SLPAT 2016 Workshop on Speech and Language Processing for Assistive Technologies
13 September 2016, San Francisco, USA

The Effect of Semantic Difference on Non-expert Judgments of Simplified Sentences

Sven Anderson, S. Rebecca Thomas, Ki Won Kwon, Wayne Zhang

Bard College
Annandale-on-Hudson, NY 12504 USA
sanderson@bard.edu,thomas@bard.edu

Abstract

Advances in text simplification depend on reliable judgments of sentence difficulty. The ability of untrained native English speakers to judge sentence difficulty in the presence of variation in semantic similarity is examined using cloze tests and a forced-choice comparison task. Judgments from participants in web-based experiments demonstrate ability to assess sentence difficulty of professionally leveled sentence pairs with 84% accuracy. The comparison task results suggest that participants' ability to judge comparative sentence difficulty is inversely related to semantic similarity; that is, contrary to our intuition, speakers appear more accurate at judging sentence difficulty for sentences that are dissimilar than for those that are similar.

1. Introduction

Text simplification aims to expand access to textual information by algorithmically reducing the reading level of text – ideally without changing its meaning. Understanding text simplification can help us to design or customize tools for Augmentative and Alternative Communication (AAC) users, language learners, and other populations who need better access to information including the wealth of information available on the web.

Many approaches to simplification use statistical machine learning which depends on large corpora of text at different levels. Of particular use, though still quite limited in size, are bi-text corpora; these corpora present the same ideas in standard/simplified text pairs and have usually been created by aligning sentences from longer text, e.g. [1]. Unfortunately, the paucity of simplified text data has hindered most efforts to optimize machine learning approaches to the problem [2]. This led us to consider whether a necessary prerequisite for the identification of simplified texts, namely reading level judgments, might be more readily obtained from untrained native speakers using crowd-sourcing methods such as Amazon's Mechanical Turk.

Although professionals who create leveled readers are able to write sentences, paragraphs, and entire book collections at a pre-specified reader grade level, it is unclear whether untrained native speakers can meaningfully assess text difficulty, particularly of short texts. In addition, when text is simplified, its content is necessarily transformed, and judgments that compare the levels of two similar sentences may be confounded by differences that go beyond complexity differences, including substantial semantic and syntactic variation unrelated to simplification. Given the difficulty of the task, it appears doubtful that untrained readers can reliably assign grade levels to texts; however, such readers might be able to determine relative difficulty of two sentences. If so, a comparison-based task might enable a crowd-source approach much like what underlies test similarity tasks like SemEval [3].

In this paper we report on experiments that seek to discover whether native speakers can compare two sentences and judge when one is "more difficult" than another, despite other differences in syntax and semantics. We expected to find that more similar sentences, having fewer lexical and syntactic differences, thereby allow greater focus on the features relevant to the task and thus more accurate assessment of sentence difficulty. Leroy et al. [4] assume that explicit judgments of difficulty are possible, and refer to these explicit judgments, collected from a population of participants, as "perceived difficulty," which they distinguish from "actual difficulty." Like these researchers we use cloze measure scores to estimate "actual difficulty," but recognize that this simply measures the predictability of words given sentence contexts. We bring these two measures together by comparing cloze scores with comparative judgments.

2. Related Work

Text simplification improves accessibility of written or spoken language by transforming it to better meet the needs of the reader, including those with cognitive and/or linguistic challenges. The primary approaches to text simplification, all of which remain open research problems, include lexical and syntactic simplification, machine translation, and explanation generation [5]. For example, lexical simplification has been used to reduce the size of an

10.21437/SLPAT.2016-10

article's vocabulary set to a smaller size that could determine which icons to use in communication boards [6]. Text simplification also applies to spoken language [7], and even non-linguistic media.

The present study focuses on the simplification of single sentences. It is not self-evident that this is the best level at which to study simplification: leveled readers, for example, are often simplified at the document level with transformations that restructure the linguistic presentation of ideas across paragraphs or sections. However, several considerations suggest the sentence as a viable unit of simplification. First, the sentence is a linguistic entity that conveys a complete proposition, and it is shared by all human languages. Sentence-level simplification involves both lexical and syntactic transformations, and even studies focused on word- or phrase-level transformations usually rely on delivery within a sentence context. Finally, simplification approaches derived from statistical machine translation depend on sentences aligned with their simplified counterparts, providing a basis for machine learning [1, 8].

