An Open Corpus of Everyday Documents for Simplification Tasks

David Pellow and **Maxine Eskenazi**
Language Technologies Institute, Carnegie Mellon University
Pittsburgh PA USA
dpellow@cs.cmu.edu, max@cs.cmu.edu

Abstract

In recent years interest in creating statistical automated text simplification systems has increased. Many of these systems have used parallel corpora of articles taken from Wikipedia and Simple Wikipedia or from Simple Wikipedia revision histories and generate Simple Wikipedia articles. In this work we motivate the need to construct a large, accessible corpus of everyday documents along with their simplifications for the development and evaluation of simplification systems that make everyday documents more accessible. We present a detailed description of what this corpus will look like and the basic corpus of everyday documents we have already collected. This latter contains everyday documents from many domains including driver's licensing, government aid and banking. It contains a total of over 120,000 sentences. We describe our preliminary work evaluating the feasibility of using crowdsourcing to generate simplifications for these documents. This is the basis for our future extended corpus which will be available to the community of researchers interested in simplification of everyday documents.

1 Introduction

People constantly interact with texts in everyday life. While many people read for enjoyment, some texts must be read out of necessity. For example, to file taxes, open a bank account, apply for a driver's license or rent a house, one must read instructions and the contents of forms, applications, and other documents. For people with limited reading ability - whether because they are not native speakers of the language, have an incomplete education, have a disability, or for some other reason - the reading level of these everyday documents can limit accessibility and affect their well-being.

The need to present people with texts that are at a reading level which is suitable for them has motivated research into measuring readability of any given text in order to assess whether automatic simplification has rendered a more difficult text into a more readable one. Readability can be measured using tools which assess the reading level of a text. We define simplification as the process of changing a text to lower its reading level without removing necessary information or producing an ungrammatical result. This is similar to the definition of (cf. (Zhu et al., 2010)), except that we avoid defining a specific, limited, set of simplification operations. The Related Work section details research into measures of readability and work on automatic simplification systems.

We have begun to construct a large, accessible corpus of everyday documents. This corpus will eventually contain thousands of these documents, each having statistics characterising its contents, and multiple readability measures. Multiple different simplifications will be collected for the original documents and their content statistics and readability measures will be included in the corpus. This type of large and accessible corpus is of vital importance in driving development of automated text simplification. It will provide training material for the systems as well as a common basis of evaluating results from different systems.

Thus far, we have collected a basic corpus of everyday documents from a wide variety of sources. We plan to extend this basic corpus to create the much larger and more structured corpus that we describe here. We have also carried out a preliminary study to evaluate the feasibility of using crowdsourcing as one source of simplifications in the extended corpus. We have used Amazon Mechanical Turk (AMT) and collected 10 simplifications each for 200 sentences from the basic cor-

Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR) @ EACL 2014, pages 84–93,
Gothenburg, Sweden, April 26-30 2014. ©2014 Association for Computational Linguistics

pus to determine feasibility, a good experimental design, quality control of the simplifications, and time and cost effectiveness.

In the next section we discuss related work relevant to creating and evaluating a large corpus of everyday documents and their simplifications. In Section 3 we further demonstrate the need for a corpus of everyday documents. Section 4 presents a description of our existing basic corpus. Section 5 describes the details of the extended corpus and presents our evaluation of the feasibility of using crowdsourcing to generate human simplifications for the corpus. Section 6 shows how the extended corpus will be made accessible. Section 7 concludes and outlines the future work that we will undertake to develop the extended corpus.

2 Related Work

2.1 Readability Evaluation

Measures of readability are important because they help us assess the reading level of any document, provide a target for simplification systems, and help evaluate and compare the performance of different simplification systems. Several measures of readability have been proposed; DuBay (2004) counted 200 such measures developed by the 1980s and the number has grown, with more advanced automated measures introduced since then.

Early measures of readability such as the Flesch-Kincaid grade level formula (Kincaid et al., 1975) use counts of surface features of the text such as number of words and number of sentences. While these older measures are less sophisticated than more modern reading level classifiers, they are still widely used and reported and recent work has shown that they can be a good first approximation of more complex measures (Štajner et al., 2012).

More recent approaches use more complicated features and machine learning techniques to learn classifiers that can predict readability. For example, Heilman et al. (2007) combine a naive Bayes classifier that uses a vocabulary-based language model with a k-Nearest Neighbors classifier using grammatical features and interpolate the two to predict reading grade level. Feng et al. (2010) and François and Miltsakaki (2012) examine a large number of possible textual features at various levels and compare SVM and Linear Regression classifiers to predict grade level. Vajjala and Meurers (2012) reported significantly higher accuracy on a similar task using Multi-level Perceptron classification.

The above two methods of measuring readability can be computed directly using the text of a document itself. To evaluate the performance of a simplification system which aims to make texts easier to read and understand, it is also useful to measure improvement in individuals' reading and comprehension of the texts. Siddharthan and Katsos (2012) recently studied sentence recall to test comprehension; and Temnikova and Maneva (2013) evaluated simplifications using the readers' ability to answer multiple choice questions about the text.

2.2 Automated Text Simplification Systems

Since the mid-90s several systems have been developed to automatically simplify texts. Early systems used hand-crafted syntactic simplification rules; for example, Chandrasekar et al. (1996), one of the earliest attempts at automated simplification. Rule-based systems continue to be used, amongst others, Siddharthan (2006), Aluisio and Gasperin (2010), and Bott et al. (2012).

