Measuring Text Simplification with the Crowd

Walter S. Lasecki
ROC HCI
Dept. of Computer Science
University of Rochester
wlasecki@cs.rochester.edu

Luz Rello
HCI Institute
School of Computer Science
Carnegie Mellon University
luzrello@cs.cmu.edu

Jeffrey P. Bigham
HCI and LT Institutes
School of Computer Science
Carnegie Mellon University
jbigham@cs.cmu.edu

ABSTRACT

Text can often be complex and difficult to read, especially for people with cognitive impairments or low literacy skills. Text simplification is a process that reduces the complexity of both wording and structure in a sentence, while retaining its meaning. However, this is currently a challenging task for machines, and thus, providing effective on-demand text simplification to those who need it remains an unsolved problem. Even evaluating the simplicity of text remains a challenging problem for both computers, which cannot understand the meaning of text, and humans, who often struggle to agree on what constitutes a good simplification.

This paper focuses on the evaluation of English text simplification using the crowd. We show that leveraging crowds can result in a collective decision that is accurate and converges to a consensus rating. Our results from 2,500 crowd annotations show that the crowd can effectively rate levels of simplicity. This may allow simplification systems and system builders to get better feedback about how well content is being simplified, as compared to standard measures which classify content into 'simplified' or 'not simplified' categories. Our study provides evidence that the crowd could be used to evaluate English text simplification, as well as to create simplified text in future work.

Categories and Subject Descriptors

K.4.2 [**Computers and Society**]: Social Issues – Assistive technologies for persons with disabilities

Keywords

Text simplification; accessibility; crowdsourcing; NLP

1. INTRODUCTION

Simplified text is crucial for some populations to read effectively, especially for cognitively impaired and low-literacy people. In fact, the United Nations [39] proposed a set of standard rules for the equalization of opportunities for persons with disabilities, including document accessibility. Given this, there are different initiatives that propose guidelines to help rewriting text to make it more comprehensible. Examples of these guidelines are *Plain Language* in the U.S. [42][1] and the *European Guidelines for the Production of Easy-to-Read Information* in Europe [24].

Text simplification is an area of Natural Language Processing (NLP) that attempts to solve this problem automatically by reducing the complexity of both wording (lexicon) and structure (syntax) in a sentence [49]. Previous efforts on automatic text simplification have focused on people with autism [21, 40], people with Down syndrome [48], people with dyslexia [43, 44], and people with aphasia [12, 19].

However, automatic text simplification is in an early stage and still is not useful for the target users [21, 44, 48]. One of the main challenges of automatic text simplification is its evaluation, which frequently relies on readability measures. Traditionally these measures are computed automatically and take into account surface features of the text, such as number of words and number of sentences [29]. More recent measures consider other features used in machine learning techniques [22]. However, these features generally do not consider gains for the end user. Most of the human evaluations of text simplifications are done by experts (normally using between two or three annotators) and consider their annotations and the agreement between them.

In this paper, we show that crowdsourcing can be used to collect input from highly-available, non-expert workers in order to provide an effective and consistent means of evaluating the simplicity of text in English. This provides a way to go beyond automatic measures, which may overlook important content changes that actually *add* difficulty to reading text.

To the best of our knowledge, our work is the first to focus on measuring text simplification for English using the crowd. By using human intelligence in a structured way, we can provide a reliable way of measuring text simplicity for end users. More generally, by addressing the simplicity feedback problem, we may be able to better facilitate the creation systems that selectively simplify textual content in previously unaddressed settings on demand, such as new content on a web page. This can provide more accessible accommodations to those who need it, when they need it.

2. RELATED WORK

The work related to our study can be grouped into (a) NLP studies on measuring text simplification, (b) crowdsourcing for accessibility, and (c) crowdsourcing for NLP.

2.1 Measuring Text Simplification in NLP

In NLP text simplification is defined as any process that reduces the syntactic or lexical complexity of a text while attempting to preserve its meaning and information content. The aim of text simplification is to make text easier to comprehend for a human user,

[1] `http://www.plainlanguage.gov/`

W4A'15 May 18 - 20, 2015, Florence, Italy.
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-3342-9/15/05 $15.00.
http://dx.doi.org/10.1145/2745555.2746658

or process by a program [49]. The quality of the simplified texts is generally evaluated by using a combination of automatic readability metrics (measuring the degree of simplification) and human assessment (measuring the grammaticality, preservation of meaning, and degree of simplification).

Traditional methods, such as the Flesch Reading Ease score [23], have used features like average sentence or word length. More recent readability measures have included other type of features. These can be classified in four broad categories [15]: (a) lexico-semantic features, *e.g.* lexical richness [36]; (b) psycholinguistics-based lexical features, *e.g.* word concreteness [52]; (c) syntactic features, *e.g.* syntactic complexity [26]; and (d) discourse-based features or higher-level semantic and pragmatic features. Machine learning approaches for readability prediction use combinations of these features to classify different levels of readability such as the feature analyses performed by Feng *et al.* [22].

In order to compare automatic and human assessment, Štajner et al. [56] used 280 pairs of original sentences and their corresponding simplified versions and compared the scores of six machine translation evaluation metrics with human judgements (three evaluators). The measures used were the grammaticality and meaning preservation of the simplified sentences. They found a strong correlation between automatic and human measures.

When using human assessment, the systems can be evaluated either by experts or by target groups such as people with autism [21, 40], people with Down syndrome [48], or people with dyslexia [43, 44]. Evaluations by expert annotators are more commonly used. Typically, three annotators are asked to grade different outputs in terms of its grammaticality, its preservation of the original meaning, and its simplicity. The agreement between the annotations is calculated by mean inter-annotation agreement metrics (using weighted kappa or Fleiss' kappa) and ranged from 0.49 to 0.69 [57] and 0.35 to 0.53 [9] for English lexical simplification, 0.33 [10] and 0.41 to 0.54 [3] for Spanish lexical simplification, and from 0.53 to 0.68 for English syntactic simplification [56]. These scores suggest that the agreement in these tasks is still not reliable since it is common practice among researchers in computational linguistics to consider 0.8 as a minimum value of acceptance [1].

2.2 Crowdsourcing Accessibility

Remote help from people has long been used to assist people with disabilities [8]. The ESP Game [54] elicited image labels from users, in part, with the goal of creating annotations for blind users for images on the web.