A major challenge to the development of usable simplification is a clearer understanding of the factors that influence perception of text difficulty. That is, the creation of reliable automated text simplification algorithms depends on being able to detect whether text has been simplified! Lasecki [9] demonstrated that judgments from untrained native speakers can be used to assess sentence level along a 7-point Likert scale. In this case, the authors assumed that sentence simplicity is correlated with the number of simplifying transformations applied to a complex sentence. Many other studies also rely on comparing sentences that are semantically related, perhaps derived from the same original complex sentence. For example, [10] demonstrated a positive effect of lexical simplification using pairs of lexically simplified sentences. One objective of our study was to determine the degree to which semantic similarity affects simplicity judgments, since text simplification often introduces significant semantic changes.

3. Background

This section provides background concerning three topics central to our methods: measures of readability, empirical measures of difficulty, and algorithmic measures of sentence similarity. To contextualize this discussion, consider a sample task from our comparison experiment shown in Figure 1. In these sentence pairs the first pair is considered semantically similar and second pair dissimilar. (Additional sentence pairs from our experiments appear in the appendix.) Each pair of sentences is drawn from texts at different author-identified reading levels and cloze scores were used as an alternate measure of actual reading level.

3.1. Measures of Reading Level

Measuring the reading grade level and readability of text is important to writers of educational materials. Dubay [11] reports that from 1940 to the 1980's, approximately 200 readability formulas appeared in the literature. Many of these formulas use simple surface measures such as word count, word length, syllable count, average syllables per word, etc. to estimate sentence readability. As one example, the widely-used Flesch-Kincaid Grade level score is given by

$$\text{grade} = 0.39\frac{n_w}{n_s} + 11.8\frac{n_\sigma}{n_w} - 15.59$$

where n_s, n_w, and n_σ are the number of sentences, words, and syllables, respectively. The result is intended to be interpreted as a grade level; thus a text written for beginning readers would score roughly 1, with more complex texts assigned higher scores.

Most formulas are intended to measure the level of longer passages, not single sentences. Fry [12] notes that most formulas require at least 300 words. He proposes a formula for 40-99 words of three or more sentences, but observes that for shorter texts this formula is unreliable. The feasibility of determining readability based exclusively on surface level measures is limited, leading some researchers to explore models that incorporate semantic content.

3.2. The Cloze Measure as Actual Difficulty

The cloze measure estimates text difficulty by relating it to the ease with which a missing word can be guessed from its context [13], usually a text of several paragraphs. In most cloze experiments every Nth word – where N is generally 5-7 – is replaced by a blank. When human subjects guess the missing word with roughly 60% accuracy or greater, the text is considered relatively easy. Human prediction of a word based on surrounding context is reminiscent of the goal of language models based on n-grams: both human and algorithm rely on context to predict a most likely word. Smith and Levy [14] compared a 5-gram continuation (cloze) task in which participants completed a short phrase with corpus statistics derived from Web 1T and scanned books. They found that cloze scores varied substantially from corpus statistics.

One challenge in this project is how to assess the "true" difficulty of a sentence, against which to compare human perceptions or ratings. Leroy et al. [4] differentiates implicit tests of text difficulty based on cloze tests or tests of comprehension from explicit reports based on comparison or Likert-scale judgments by participants. The authors call the former *actual difficulty* and the latter *perceived difficulty*. We adopt the same terminology in this paper, acknowledging that the cloze score is merely our best approximation to actual difficulty.

Similar	S: North of Cairo, Egypt, the Nile enters the region called the delta. C: North of Cairo the Nile enters the delta region, a level triangular lowland. SEMILAR score: 0.883
Dissimilar	S: Fish come from the sea or from fish farms. C: Small clustered fishing villages are found along the coastline. SEMILAR score: 0.349

Figure 1: Example sentences; in each pair, the simpler sentence is marked with S. See appendix for more.

