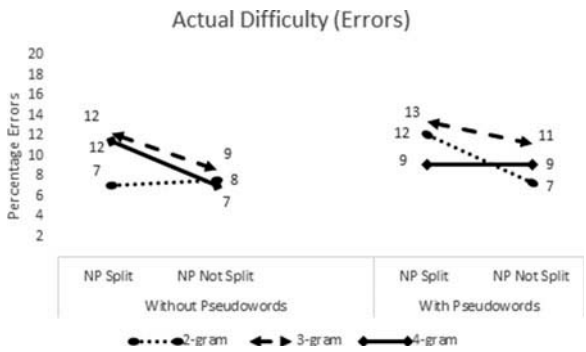

Fig. 3. Actual difficulty of sentences (higher score = more errors). NP = noun phrase.

when there were no pseudowords and for 2-grams and 3-grams when there were pseudowords.

The results also showed a second main effect of noun phrase length, $F(2, 11937) = 9.413$, $p < .001$. However, there was 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 decreased accuracy, and the error was 9.1% for sentences without pseudowords and 10.4% for sentences with pseudowords.

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 longer noun phrases were split into shorter ones. The most commonly used preposition was *of* followed by *in*. These numbers can help to interpret Figure 4 and Figure 5.

Figure 4 and Figure 5 show the perceived and actual difficulty based on the preposition used to split the noun phrase. By looking at the prepositions at this level of granularity, we see some patterns emerge and see that splitting using certain prepositions did result in easier looking text *and* text that was easier to understand. There were several prepositions for which we found improvements. For example, there was a consistent effect when phrases were split 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. A similar pattern emerged for 2-grams with *for* and *in* and 3-grams with *during* and *involving*, and both the perceived and actual difficulty improved after splitting. For other prepositions, the perceived difficulty was not as good an indicator of actual difficulty. For example, splitting a 2-gram using *by* was 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 here 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 versions. 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.

Table 3. Prepositions used to split the noun phrases

	Noun phrase length			
Word used	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

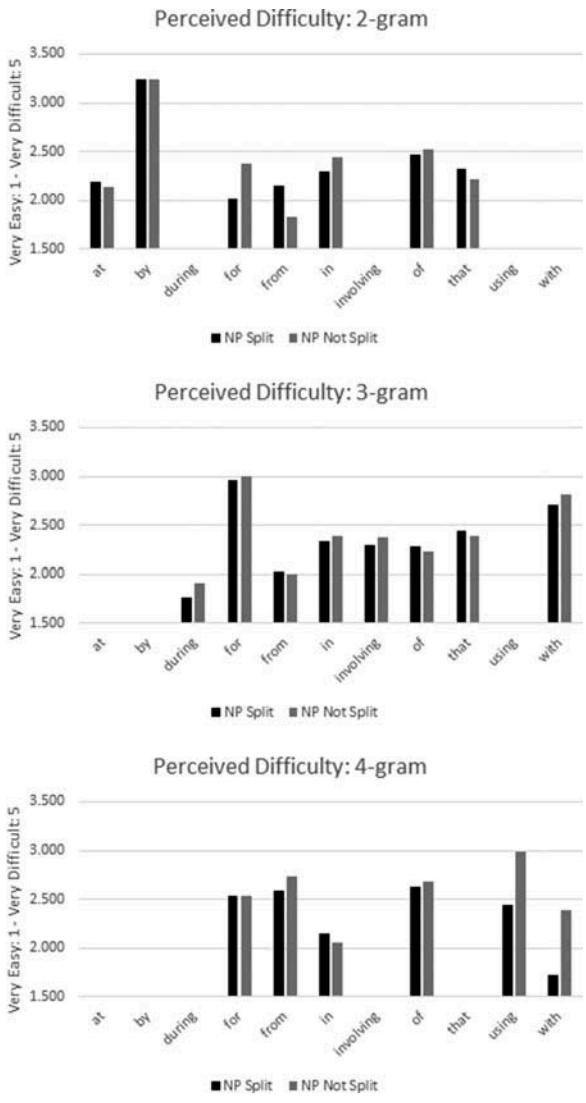

Fig. 4. Perceived difficulty by split word (higher score = more difficult). NP = noun phrase.