Recently, crowdsourcing has provided a means of supporting on-demand assistance. VizWiz [7] used the crowd to provide nearly real-time answers to visual questions asked by blind users from a phone or other mobile device. Chorus:View [32] extended the basic VizWiz model by allowing people to interact conversationally with a question answering system by viewing a user's live video stream in order to answer questions during a chat conversation.

More recently, real-time crowdsourcing has allowed for the creation of on-demand accessibility tools, as well as new tools for creating interactive experiences powered by a combination of human and machine intelligence have been created. Seaweed [14] recruited workers and asked them to wait for a task to begin in order to complete tasks synchronously. Adrenaline [4] showed that workers could be directed to a task in under two seconds from such a model, and with proper optimization, even under a second [5].

Legion [31] introduced the idea of *continuous* crowdsourcing, which engaged workers for longer tasks to allow them to maintain context and respond even faster. Scribe [30] built on this work to create a system that provides real-time captions to deaf and hard-of-hearing users with a per-word latency of under 5 seconds.

Our work explores using crowdsourcing as a means of quickly and easily evaluating textual simplicity. By building off of prior work, our goal is to enable systems that support assistive technologies by evaluating their performance in real time, allowing them to better adapt to the environment in which they are operating.

2.3 Crowdsourcing for NLP

Crowdsourcing draws on large highly-available, dynamic groups of human workers to solve tasks that automated systems cannot yet accomplish alone [35, 53]. Integrating human intelligence into computational workflows has allowed problems ranging from image labeling [54], to protein folding [16], taxonomy structuring [13], and trip planning [58] to be solved. In NLP crowdsourcing has been used to evaluate and edit writing [6], identify textual features [51], and even hold conversations [34]. Machine translation, which aims to automatically convert one language to another, has also been supported by the crowd [11]. Machine translation is in many ways similar to the problem of converting complex text into simple text.

Regarding text simplification for Dutch, De Clercq *et al.* [18] used of crowdsourcing to obtain readability assessments. The authors used a corpus of written Dutch generic texts and a tool to obtained readability ranking from the expert readers. From the crowd workers the authors obtained pairwise comparisons. Their results show that the assessments collected through both methodologies are highly consistent.

Finally, Pellow and Eskenazi [41] used crowdsourcing to generate simplifications of everyday documents, *e.g.* driver's licensing or government documents. They created a corpus or 120,000 sentences and demonstrated the feasibility of using crowdsourced simplifications to simplify documents.

2.4 What is missing?

In this paper, we explore how recruiting large groups of people to analyze text simplification approaches can effectively measure more fine-grained classes of simplification that are usually studied, without the need for expensive experts who are not always available when needed. This is especially important for automated systems that may need immediate feedback on the simplicity of their output. Our approach complements this work addressing whether crowdsourcing could be used to measure text simplification in English for legal texts.

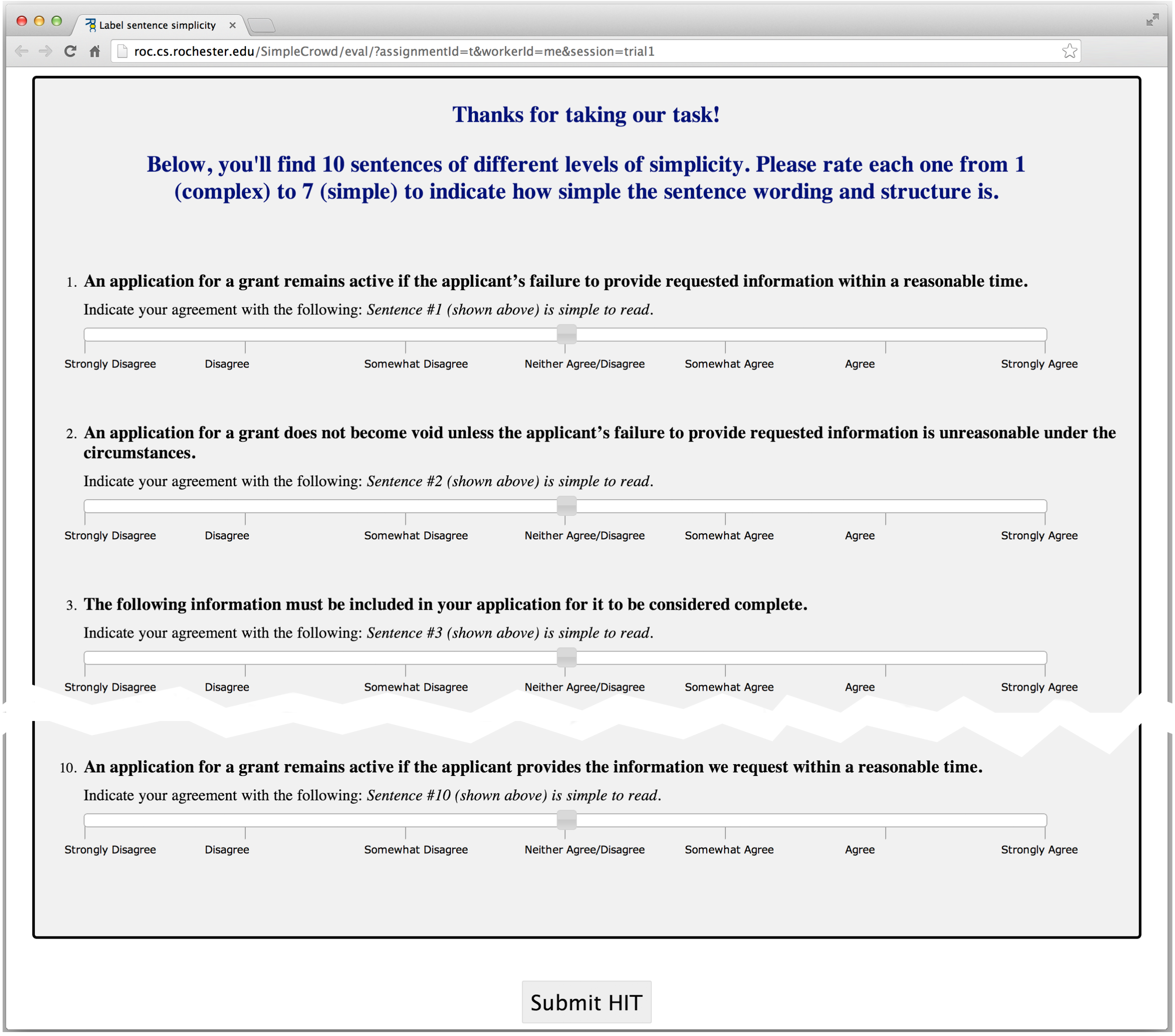
3. MEASURING SIMPLICITY

To collect simplicity ratings from the crowd, we created a simple text-rating interface (Figure 1) that asked workers recruited from Amazon's Mechanical Turk to rate a series of 10 sentences on a 7-point Likert scale.