3.3. Semantic Similarity of Sentences

We employed the SEMILAR Toolkit [15] to measure the semantic similarity of sentence pairs. We measured the semantic similarity of each sentence pair using SEMILAR, with word similarity measured by LSA using the TASA model provided with the SEMILAR download. This variant of SEMILAR was selected based on empirical tests of several similarity measures as applied to three datasets: O'Shea [16], Sem* STS 2012 [17], and Sem* STS 2013 [3]; it scored the closest to human judgments for all three. The sentence pairs used in this study were selected to be of similar length. Thus sentence pairs with high semantic similarity are likely to share numerous similar word pairs, whereas sentences with low semantic similarity are likely to have a large number of word-to-word differences.

4. Methods

The sentences used for all experiments were drawn from Brittanica School articles. Topic-matched articles from Level 1 (elementary school level) and Level 3 (high school level) were mined to find sentence pairs differing markedly in difficulty, one drawn from each level. Aligned sentences extracted from these articles were retained only if their lengths (in words) differed by less than 20% to avoid confounding effects. For each topic, at most one sentence pair was retained in order to avoid semantic contamination across different sentence pairs. Sentence pairs were chosen to provide a group of highly similar and highly dissimilar sentences. The SEMILAR score falls in the range of 0.0 to 1.0, with 1.0 meaning identical sentences and 0.0 meaning no similarity. In order to facilitate comparing results on high similarity pairs vs. low similarity pairs, sentence pairs were ranked according to SEMILAR scores; the pairs falling in the 85-87.5th percentile and the 12.5-15th percentile were used in these experiments. The low-similarity pair SEMILAR scores ranged from 0.0 to 0.36; the high-similarity pair SEMILAR scores ranged from 0.83 to 0.97.

All experiments were run using a custom, web-based LimeSurvey [18] redirected from Amazon Mechanical Turk. Participants were paid a small amount for completing the experiment. The first two questions were "dummy" questions, with responses collected but not analyzed. The third question was designed with an unambiguous correct answer, and any participant who answered this incorrectly was not allowed to proceed with the experiment, to eliminate inattentive subjects. Demographic information was collected, and participants were told that only native English speakers could participate. In addition, each participant was asked about current English usage, e.g. whether they primarily or exclusively spoke English in their home.

5. Cloze Experiment

A total of three paired cloze experiments were run, making a total of six different sets of sentences, each of which was completed by twenty participants. Each experiment pair used the same set of 38 sentence pairs, which were taken from topic matched, level-differing articles as described above. The sentence pairs were divided into tasks A and B, such that for each sentence pair, the Level 1 sentence was randomly assigned to either task A or B, and the Level 3 sentence assigned to the other. Every seventh word in each sentence was deleted, starting at a specified position in the sentence: the fourth in the first experiment pair, the third in the second experiment pair, and the fifth in the third experiment pair. Participants provided responses to only one of the six different tasks and sentence order was randomized for each participant. A participant's response was considered correct only if it was a case-insensitive exact match. In total, participants produced 8,160 individual cloze responses, approximately 68 words per participant.

5.1. Cloze Results

Our sentences have an average of 12.8 words; a sentence will have one to four blank cloze positions on any trial. The proportion of a sentence's blanks correctly filled in by all participants, its cloze score, is highly variable, since scores depend largely on whether the cloze position happens to fall on an easily guessed word. To overcome this source of noise we average scores for each sentence over the three different blank positions and call this the *average cloze score*. The distribution of average cloze scores for all sentences shown in Figure 2 suggests that sentence difficulty is quite variable ($\overline{x} = 0.42$; $\sigma = 0.17$).

The original sentences were drawn from texts at elementary (Level 1) and high-school (Level 3) grade levels. The average cloze scores of sentences drawn from texts

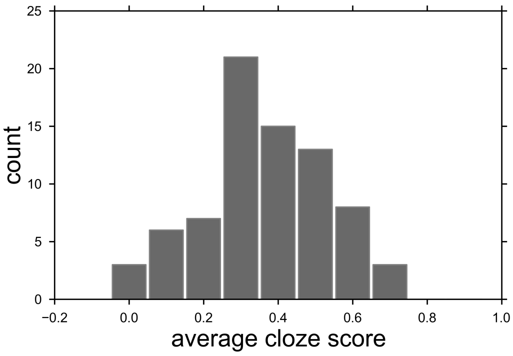

Figure 2: Histogram of average cloze scores for sentences. Averages are over three different initial positions in each sentence, with that and every seventh subsequent word deleted.

at the elementary (49.3%) versus those at the high-school level (35.2%) are significantly different ($p < 0.0001$; $t = 4.05$), reflecting agreement between professional leveling of the articles and sentence difficulty as estimated by the cloze measure.