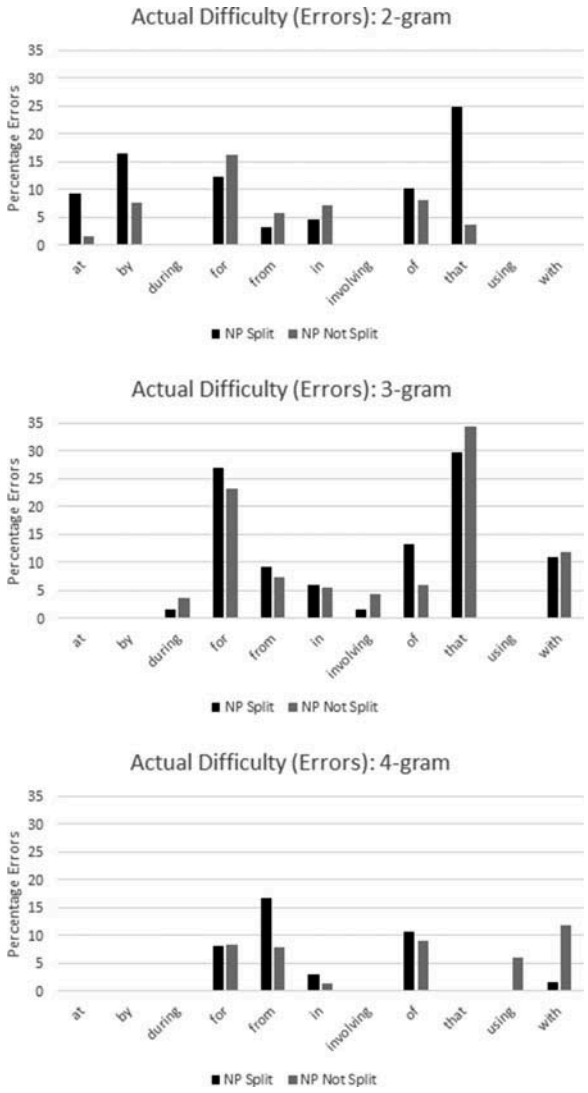

Fig. 5. Actual difficulty by split word (higher score = more errors). NP = noun phrase.

Participants chose whether they preferred the split version, preferred the original version, or did not have a preference. We added four qualification tasks in which the workers were instructed to choose a certain option. This allowed us to remove workers from our data sets who did not take the time to read each task.

For each phrase, we collected evaluations from 15 different participants. The study was completed in the spring of 2015 by 30 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 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, accuracy was higher (fewer errors). Figure 6 shows the scatterplot for the significant correlation.

Discussion

Our overall goal is the development of a text simplification tool in support of providers of health care 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., the 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 semiautomated simplification approaches.

This study demonstrates 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 (Leroy & Endicott, 2012), 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, to our surprise, we found that splitting noun phrases did not increase comprehension of the sentence. Although the sentences with the split noun phrases were perceived as being 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 being split. However, in some cases, a certain awkwardness was introduced that was clear to us. 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 of pseudowords on both perceived and actual difficulty: pseudowords increased difficulty. The effect of the pseudowords seemed to overshadow the effect of splitting the

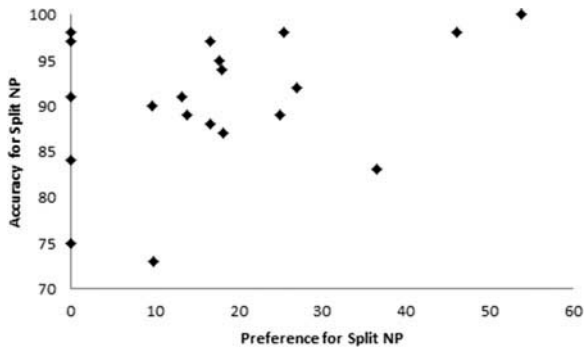

Fig. 6. Scatter plot of Accuracy × Preference for Splitting Phrase. NP = noun phrase.

noun phrases. When pseudowords were present, the perceived difficulty of the sentences remained the same regardless of whether the noun phrase was split.

Implications for practice and future research from the present study include the following:

- 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 what is natural, and to automate the discovery of such phrases. *Natural* was defined here by native English speakers.
- Terms that are not understood (e.g., because 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 to split the noun phrase. We saw some initial patterns in our sample, but a study with a larger set of noun phrases may help make the role of the preposition clearer.

Conclusion

Overall, text difficulty affects how well readers can learn the content of a text. Two aspects of text difficulty are perceived and actual difficulty. We measure both separately in all of 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. We only found evidence that splitting noun phrases increased comprehension in a few cases, frequently when the split noun phrase was identified as sounding more natural than the original.

Acknowledgments

We thank the 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.