The system we developed for this evaluation automatically pulls sentences that have been added to a queue and serves them to workers as they arrive. This means that response speed is dependent only on the arrival speed of workers, which can be increased by using a retainer model [4] and/or increasing pay to encourage workers to take tasks quicker if a rapid response is desired. In our trials, we focus on the ability of workers to accurately evaluate text simplicity, rather than optimizing for response speed.

3.1 Evaluation Dataset

Our evaluation explored how well the crowd was able to judge the simplicity of sentences. For our evaluation dataset, we chose a set of sentences from the PLAIN language guidelines [42]. We decided to chose these sentences because they have a considerable


Label sentence simplicity
roc.cs.rochester.edu/SimpleCrowd/eval/?assignmentId=t&workerId=me&session=trial1

Thanks for taking our task!

Below, you'll find 10 sentences of different levels of simplicity. Please rate each one from 1 (complex) to 7 (simple) to indicate how simple the sentence wording and structure is.

1. **An application for a grant remains active if the applicant's failure to provide requested information within a reasonable time.**

   Indicate your agreement with the following: *Sentence #1 (shown above) is simple to read*.

   Strongly Disagree | Disagree | Somewhat Disagree | Neither Agree/Disagree | Somewhat Agree | Agree | Strongly Agree

2. **An application for a grant does not become void unless the applicant's failure to provide requested information is unreasonable under the circumstances.**

   Indicate your agreement with the following: *Sentence #2 (shown above) is simple to read*.

   Strongly Disagree | Disagree | Somewhat Disagree | Neither Agree/Disagree | Somewhat Agree | Agree | Strongly Agree

3. **The following information must be included in your application for it to be considered complete.**

   Indicate your agreement with the following: *Sentence #3 (shown above) is simple to read*.

   Strongly Disagree | Disagree | Somewhat Disagree | Neither Agree/Disagree | Somewhat Agree | Agree | Strongly Agree

10. **An application for a grant remains active if the applicant provides the information we request within a reasonable time.**

    Indicate your agreement with the following: *Sentence #10 (shown above) is simple to read*.

    Strongly Disagree | Disagree | Somewhat Disagree | Neither Agree/Disagree | Somewhat Agree | Agree | Strongly Agree

Figure 1: Our interface for collecting crowd ratings of sentence simplicity asked each participant to rate the simplicity of 10 sentences on a 7-point Likert scale. We did not specify the concept of "simple" in terms of rules because we thought this could bias workers' perception of simplicity and may also be too complex.

level of complexity (legal genre), but at the same time, are used quite frequently in official documents.

Each of the selected sentences could be simplified using several of the rules suggested in the guidelines. Because we do not only want to determine if people can tell if a sentence has "been simplified", but also *how simple* the result is, we apply rules in the following ways over five conditions. This allowed us to measure the crowd's ability not only to differentiate simple from complex, but also to rate a continual notion of simplicity between extremes.

- **0 or original:** No rules applied. This is the original sentence from the guidelines.
- **1:** A single rule from the guideline-suggested set, applied once in the sentence.
- **1:** Another single rule from the guideline-suggested set, applied once in the sentence. This might be a different application of the same rule (in a different place from above), or the application of a different rule.
- **2:** Both of the single rules from above applied to the same sentence. This may be two applications of the same rule in a sentence in difference locations, or the application of two different rules.
- **3:** Three rules applied to the same sentence. We begin with the two rules applied in the previous condition, but now add a third rule (or repeated instance of a prior rule) to the set.

Our goal is to create a dataset with incremental simplifications that can serve as a gold standard. To illustrate our dataset, we include two example sentences and simplification change sets below. The simplification changes are shown in parentheses. The dataset used in the experiment was composed of 60 sentences. See the complete dataset in the Appendix.

Example 1:

0 or original: **The following information must be included in the application for it to be considered complete.**

1: *You must include the following information in the application for it to be considered complete.* (must be included → you must; Guideline "Use active voice" [42].)

1: *The following information must be included in your application for it to be considered complete.* (the → your) Guideline "Use pronouns to speak directly to readers" [42].

2: *You must include the following information in your application for it to be considered complete.* The combination of the two previous simplification changes.

3: *You must include the following information in your application.* (for it to be considered complete → omission) Guideline "Omit unnecessary words" [42].

Example 2:

Original: **Bonds will be withheld in cases of noncompliance with all permits and conditions.**

1: *Bonds will be withheld if you don't comply with all permits and conditions.* (in cases of non-compliance → if you don't comply) Guidelines "Don't turn verbs into nouns" and "Use pronouns to speak directly to readers" [42].

1: *Your bond will be withheld in cases of non-compliance with all permits and conditions.* (Bonds → Your bond) Guideline "Use pronouns to speak directly to readers" [42].

2: *Your bond will be withheld if you don't comply with all permits and conditions.* The combination of the two previous simplification changes.

3: *We will withhold your bond if you don't comply with all permit terms and conditions.* (Bonds will be withheld → We will withhold your) Guideline "Use active voice" [42].

3.2 Rating Results

We elicited 2,500 individual ratings from 250 workers. Each worker was shown a set of 10 randomly-ordered sentence variants from two of the sentences from our dataset, that is two complex sentences and their corresponding simplifications.

Figure 2 shows worker ratings of each of the 10 sentences that we had rated, over the different number of simplification rules applied. While at first the results appear mixed, our inspection of the dataset shows that the diversity of responses mirrors the diversity of the sentences used and changes made to them. To control for differing complexity and get better insight into the types of text changes workers captured in their ratings, we divided the initial dataset into two groups: those sentences which were significantly changed (defined as more than 10 changes in terms of word-level edit distance), and those that were not.