Given this result, we might ask whether other measures, such as unigram frequency or common readability measures show greater agreement with professional judgments of reading level than cloze scores do. To explore the utility of unigram frequency, word probability for each cloze word was estimated from the Web1T corpus. Unigram probabilities and average cloze scores have a correlation of $r = 0.506$, indicating some shared information. We then evaluated the ability of a logistic regression model to predict reading level based on both average cloze scores and unigram probabilities. Only the average cloze score is a significant ($p < 0.001$) predictor of level, suggesting that average cloze scores are superior predictors compared to unigram frequency.

The readability measures we examined, Flesch-Kincaid grade level and Lexile Score, do not generate significantly different values for Level 1 versus Level 3 sentences. The Flesch-Kincaid grade level has often been used to measure text difficulty, even in single sentences [4, 10]. Average Flesch-Kincaid grade levels are 7.2 and 8.6 for the Level 1 and 3 sentences, respectively; however, this difference does not reach significance ($p = 0.11$). This result is consistent with findings of Leroy et al. [10] who found that Flesch-Kincaid grade level did not differ significantly for sentence pairs that had been lexically simplified. We obtained similar results when using Lexile scores; we note that Lexile is intended for much longer texts. In this case average Lexile Scores [19] were 897.9 and 932.9 for Levels 1 and 3, respectively. The scores, grouped by reading level, were not significantly different ($p = 0.55$). Thus cloze scores better predict reading level than either unigram frequency or these reading level measures.

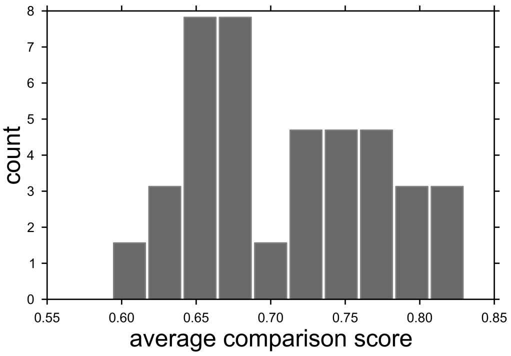

Figure 3: Average comparison accuracy per participant. Comparison is correct if Level 1 is judged simpler than Level 3.

6. Comparison Experiment

The sentence comparison task tests whether participants are able to compare two sentences and determine which is simpler, despite variation in semantic similarity. Each of the 38 randomly ordered sentence pairs was presented in turn to the participant. The order in which the Level 1 and Level 3 sentences were presented was also randomized as the experiment was run. The forced-choice task requires participants to select the less difficult of two sentences.

6.1. Comparison Results

Using author-assigned text level to measure sentence difficulty, 27 of 29 participants responded above 60% correct. The scores of two participants were markedly lower outliers, within 5% of chance; thus, their results are removed from subsequent analysis. The mean score for the remaining 27 participants was 72.1% (Figure 3).

Using author-assigned text level as the measure of actual sentence difficulty, the per-sentence comparison accuracy results are shown in Figure 4. 84.2% (32/38) of sentence pairs had comparison judgments consistent with the author-assigned reading level; that is, the Level 1 sentence was more often labeled simpler than the Level 3 sentence. If instead we use the average cloze scores as the measure of actual sentence difficulty, only 76.3% (29 of 38) of comparison judgments match.

There are nine sentence pairs for which participants had better cloze accuracy on the Level 3 sentence than on the Level 1 sentence, which is not the expected result. This may, in part, be an artifact of having used only three of the possible seven initial positions for deleting words; in some sentences the deleted words may have been the hardest or easiest to guess. Another potential explanation is that the reading level of individual sentences may not always reflect the reading level of an entire text; a com-

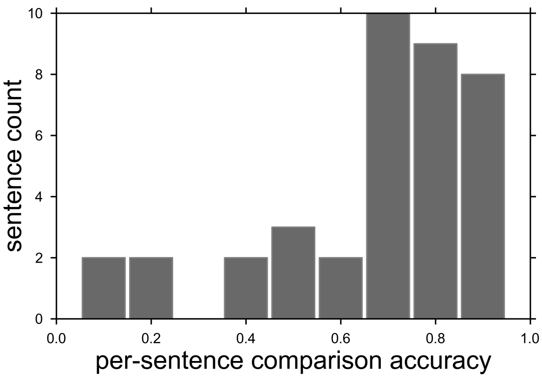

Figure 4: Histogram of per-sentence average comparison accuracy.