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

References

Ahmed, O. H., Sullivan, S. J., Schneiders, A. G., & McCrory, P. R. (2012). Concussion information online: Evaluation of information quality, content and readability of concussion-related websites. *British Journal of Sports Medicine*, *46*(9), 675–683. doi:10.1136/bjsm.2010.081620

Bailey, M. A., Coughlin, P. A., Sohrabi, S., Griffin, K. J., Rashid, S. T., Troxler, M. A., & Scott, D. J. (2012). Quality and readability of online patient information for abdominal aortic aneurysms. *Journal of Vascular Surgery*, *56*(1), 21–26. doi:10.1016/j.jvs.2011.12.063

Cameron, K. A. (2009). A practitioner's guide to persuasion: An overview of 15 selected persuasion theories, models and frameworks. *Patient Education and Counseling*, *74*, 309–317. doi:10.1016/j.pec.2008.12.003

Centers for Disease Control and Prevention. (2011, February). *HIV surveillance report*. Retrieved from http://www.cdc.gov/hiv/pdf/statistics_2011_hiv_surveillance_report_vol_23.pdf

Davis, T., Long, S., Jackson, R., Mayeaux, E., George, R., Murphy, P., & Crouch, M. (1993). Rapid Estimate of Adult Literacy in Medicine: A shortened screening instrument. *Family Medicine*, *25*(6), 391–395.

Didonet, J., & Mengue, S. S. (2008). Drug labels: Are they a readable material? *Patient Education and Counseling*, *73*, 141–145. doi:10.1016/j.pec.2008.05.004

DuBay, W. H. (2004). *The principles of readability*. Retrieved from http://www.impact-information.com/impactinfo/readability02.pdf

Fox, S. (2011). Health topics. *Pew Internet & American Life Project*. Retrieved from http://www.pewinternet.org/2011/02/01/health-topics-2/

Friedman, D. B., Corwin, S. J., Dominick, G. M., & Rose, I. D. (2009). 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*, *34*, 449–460. doi:10.1007/s10900-009-9167-3

Goto, A., Rudd, R. E., Lai, A. Y., & Yoshida-Komiya, H. (2014). Health literacy training for public health nurses in Fukushima: A case-study of program adaptation, implementation and evaluation. *Japan Medical Association Journal*, *57*(3), 146–153.

Graesser, A. C., McNamara, D. S., & Kulikowich, J. M. (2011). Coh-Metrix: Providing multilevel analyses of text characteristics. *Educational Researcher*, *40*, 223–234. doi:10.3102/0013189X11413260

Hendrick, P. A., Ahmed, O. H., Bankier, S. S., Chan, T. J., Crawford, S. A., Ryder, C. R., & Schneiders, A. G. (2012). Acute low back pain information online: An evaluation of quality, content accuracy and readability of related websites. *Manual Therapy*, *17*(4), 318–324. doi:10.1016/j.math.2012.02.019

Institute of Medicine. (Ed.). (2004). *Health literacy: A prescription to end confusion*. Washington, DC: National Academies Press.

Keselman, A., & Smith, C. A. (2012). A classification of errors in lay comprehension of medical documents. *Journal of Biomedical Informatics*, *45*, 1151–1163. doi:10.1016/j.jbi.2012.07.012

Keuleers, E., & Brysbaert, M. (2010). Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, *42*(3), 627–633. doi:10.3758/BRM.42.3.627

Kittur, A., Chi, E. H., & Suh, B. (2008). Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 452–456). New York, NY: Association for Computing Machinery.

Kwak, M., Leroy, G., Martinez, J. D., & Harwell, J. (2013). Development and evaluation of a biomedical search engine using a predicate-based vector space model. *Journal of Biomedical Informatics*, *15*(7), e144.

Lam, C. G., Roter, D. L., & Cohen, K. J. (2013). Survey of quality, readability, and social reach of websites on osteosarcoma in adolescents. *Patient Education and Counseling*, *90*, 82–87. doi:10.1016/j.pec.2012.08.006

Lease, M., Hullman, J., Bigham, J. P., Bernstein, M. S., Kim, J., Lasecki, W., & Miller, R. C. (2013). Mechanical Turk is not anonymous. *Social Science Research Network*. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2228728

Leroy, G., & Endicott, J. E. (2011, October). *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, Beijing, China.

Leroy, G., & Endicott, J. E. (2012, January). *Combining NLP with evidence-based methods to find text metrics related to perceived and actual text difficulty*. Paper presented at the Second ACM SIGHIT International Health Informatics Symposium, Miami, FL.

Leroy, G., Endicott, J. E., Kauchak, D., Mouradi, O., & Just, M. (2013). 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*, *15*(7), e144. doi:10.2196/jmir.2569

McLaughlin, G. H. (1969). SMOG grading: A new readability formula. *Journal of Reading*, *12*, 636–646.