Figure 3 shows the aggregated results for the two sets. Sentences which changed very little after the standard simplification rules were applied saw no significant change overall, growing just 6.6% ($p > .05$). Sentences with more changes, on the other hand, saw a significant increase of 76.5% ($p < .05$). These results confirm that workers are able to accurately rate the simplifying changes that we added to our set of sentences. Further checks of the data confirmed that borderline cases, where roughly 10 changes are made, result in a minor but positive increase in simplicity rating.

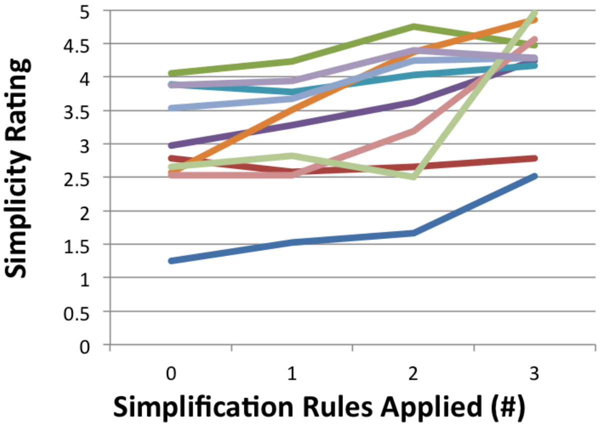


Figure 2: Average worker ratings of all 10 trial sentences as different canonical rules are applied. Different rules have different effects although generally the trend is to increase simplicity as expected.

Unsurprisingly, the set of sentences with smaller sets of changes were also those that were initially rated higher (Figure 3) by the higher initial value when 0 changes are present. This further verifies that workers' ratings are in line with our expected measurement outcomes.

3.3 Rating Consistency

The results presented above demonstrate that a crowd of 50 workers can accurately rate the simplicity of text. However, requiring 50 workers to redundantly code each sentence that needs to be rated is likely going to be impractical in many cases, even if each individual rating is cheap and easy to acquire. We next investigated the number of workers that might be needed in a real system by looking at the variation in answers when sampling a part of our full dataset.

Figure 4 shows the convergence toward the whole group's answer (separated into large-difference, small-difference, and combined conditions) as we sample a randomly selected 25%, 50%, and 75% of the contributing workers. The initial separation between the two conditions that we observed at a 25% sampling rate is not significant ($p > .5$). After an initially large drop of 0.31 (from .44 to .195, a difference of 55.7% of the total average difference), we see a more modest drop of 0.1 (22.7% of the total average difference). At 75%, we are within 0.1 of the final collective answer.

To verify these results are not dependant on absolute values, we next check for convergence in terms of the percent difference between answers over the same four worker response sampling rates (Figure 5). The first thing to note is that the small difference in starting value (shown above to be insignificant) almost completely disappears when viewed as a percent of the overall score. This suggests that the relative error even at just 25% of workers is reasonably stable.

We again observed a similar pattern to above – the initial step from sampling 25% to 50% of workers results in a relatively large step down from 11.5% difference to just 4.5% (a difference of 61.1% of the overall error), and a smaller decrease of 2.3% (20.3%

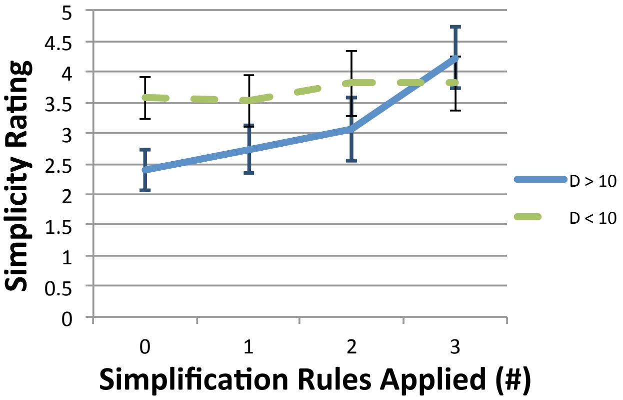


Figure 3: Average worker ratings (with standard deviation shown) of our 10 trial sentences, divided into two groups: high-change sentences (6) and low-change sentences (4) as different canonical rules are applied. A clear effect is observed when applying each rule in settings where there is a larger difference in the resulting sentence, suggesting workers are able to accurately pick up on successful rule-base simplicity changes.

of the overall difference) is seen between 50% and 75%. The final step from 75% to 100% of our 50 workers was again small, at just 2.1% (18.5% of the overall difference).

To select a cutoff value, we noted the reduction in convergence rate, and concluded that approximately 50%, or 25 workers, is suggested as a reasonable threshold by the data. This analysis confirmed that workers collectively provide consistent results, because it shows that workers monotonically converge to a more stable answer as more responses are added.

4. DISCUSSION

In this paper, we have explored how the non-expert crowd can measure the simplicity of text. Our results are consistent with the findings of De Clercq *et al.* [18] for generic texts in Dutch. In their study the crowd was presented with pairwise comparisons. Their findings suggest that crowdsourcing could be used as an alternative to expert text labeling for text simplification. Our results advance these findings in two aspects. First previous results can be extended to English language. This contribution is not trivial because text complexity perception could be language dependent, as different languages present different degrees of complexity in different language levels (*e.g.* morphology *vs.* syntax). Second, we show that the crowd is sensitive to different levels of simplicity. This allows fine-grained impact of changes to be measured.

Next, we discuss the results in relationship with target populations and readability measures. Later we discuss how crowdsourcing applies to text simplification in terms of fine-grained answers, workforce, response speed, and potential applications. Finally we discuss the limitations of the study.

4.1 Readability in Target Populations

The ratings that were given by the crowd are consistent with previous studies on experimental psychology that have studied how word and sentence length impact the performance of readers with special needs. As with the crowd, short sentences and short words are more readable not only by general readers but also by people with cognitive disabilities. For instance, Simmons and Singleton show that long and complex sentences have a negative effect on the comprehension of people with dyslexia [50]. Similarly, Rello *et al.* [45] demonstrate how text containing less frequent and long words were less understandable by people with dyslexia than the same texts when the long words were substituted by shorter synonyms. At the same time, the presence of numerical expressions make texts more challenging to read for people with and without dyslexia [46]. Minshew and Goldstein [38] present that people with autism spectrum disorders have difficulty understanding complex sentences. Finally, Baddeley *et al.* [2] show how children with Down syndrome can remember a greater number of shorter words than long words. The more syllables the word have the less words children with Down syndrome could retain.