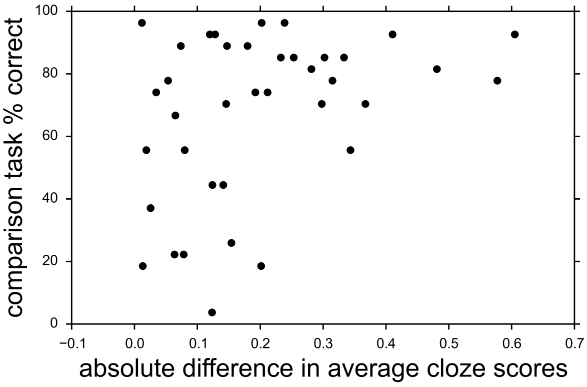

Figure 5: Scatter plot of percent correct perceived difficulty vs. absolute difference in average cloze.

plex text may contain relatively simple individual sentences. In selecting our experimental data, we assumed that sentences drawn from a text reflected the level of that text. The variation of average cloze scores and the variation in comparison scores both suggest that text level and sentence level are not in complete agreement.

6.2. Combined Results

The primary focus of this study was to determine whether the comparative difficulty of sentences can be assessed without regard to semantic or syntactic differences. The sentence pairs in this study belong to one of two groups: high similarity or low similarity. We compared the degree to which comparison scores agreed with author-assigned reading level for the high similarity (67.7%) versus low similarity (77.0%) groups; a t-test indicates that the comparison scores for these groups are not significantly different ($p = 0.158$; $t = -1.44$). That is, the ability of participants to compare sentence level was not significantly affected by the semantic and syntactic differences between those sentence pairs. However, when we compared the degree to which comparison scores agreed with comparative average cloze scores, the high similarity pairs average 58.4% versus 75.7% for low similar sentence pairs (one-tailed, $t = -2.12$, $p = 0.02$), reaching significance. Surprisingly, whether we assess actual sentence difficulty based on average cloze score or on author-assigned level, lower semantic similarity appears to aid the ability to judge comparative sentence difficulty.

Although not a simple linear relationship, the relation between perceived and actual difficulty shown in Figure 5 suggests that the ability to judge differences in sentence difficulty improves as cloze difference between two sentences increases.

Other factors that have been shown to be related to perceived difficulty include function word density, the occurrence of difficult words as measured by Dale-Chall, and noun-phrase complexity [4]. We measured noun-phrase complexity using the maximal depth of all noun phrases in a sentence. Function word density was based on English stopwords listed in NLTK. A multiple regression test including these variables, average cloze scores, and ratios of cloze scores was performed. The overall adjusted R-squared value was 0.17 ($p = 0.12$). The only significant variable was the Dale-Chall score of the difficult sentence ($p = 0.05$). The average cloze score of the Level 1 sentence had the next smallest p-value ($p = 0.10$). We have found no variables that are significant predictors of perceived difficulty scores.

7. Discussion

One goal of this study was to determine whether sentence level judgements of untrained native speakers could be used to replace professional judgments of sentence difficulty. Experience with simplifications in the Simple English Wikipedia suggests that non-professional simplification is much less reliable than that obtained from professionals [2]. However, [9] showed that crowd-source methods could accurately rate the number of simplifying transformations made to single complex sentences. These results, from a slightly smaller set of sentences, are most directly comparable to the results reported above. Note that in [9] sentence simplicity was based on a count of transformations to a base sentence and did not incorporate the wide variation in semantic similarity we investigate here. Our cloze results support the claim in [9] that untrained speakers can implicitly judge sentence difficulty in agreement with expert judgments, but imperfectly so. Moreover, perceived difficulty, as measured by the sentence comparison task somewhat similar to that in [9], is not clearly related to cloze scores on the same sentences. This suggests that actual and perceived difficulty do not measure identical sentence attributes. Additionally, neither directly represents the benefit in comprehension that such a simplification might afford the reader. The best metric for measuring text difficulty depends on the task, and the field has not settled on a clear methodology for such measures.