Meade, C. D., & Smith, C. F. (1991). Readability formulas: Cautions and criteria. *Patient Education and Counseling*, *17*, 153–158. doi:10.1016/0738-3991(91)90017-Y

Mouradi, O., Leroy, G., Kauchak, D., & Endicott, J. E. (2013, January). *Influence of text and participant characteristics on perceived and actual text difficulty*. Paper presented at the Hawaii International Conference on System Sciences, Maui, HI.

Mullan, J., Crookes, P. A., & Yeatman, H. (2003). Rain, fog, smog and printed educational material. *Journal of Pharmacy Practice and Research*, *33*(4), 284–286.

Norman, C., & Skinner, H. (2006). eHEALS: The eHealth Literacy Scale. *Journal of Medical Internet Research*, *8*(4), e27. doi:10.2196/jmir.8.4.e27

Nurss, J. R., Parker, R. M., Williams, M. V., & Baker, D. W. (1995). *Test of Functional Health Literacy in Adults*. Hartford, MI: Peppercorn Books & Press.

Ogden, C. L., Carroll, M. D., Kit, B. K., & Flegal, K. M. (2012). *Prevalence of obesity in the United States, 2009–2010* (NCHS Data Brief No. 82). Hyattsville, MD: National Center for Health Statistics.

Pichert, J. W., & Elam, P. (1985). Readability formulas may mislead you. *Patient Education and Counseling*, *7*(2), 181–191. doi:10.1016/0738-3991(85)90008-4

Pignone, M., DeWalt, D. A., Sheridan, S., Berkman, N., & Lohr, J. N. (2005). Interventions to improve health outcomes for patients with low literacy: A systematic review. *Journal of General Internal Medicine*, *20*(2), 185–192. doi:10.1111/j.1525-1497.2005.40208.x

Polishchuk, D. L., Hashem, J., & Sabharwal, S. (2012). Readability of online patient education materials on adult reconstruction Web sites. *Journal of Arthroplasty*, *27*(5), 716–719. doi:10.1016/j.arth.2011.08.020

Ross, J., Irani, L., Silberman, M., Zaldivar, A., & Tomlinson, B. (2010). *Who are the crowdworkers? Shifting demographics in Mechanical Turk*. In *CHI EA '10: CHI '10 Extended Abstracts on Human Factors in Computing Systems* (pp. 2863–2872). New York, NY: Association for Computing Machinery,

Safran, N. (2012). *Wikipedia in the SERPs*. Retrieved 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. (2002, July). *Resolving attachment and clause boundary ambiguities for simplifying relative clause constructs*. Paper presented at the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, PA.

Tanaka, S., Jatowt, A., Kato, M. P., & Tanaka, K. (2013, February). *Estimating content concreteness for finding comprehensible documents*. Paper presented at the Sixth ACM International Conference on Web Search and Data Mining, Rome, Italy.

Taylor, W. L. (1953). Cloze procedure: A new tool for measuring readability. *Journalism Quarterly*, *30*, 415–433.

Vallance, J. K., Taylor, L. M., & Lavallee, C. (2008). Suitability and readability assessment of educational print resources related to physical activity: Implications and recommendations for practice. *Patient Education and Counseling*, *72*, 342–349. doi:10.1016/j.pec.2008.03.010

Wang, L.-W., Miller, M. J., Schmitt, M. R., & Wen, F. K. (2013). Assessing readability formula differences with written health information materials: Application, results, and recommendations. *Research in Social & Administrative Pharmacy*, *9*(5), 503–516.

Wang, Y. (2006, June). *Automatic recognition of text difficulty from consumers health information*. Paper presented at the 19th IEEE International Symposium on Computer-Based Medical Systems, Salt Lake City, UT.

Weiss, B. D. (2007). *Health literacy and patient safety: Help patients understand. Manual for clinicians* (2nd ed.). Chicago, IL: American Medical Association and American Medical Association Foundation.

Wubben, S., Van den Bosch, A., & Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *ACL '12: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Paper* (Vol. 1, pp. 1015–1024). Stroudsburg, PA: Association for Computational Linguistics.

Zaidan, O. F., & Callison-Burch, C. (2011). Crowdsourcing translation: Professional quality from non-professionals. In HLT '11: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol.1, pp. 1220–1229). Stroudsburg, PA: Association for Computational Linguistics.

Zarcadoolas, C. (2011). The simplicity complex: Exploring simplified health messages in a complex world. *Health Promotion International*, *26*(3), 338–350. doi:10.1093/heapro/daq075