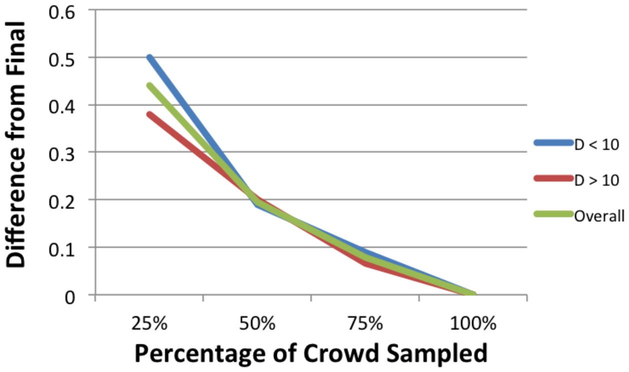


Figure 4: Crowd sampling rate *vs.* the difference from the final answer. Overall, the difference from 25% of the 50 crowd workers polled to all of them was only 0.5 points.

4.2 Readability Measures

The crowd measured as simpler the sentences with more simplification changes, which, in most of the cases, are also the shorter ones. Their ratings support the readability measures used in NLP to address text complexity.

There are over 200 readability measures [20]. If we take into account four of the most frequently used readability metrics, they all account factors related to word and sentence length [55]. For instance, the Flesch Reading Ease score [23] and its simplified version, the Flesch-Kincaid readability formula [29], consider the average sentence length and the average number of syllables per word. The Fog Index [27] accounts average sentence length and the number of words containing more than two syllables. Finally, the SMOG grading [37] take into account the number of words with more than two syllables every 30 sentences.

4.3 Fine-Grained Answers

One of the advantages of crowd-based simplification measurements is the granularity of responses that are possible elicit. Unlike prior approaches, our crowd responses are consistent over varying levels of simplicity, allowing fine-grained impact of changes to be measured. This may allow for better find-tuning of simplication algorithms or other approaches.

4.4 Workforce

Our results suggest that a relatively large number of workers may be necessary to get reliable results, as smaller samples from our data (*e.g.*, 10%) showed high variance. Our task paid 20 cents to rate 10 sentences. Workers typically could complete this task in less than 90 seconds, making their effective pay rate over $8 per hours – a relatively high wage for Mechanical Turk. As such, these tasks

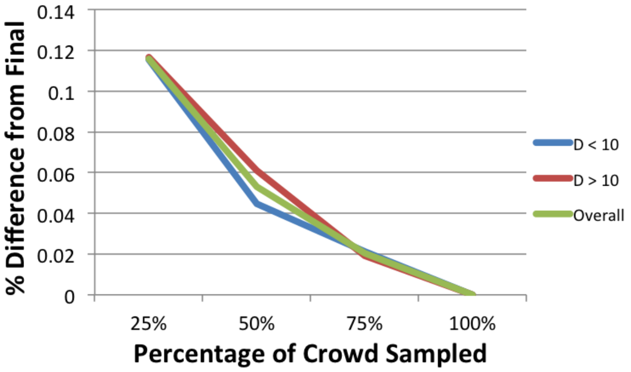


Figure 5: Crowd sampling rate *vs.* difference from final answer. Even a small fraction of the 50 crowd workers polled for this study were able to approximate the final result from them all.

were taken quickly. Assuming that we need around 25 workers as suggested by our results, this brings the total cost per sentence to around 50 cents.

4.5 Response Speed

Some applications may want to rate the simplicity of content quickly and on-the-fly. The method introduced here is able to do that well. By dividing up a block of text into multiple pieces and asking different workers to assess each in parallel, the overall response time can be only a small fraction of the time it would take any one person. While the $0.50 USD per sentence that we paid is somewhat high (although we did not optimize it), it may still be cheaper —and certainly parallelized to be faster than —an expert.

4.6 Potential Applications

We believe that our crowd-based measurement of simplicity has many applications. For one, it may allow researchers in natural language processing to proceed faster than they would otherwise be able to by providing a way for them to gauge the simplicity of the sentences that they produce. Second, it may allow for new kinds of tools that determine on-the-fly the simplicity of text that is being viewed and then perform some process to mitigate any problems that it may cause, without requiring users to explicitly request this additional functionality.

4.7 Study Limitations

One limitation of our study is that we did not measure whether the crowd actually comprehended the text. This is a common flaw of the human evaluations that address text simplification with annotators [9, 10, 57], maybe be due to the fact that comprehension and readability normally are so closely related that sometimes both terms have been used interchangeably [28]. Another limitation of the study is that our testing dataset in restricted to the legal domain. Since most of the NLP tasks are domain dependent, including text simplification [49], our results do not necessarily extend to other language domains. However, because the underlying power for our approach comes from human understanding, even other domains that cannot be accurately measured by Mechanical Turk workers might be successfully measured by others crowds with different domain expertise or knowledge.

5. CONCLUSIONS AND FUTURE WORK

The main conclusion of this study is that the non-expert crowd recruited from Amazon Mechanical Turk can perceive and measure different levels of simplicity in text – demonstrating that crowdsourcing may be a viable tool for evaluating the accessibility of text. The measurements of the crowd can support the needs of readers with cognitive disabilities, as well measure general readability issues in textual accessibility. This may be due to the large overlapping of the language symptoms between different developmental language disorders such as autism spectrum disorders or Down syndrome [47]. Since text simplification is beneficial also for non–native language learners [17] and low literacy people [25], the results of this study could be extended to other fields besides accessibility.

Future work may include defining new readability measures based on the crowd's judgements, as well as the integration of crowdsourcing evaluations in NLP simplification methods. By integrating human evaluations, automatic simplifications could come closer to the real needs or users. It would also allow developers of these technologies to quickly iterate and test out designs.

Eventually, it may even be possible to use the crowd to *generate* simplified text. The crowd may also bring diverse insights into what makes text simpler, and even help formalize the process of simplifcation for automated systems, allowing them to more quickly learn how to complete this task – similar to how crowd understanding of tasks has been used in other domains [33].

While the goal of this paper was to measure how simple existing text is, an obvious next step is to ask the crowd to actually generate the simplified text itself. Crowdsourcing is likely to be able to simplify text more accurately than automated approaches due to people's understanding of the content, and may also be able to target the text that they produce toward the current context or user abilities.