Our unexpected finding that comparing sentence difficulty is improved by semantic dissimilarity is one that must be more carefully studied. Semantic priming presents one possible explanation. By this account, when participants are presented with a sentence pair, the second sentence may seem easier because the first sentence has semantically primed the participant for the second sentence. Of course, this is likely to be a much stronger effect for semantically similar sentences. If so, we would predict that there may be a consistent bias toward underestimating the complexity of the second sentence in high-similarity pairs. Unfortunately, our experimental design was such that the order in which the two sentences were presented was randomized for each participant at run time, and the order of presentation was not recorded. Thus the experimenters cannot assess whether such a bias actually was observed.

7.1. Crowd-sourced Effort to Judge Difficulty

One advantage of both the cloze procedure and the forced-choice comparison task is that they can be undertaken by readers of a language with very little training and are therefore suitable for crowd-sourced data collection. The comparison task requires a single decision that ranks one sentence relative to another. By contrast, the cloze procedure requires multiple lexical inputs, but it yields a percentage cloze score that provides a total order for all sentences. Which approach is best?

In cases where experimenters desire a total order over a particular set of sentences, a simple analysis provides a means to compare the effort required for each of these two methods. If we have n sentences of average length m, the comparison task is essentially sorting by binary decisions and will provide a total order of the sentences with $nlog(n)$ comparisons. Assuming we must provide blanks on approximately half of the words to estimate the cloze score, the cloze Procedure will require $\frac{m}{2} \times n$ blanks to be completed by a set of participants. Thus, the number of inputs provided by participants is numerically similar when $log(n) = \frac{m}{2}$. For sentences of about 20 words, this implies $n \approx 1024$. That is, the comparison procedure will require fewer human decisions until approximately 1000 sentences are to be completely ordered. Note that this ignores the relative difficulty of, and time required to make, comparison vs. word-choice decisions.

In this study, the cloze procedure provided a total order using about 210 word choices per participant, but took significantly more effort for both participants and the experimenters. In contrast, the comparison task required only 38 comparisons per participant but was not designed to provide a total order. If a total order is desired, further comparison tasks might be generated for additional test subjects, guided by an $O(nlog(n))$ sorting algorithm such as mergesort.

7.2. Conclusion

The methodology by which sentence difficulty is measured has direct consequences for the creation of corpora (e.g., [20, 8, 1, 10]) underlying future development of text simplification. The genesis of this project was the authors' awareness of the need for more bi-text data at different levels of reading complexity, which might hypothetically be used to train machine translation systems to perform text simplification, or to train systems that measure text readability. Naive users could presumably judge relative readability more easily and more quickly than they could perform cloze exercises. If their judgments were sound, then one might use crowdsourcing to efficiently sort sentences by readability. Our results indicate that there is no need to match the sentences by content or even by topic; in fact, it appears be an advantage not to do so. Future research should clarify whether this is more generally true.

8. Appendix: Sample sentence pairs

Sample high-similarity pairs

S: The two openings in the nose are called nostrils.
C: The external openings are known as nares or nostrils.
SEMILAR score: 0.828

S: Northern Ireland is often called Ulster because it includes six of the nine counties that made up the ancient kingdom of Ulster.
C: Northern Ireland is sometimes referred to as Ulster, although it includes only six of the nine counties which made up that historic Irish province.
SEMILAR score: 0.869

S: Four main aerodynamic forces act on an airplane in flight.
C: An aircraft in straight-and-level unaccelerated flight has four forces acting on it.
SEMILAR score: 0.843

Sample low-similarity pairs

S: Guglielmo Marconi was an Italian scientist and inventor.
C: Marconi's great triumph was, however, yet to come.
SEMILAR score: 0.010

S: Unlike many plants, cacti do not have deep roots.
C: The fruit is usually a berry and contains many seeds.
SEMILAR score: 0.013

S: In nearly all mammals, the female carries the developing young in her body after mating.
C: The winter dormancy of bears at high latitudes is an

analogous phenomenon and can not be considered true hibernation.
SEMILAR score: 0.113

9. References

[1] W. Coster and D. Kauchak, "Simple English wikipedia: a new text simplification task," in *Proc. of the 49th Association for Computational Linguistics*, 2011, pp. 665–669.