Acknowledgments

This work has been supported by a Microsoft Research Ph.D. Fellowship, an Alfred P. Sloan Foundation Fellowship, National Science Foundation Awards #IIS-1149709 and #IIS-1116051, and Google.

6. REFERENCES

[1] R. Artstein and M. Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[2] A. D. Baddeley, N. Thomson, and M. Buchanan. Word length and the structure of short-term memory. *Journal of verbal learning and verbal behavior*, 14(6):575–589, 1975.

[3] R. Baeza-Yates, L. Rello, and J. Dembowski. A context-aware synonym simplification algorithm: Cassa. In *Proc. NAACL '15*, Denver, Colorado, USA, 2015. ACM.

[4] M. S. Bernstein, J. Brandt, R. C. Miller, and D. R. Karger. Crowds in two seconds: Enabling realtime crowd-powered interfaces. In *Proc. UIST '11*, pages 33–42, New York, NY, USA, 2011. ACM.

[5] M. S. Bernstein, D. R. Karger, R. C. Miller, and J. Brandt. Analytic methods for optimizing realtime crowdsourcing. *CoRR*, abs/1204.2995, 2012.

[6] M. S. Bernstein, G. Little, R. C. Miller, B. Hartmann, M. S. Ackerman, D. R. Karger, D. Crowell, and K. Panovich. Soylent: A word processor with a crowd inside. In *Proc. UIST '10*, pages 313–322, New York, NY, USA, 2010. ACM.

[7] J. P. Bigham, C. Jayant, H. Ji, G. Little, A. Miller, R. C. Miller, R. Miller, A. Tatarowicz, B. White, S. White, and T. Yeh. Vizwiz: Nearly real-time answers to visual questions. In *Proceedings of the 23Nd Annual ACM Symposium on*

User Interface Software and Technology, UIST '10, pages 333–342, New York, NY, USA, 2010. ACM.

[8] J. P. Bigham and R. E. Ladner. What the disability community can teach us about interactive crowdsourcing. *interactions*, 18(4):78–81, July 2011.

[9] O. Biran, S. Brody, and N. Elhadad. Putting it simply: a context-aware approach to lexical simplification. In *Proc. ACL'11*, pages 496–501, Portland, Oregon, USA, 2011.

[10] S. Bott, L. Rello, B. Drndarevic, and H. Saggion. Can Spanish be simpler? LexSiS: Lexical simplification for Spanish. In *Proc. Coling '12*, Mumbay, India, 2012.

[11] C. Callison-Burch. Fast, cheap, and creative: Evaluating translation quality using amazon's mechanical turk. In *Proc. EMNLP '09*, pages 286–295, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[12] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. Practical Simplification of English Newspaper Text to Assist Aphasic Readers. In *Proc. of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10, 1998.

[13] L. B. Chilton, G. Little, D. Edge, D. S. Weld, and J. A. Landay. Cascade: Crowdsourcing taxonomy creation. In *Proc. CHI '13*, pages 1999–2008, New York, NY, USA, 2013. ACM.

[14] L. B. Chilton, C. T. Sims, M. Goldman, G. Little, and R. C. Miller. Seaweed: A web application for designing economic games. In *Proceedings of the ACM SIGKDD Workshop on Human Computation*, HCOMP '09, pages 34–35, New York, NY, USA, 2009. ACM.

[15] K. Collins-Thompson. Computational assessment of text readability: A survey of current and future research (working draft), 2014.

[16] S. Cooper, F. Khatib, A. Treuille, J. Barbero, J. Lee, M. Beenen, A. Leaver-Fay, D. Baker, Z. Popović, et al. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

[17] S. Crossley, M. Louwerse, P. McCarthy, and D. McNamara. A linguistic analysis of simplified and authentic texts. *The Modern Language Journal*, 91(1):15–30, 2007.

[18] O. De Clercq, V. Hoste, B. Desmet, P. Van Oosten, M. De Cock, and L. Macken. Using the crowd for readability prediction. *Natural Language Engineering*, pages 1–33, 2013.

[19] S. Devlin and G. Unthank. Helping aphasic people process online information. In *Proc. ASSETS '06*, pages 225–226. ACM, 2006.

[20] W. H. Dubay. The principles of readability a brief introduction to readability research, 2004.

[21] R. Evans, C. Orasan, and I. Dornescu. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) at EACL*, pages 131–140, 2014.

[22] L. Feng, M. Jansche, M. Huenerfauth, and N. Elhadad. A comparison of features for automatic readability assessment. In *Proc. ACL '10*, pages 276–284. Association for Computational Linguistics, 2010.

[23] R. Flesch. A new readability yardstick. *Journal of applied psychology*, 32(3):221, 1948.

[24] G. Freyhoff, G. Hess, L. Kerr, E. Menzel, B. Tronbacke, and K. V. D. Veken. European guidelines for the production of easy-to-read information for people with learning disability, 1998.

[25] C. Gasperin, E. Maziero, L. Specia, T. Pardo, and S. Aluisio. Natural language processing for social inclusion: a text simplification architecture for different literacy levels. *the Proceedings of SEMISH–XXXVI Seminário Integrado de Software e Hardware*, pages 387–401, 2009.

[26] E. Gibson. Linguistic complexity: Locality of syntactic dependencies. *Cognition*, 68(1):1–76, 1998.

[27] R. Gunning. *Technique of clear writing*. McGraw-Hill, New York, 1952.

[28] K. Inui, A. Fujita, T. Takahashi, R. Iida, and T. Iwakura. Text simplification for reading assistance: A project note. In *Proceedings of the second international workshop on Paraphrasing-Volume 16*, pages 9–16. Association for Computational Linguistics, 2003.

[29] J. P. Kincaid, R. P. Fishburne Jr, R. L. Rogers, and B. S. Chissom. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for navy enlisted personnel. Technical report, DTIC Document, 1975.

[30] W. Lasecki, C. Miller, A. Sadilek, A. Abumoussa, D. Borrello, R. Kushalnagar, and J. Bigham. Real-time captioning by groups of non-experts. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 23–34, New York, NY, USA, 2012. ACM.

[31] W. S. Lasecki, K. I. Murray, S. White, R. C. Miller, and J. P. Bigham. Real-time crowd control of existing interfaces. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology*, UIST '11, pages 23–32, New York, NY, USA, 2011. ACM.

[32] W. S. Lasecki, P. Thiha, Y. Zhong, E. Brady, and J. P. Bigham. Answering visual questions with conversational crowd assistants. In *Proceedings of the 15th International ACM SIGACCESS Conference on Computers and Accessibility*, ASSETS '13, pages 18:1–18:8, New York, NY, USA, 2013. ACM.

[33] W. S. Lasecki, L. Weingard, G. Ferguson, and J. P. Bigham. Finding dependencies between actions using the crowd. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 3095–3098, New York, NY, USA, 2014. ACM.

[34] W. S. Lasecki, R. Wesley, J. Nichols, A. Kulkarni, J. F. Allen, and J. P. Bigham. Chorus: A crowd-powered conversational assistant. In *Proceedings of the 26th Annual ACM Symposium on User Interface Software and Technology*, UIST '13, pages 151–162, New York, NY, USA, 2013. ACM.

[35] G. Little, L. B. Chilton, M. Goldman, and R. C. Miller. Turkit: Human computation algorithms on mechanical turk. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 57–66, New York, NY, USA, 2010. ACM.

[36] D. Malvern and B. Richards. Measures of lexical richness. *The Encyclopedia of Applied Linguistics*, 2012.

[37] G. H. McLaughlin. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[38] N. J. Minshew and G. Goldstein. Autism as a disorder of complex information processing. *Mental Retardation and Developmental Disabilities Research Reviews*, 4(2):129–136, 1998.

[39] U. Nations. Standard Rules on the Equalization of Opportunities for Persons with Disabilities, 1994.

[40] C. Orasan, R. Evans, and I. Dornescu. *Towards Multilingual Europe 2020: A Romanian Perspective*, chapter Text Simplification for People with Autistic Spectrum Disorders, pages 287–312. Romanian Academy Publishing House, Bucharest, 2013.

[41] D. Pellow and M. Eskenazi. An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, pages 84–93, 2014.

[42] Plain Language Action and Information Network (PLAIN). *Federal Plain Language Guidelines*. US Government, 2011. `http://www.plainlanguage.gov/`.

[43] L. Rello and R. Baeza-Yates. Evaluation of Dyswebxia: A reading app designed for people with dyslexia. In *Proc. W4A '14*, Seoul, Korea, 2014.

[44] L. Rello, R. Baeza-Yates, S. Bott, and H. Saggion. Simplify or help? Text simplification strategies for people with dyslexia. In *Proc. W4A '13*, Rio de Janeiro, Brazil, 2013.

[45] L. Rello, R. Baeza-Yates, L. Dempere, and H. Saggion. Frequent words improve readability and short words improve understandability for people with dyslexia. In *Proc. INTERACT '13*, Cape Town, South Africa, 2013.

[46] L. Rello, S. Bautista, R. Baeza-Yates, P. Gervás, R. Hervás, and H. Saggion. One half or 50%? An eye-tracking study of number representation readability. In *Proc. INTERACT '13*, Cape Town, South Africa, 2013.

[47] M. L. Rice, S. F. Warren, and S. K. Betz. Language symptoms of developmental language disorders: An overview of autism, down syndrome, fragile x, specific language impairment, and williams syndrome. *Applied psycholinguistics*, 26(01):7–27, 2005.

[48] H. Saggion, S. Štajner, S. Bott, S. Mille, L. Rello, and B. Drndarevic. Making it Simplext: Implementation and evaluation of a text simplification system for spanish. *ACM Transactions on Accessible Computing (TACCESS)*, In Press, 2015.

[49] A. Siddharthan. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4(1):77–109, 2006.

[50] F. Simmons and C. Singleton. The reading comprehension abilities of dyslexic students in higher education. *Dyslexia*, 6(3):178–192, 2000.

[51] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proc. EMNLP '08*, pages 254–263, Stroudsburg, PA, USA, 2008. Association for Computational Linguistics.

[52] S. Tanaka, A. Jatowt, M. P. Kato, and K. Tanaka. Estimating content concreteness for finding comprehensible documents. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 475–484. ACM, 2013.

[53] L. von Ahn. *Human Computation*. PhD thesis, Carnegie Mellon University, Pittsburgh, PA, 2005.

[54] L. von Ahn and L. Dabbish. Labeling images with a computer game. In *Proc. CHI '04*, pages 319–326, New York, NY, USA, 2004. ACM.

[55] S. Štajner, R. Evans, C. Orasan, and R. Mitkov. What can readability measures really tell us about text complexity. In *Proceedings of the the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*, 2012.

[56] S. Štajner, R. Mitkov, and H. Saggion. One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)@ EACL*, pages 1–10, 2014.

[57] M. Yatskar, B. Pang, C. Danescu-Niculescu-Mizil, and L. Lee. For the sake of simplicity: Unsupervised extraction of lexical simplifications from Wikipedia. In *Proc. ACL'10*, pages 365–368, Uppsala, Sweden, 2010.

[58] H. Zhang, E. Law, R. Miller, K. Gajos, D. Parkes, and E. Horvitz. Human computation tasks with global constraints. In *Proc. CHI '12*, pages 217–226, New York, NY, USA, 2012. ACM.

APPENDIX

Dataset used in the experiment. Numbers "1", "2" and "3" refer to the number of manual simplifications performed to the original sentence.

Sentence 1, original: **An application for a grant does not become void unless the applicant's failure to provide requested information is unreasonable under the circumstances.**

1: *An application for a grant does not become void if the applicant's failure to provide requested information is unreasonable under the circumstances.*

1: *An application for a grant remains active unless the applicant's failure to provide requested information is unreasonable under the circumstances.*

2: *An application for a grant remains active if the applicant's failure to provide requested information within a reasonable time.*

3: *An application for a grant remains active if the applicant provides the information we request within a reasonable time.*

Sentence 2, original: **The following information must be included in the application for it to be considered complete.**

1: *You must include the following information in the application for it to be considered complete.*

1: *The following information must be included in your application for it to be considered complete.*

2: *You must include the following information in your application for it to be considered complete.*

3: *You must include the following information in your application.*

Sentence 3, original: **Bonds will be withheld in cases of non-compliance with all permits and conditions.**

1: *Bonds will be withheld if you don't comply with all permits and conditions.*

1: *Your bond will be withheld in cases of non-compliance with all permits and conditions.*

2: *Your bond will be withheld if you don't comply with all permits and conditions.*