[2] W. Xu, C. Callison-burch, and C. Napoles, "Problems in current text simplification research : New data can help," *Transactions of the ACL*, vol. 3, pp. 283–297, 2015.

[3] E. Agirre, D. Cer, M. Diab, A. Gonzalez-Agirre, and W. Guo, "SEM 2013 shared task: Semantic textual similarity, including a pilot on typed-similarity," in *In *SEM 2013: The Second Joint Conference on Lexical and Computational Semantics. Association for Computational Linguistics*, 2013.

[4] G. Leroy, S. Helmreich, and J. R. Cowie, "The influence of text characteristics on perceived and actual difficulty of health information," *International journal of medical informatics*, vol. 79, no. 6, pp. 438–49, 2010.

[5] M. Shardlow, "A survey of automated text simplification," *International Journal of Advanced Computer Science and Applications*, vol. 4, no. 1, pp. 58–70, 2014.

[6] S. Anderson, S. Thomas, C. Segal, and Y. Wu, "Automatic reduction of a document-derived noun vocabulary," in *Twenty-Fourth International FLAIRS Conference*, 2011.

[7] D. J. Higginbotham, G. W. Lesher, B. J. Moulton, and B. Roark, "The application of natural language processing to augmentative and alternative communication," *Assistive Technology*, vol. 24, no. 1, pp. 14–24, 2012.

[8] Z. Zhu, D. Bernhard, and I. Gurevych, "A monolingual tree-based translation model for sentence simplification," in *Proceedings of the 23rd International Conference on Computational Linguistics*, 2010, pp. 1353–1361.

[9] W. S. Lasecki, L. Rello, and J. P. Bigham, "Measuring text simplification with the crowd," *Proceedings of the 12th Web for All Conference (W4A '15)*, pp. 4:1—4:9, 2015.

[10] G. Leroy, D. Kauchak, and O. Mouradi, "A user-study measuring the effects of lexical simplification and coherence enhancement on perceived and actual text difficulty," *International journal of medical informatics*, vol. 82, no. 8, pp. 717–30, Aug. 2013.

[11] W. DuBay, "The principles of readability," *Costa Mesa: Impact Information*, 2004.

[12] E. Fry, "A readability formula for short passages," *Journal of Reading*, vol. 33, no. 8, pp. 594–597, 1990.

[13] W. L. Taylor, "Cloze procedure: a new tool for measuring readability." *Journalism and Mass Communication Quarterly*, vol. 30, no. 4, p. 415, 1953.

[14] N. J. Smith and R. Levy, "Cloze but no cigar: The complex relationship between cloze, corpus, and subjective probabilities in language processing," *Proceedings of the 33rd Annual Meeting of the Cognitive Science Conference*, pp. 1637–1642, 2011.

[15] V. Rus, M. C. Lintean, R. Banjade, N. B. Niraula, and D. Stefanescu, "Semilar: The semantic similarity toolkit." in *ACL*, 2013, pp. 163–168.

[16] J. O'Shea, Z. Bandar, K. Crockett, and D. McLean, "Benchmarking short text semantic similarity," *International Journal of Intelligent Information and Database Systems*, vol. 4, no. 2, p. 103, 2010.

[17] E. Agirre, M. Diab, D. Cer, and A. Gonzalez-Agirre, "Semeval-2012 task 6: A pilot on semantic textual similarity," in *Proc. First Joint Conference on Lexical and Computational Semantics*, 2012, pp. 385–393.

[18] C. Schmitz, *LimeSurvey:An Open Source Survey Tool*, 2015, http://www.limesurvey.org.

[19] A. J. Stenner, H. Burdick, E. E. Sanford, and D. S. Burdick, "The Lexile framework," Durham, NC: MetaMetrics, Tech. Rep., 2007.

[20] R. Chandrasekar and B. Srinivas, "Automatic induction of rules for text simplification," *Knowledge-Based Systems*, vol. 10, no. 3, pp. 183–190, 1997.