3: *We will withhold your bond if you don't comply with all permit terms and conditions.*

Sentence 4, original: **These sections describe types of information that would satisfy the application requirements of Circular A-110 as it would apply to this grant program.**

1: *These sections tell you the information that would satisfy the application requirements of Circular A-110 as it would apply to this grant program.*

1: *These sections describe types of information to meet requirements of Circular A-110 as it would apply to this grant program.*

2: *These sections tell you the information to meet the requirements of Circular A-110 as it would apply to this grant program.*

3: *These sections tell you how to meet the requirements of Circular A-110 for this grant program.*

Sentence 5, original: **The production of accurate statistics is important for the committee in the assessment of our homelessness policy.**

1: *Producing accurate statistics is important for the committee in the assessment of our homelessness policy.*

1: *The production of accurate statistics is important for the committee in the assessment of our policy on homelessness.*

2: *Producing accurate statistics is important for the committee in the assessment of our policy on homelessness.*

3: *Producing accurate statistics is important to the committee in assessing our policy on homelessness.*

Sentence 6, original: **The applicant shall be notified by registered mail in all cases where the permit applied for is not granted, and shall be given 30 days within which to appeal such decision.**

1: *The applicant shall be notified by registered mail if we reject your application, and shall be given 30 days within which to appeal such decision.*

1: *The applicant shall be notified by registered mail in all cases where the permit applied for is not granted. You must file an appeal of that decision within 30 days.*

2: *The applicant shall be notified by registered mail if we reject your application. You must file an appeal of that decision within 30 days.*

3: *We will notify you by registered mail if we reject your application. You must file an appeal of that decision within 30 days.*

Sentence 7, original: **Total disclosure of all facts is very important to make sure we draw up a total and completely accurate picture of the Agency's financial position.**

1: *Total disclosure of all facts is important to make sure we draw up a total and accurate picture of the Agency's financial position.*

1: *Disclosing all facts is very important to make sure we draw up a total and completely accurate picture of the Agency's financial position.*

2: *Disclosing all facts is important to make sure we draw up a total and accurate picture of the Agency's financial position.*

3: *Disclosing all facts is important to creating an accurate picture of the Agency's financial position.*

Sentence 8, original: **If the State Secretary finds that an individual has received a payment to which the individual was not entitled, whether or not the payment was due to the individual's fault or misrepresentation, the individual shall be liable to repay to State the total sum of the payment to which the individual was not entitled.**

1: *If the State Secretary finds that an individual has received a payment to which the individual was not entitled, the individual shall be liable to repay to State the total sum of the payment to which the individual was not entitled.*

1: *If the State Secretary finds that an individual has received a payment to which the individual was not entitled, whether or not the payment was due to the individual's fault or misrepresentation, the individual shall be liable to repay to State the total sum back.*

2: *If the State Secretary finds that an individual has received a payment to which the individual was not entitled, the individual shall be liable to repay to State the total sum back.*

3: *If the State Secretary finds that you received a payment that you weren't entitled to, you must pay the entire sum back.*

Sentence 9, original: **Most refractory coatings to date exhibit a lack of reliability when subject to the impingement of entrained particulate matter in the propellant stream under extended firing durations.**

1: *The coating of most existing ceramics exhibit a lack of reliability when subject to the impingement of entrained particulate matter in the propellant stream under extended firing durations.*

1: *Most refractory coatings to date get damaged when subject to the impingement of entrained particulate matter in the propellant stream under extended firing durations.*

2: *The coating of most existing ceramics get damaged when subject to the impingement of entrained particulate matter in the propellant stream under extended firing durations.*

3: *The exhaust gas eventually damages the coating of most existing ceramics.*

Sentence 10, original: **Once the candidate's goals are established, one or more potential employers are identified. A preliminary proposal for presentation to the employer is developed. The proposal is presented to an employer who agrees to negotiate an individualized job that meets the employment needs of the applicant and real business needs of the employer.**

1: *Once the candidate's goals are established, one or more potential employers are identified. A preliminary proposal for presentation to the employer is developed. The proposal is presented to an employer who agrees to negotiate a job that meets both his and your employment needs.*

1: *Once the candidate's goals are established, one or more potential employers are identified. We prepare a preliminary proposal to present to an employer who agrees to negotiate an individualized job that meets the employment needs of the applicant and real business needs of the employer.*

2: *Once the candidate's goals are established, one or more potential employers are identified. We prepare a preliminary proposal to present to an employer who agrees to negotiate a job that meets both his and your employment needs.*

3: *Once we establish your goals, we identify one or more potential employers. We prepare a preliminary proposal to present to an employer who agrees to negotiate a job that meets both his and your employment needs.*

Sentence 11, original: **If any member of the board retires, the company, at the discretion of the board, and after notice from the chairman of the board to all the members of the board at least 30 days before executing this option, may buy, and the retiring member must sell, the member's interest in the company.**

1: *The company, at the discretion of the board, and after notice from the chairman of the board to all the members of the board at least 30 days before executing this option, may buy, and the retiring member must sell, the member's interest in the company.*

1: *If any member of the board retires, the company, at the discretion of the board, and after notice from the chairman of the board to all the members of the board at least 30 days before executing this option, may buy, and the retiring member must sell.*

2: *The company, at the discretion of the board, and after notice from the chairman of the board to all the members of the board at least 30 days before executing this option, may buy, and the retiring member must sell.*

3: *The company may buy a retiring member's interest.*

Sentence 12, original: **Applicants may be granted a permit to prospect for geothermal resources on any federal lands except lands in the National Park System, unless the applicant holds valid existing rights to the geothermal resources on the National Park System lands listed in the application.**

1: *You may be granted a permit to prospect for geothermal resources on any federal lands except lands in the National Park System, unless the applicant holds valid existing rights to the geothermal resources on the National Park System lands listed in the application.*

1: *Applicants may be granted a permit to prospect for geothermal resources on any federal lands. This includes lands in the National Park System only if you hold valid existing rights to the geothermal resources on the National Park System lands listed in the application.*

2: *You may be granted a permit to prospect for geothermal resources on any federal lands. This includes lands in the National Park System only if you hold valid existing rights to the geothermal resources on the National Park System lands listed in the application.*

3: *You may be granted a permit to prospect for geothermal resources on any federal lands. This includes lands in the National Park System only if you hold valid existing rights to the park lands listed in your application.*