Many of the more recent systems are statistically-based adapting techniques developed for statistical machine translation. Zhu et al. (2010) train a probabilistic model of a variety of sentence simplification rules using expectation maximization with a parallel corpus of aligned sentences from Wikipedia and Simple Wikipedia. Woodsend and Lapata (2011) present a system that uses quasi-synchronous grammar rules learned from Simple Wikipedia edit histories. They solve an integer linear programming (ILP) problem to select both which sentences are simplified (based on a model learned from aligned Wikipedia-Simple Wikipedia articles) and what the best simplification is. Feblowitz and Kauchak (2013) use parallel sentences from Wikipedia and Simple Wikipedia to learn synchronous tree substitution grammar rules.

2.3 Corpora for Text Simplification

Presently there are limited resources for statistical simplification methods that need to train on a parallel corpus of original and simplified texts. As mentioned in the previous section, common data sources are Simple Wikipedia revision histories and aligned sentences from parallel Wikipedia and Simple Wikipedia articles. Petersen and Ostendorf

(2007) present an analysis of a corpus of 104 original and abridged news articles, and Barzilay and Elhadad (2003) present a system for aligning sentences trained on a corpus of parallel Encyclopedia Britannica and Britannica Elementary articles. Other work generates parallel corpora of original and simplified texts in languages other than English for which Simple Wikipedia is not available. For example, Klerke and Søgaard (2012) built a sentence-aligned corpus from 3701 original and simplified Danish news articles, and Klaper et al. (2013) collected 256 parallel German and simple German articles.

2.4 Crowdsourcing for Text Simplification and Corpus Generation

Crowdsourcing uses the aggregate of work performed by many non-expert workers on small tasks to generate high quality results for some larger task. To the best of our knowledge crowdsourcing has not previously been explored in detail to generate text simplifications. Crowdsourcing has, however, been used to evaluate the quality of automatically generated simplifications. Feblowitz and Kauchak (2013) used AMT to collect human judgements of the simplifications generated by their system and De Clercq et al. (2014) performed an extensive evaluation of crowdsourced readability judgements compared to expert judgements.

Crowdsourcing has also been used to generate translations. The recent statistical machine translation-inspired approaches to automated simplification motivate the possibility of using crowdsourcing to collect simplifications. Ambati and Vogel (2010) and Zaidan and Callison-Burch (2011) both demonstrate the feasibility of collecting quality translations using AMT. Post et al. (2012) generated parallel corpora between English and six Indian languages using AMT.

3 The Need for a Corpus of Everyday Documents

A high quality parallel corpus is necessary to drive research in automated text simplification and evaluation. As shown in the Related Work section, most statistically driven simplification systems have used parallel Wikipedia - Simple Wikipedia articles and Simple Wikipedia edit histories. The resulting systems take Wikipedia articles as input and generate simplified versions of those articles. While this demonstrates the possibility of automated text simplification, we believe that a primary goal for simplification systems should be to increase accessibility for those with poor reading skills to the texts which are most important to them. Creating a corpus of everyday documents will allow automated simplification techniques to be applied to texts from this domain. In addition, systems trained using Simple Wikipedia only target a single reading level - that of Simple Wikipedia. A corpus containing multiple different simplifications at different reading levels for any given original will allow text simplification systems to target specific reading levels.

The research needs that this corpus aims to meet are:

- A large and accessible set of original everyday documents to:
    - provide a training and test set for automated text simplification
- A set of multiple human-generated simplifications at different reading levels for the same set of original documents to provide:
    - accessible training data for automated text simplification systems
    - the ability to model how the same document is simplified to different reading levels
- An accessible location to share simplifications of the same documents that have been generated by different systems to enable:
    - comparative evaluation of the performance of several systems
    - easier identification and analysis of specific challenges common to all systems

4 Description of the Basic Corpus of Everyday Documents

We have collected a first set of everyday documents. This will be extended to generate the corpus described in the following section. The present documents are heavily biased to the domain of driving since they include driving test preparation materials from all fifty U.S. states. This section presents the information collected about each document and its organisation in the basic corpus. The basic corpus is available at: `https://dialrc.org/simplification/data.html`.

4.1 Document Fields

Each document has a name which includes information about the source, contents, and type of document. For example the name of the Alabama Driver Manual document is `al_dps_driver_man`. The corpus entry for each document also includes the full title, the document source (url for documents available online), the document type and domain, the date retrieved, and the date added to the corpus. For each document the number of sentences, the number of words, the average sentence length, the Flesch-Kincaid grade level score, and the lexical (L) and grammatical (G) reading level scores described in Heilman et al. (2007) are also reported. An example of an entry for the Alabama Driver Manual is shown in Table 1. The documents are split so that each sentence is on a separate line to enable easy alignments between the original and simplified versions of the documents.

Document Name	al_dps_driver_man
Full Title	Alabama Driver Manual
Document Type	Manual
Domain	Driver's Licensing
# Sentences	1,626
# Words	28,404
Avg. # words/sent	17.47
F-K Grade Level	10.21
Reading Level (L)	10
Reading Level (G)	8.38
Source	`http://1.usa.gov/1jjd4vw`
Date Added	10/01/2013
Date Accessed	10/01/2013

Table 1: Example basic corpus entry for Alabama Driver Manual

4.2 Corpus Statistics

There is wide variation between the different documents included in the corpus, across documents from different domains and also for documents from the same domain. This includes variability in both document length and reading level. For example, the driving manuals range from a lexical reading level of 8.2 for New Mexico to 10.4 for Nebraska. Table 2 shows the statistics for the different reading levels for the documents which have been collected, using the lexical readability measure and rounding to the nearest grade level. Table 3 shows the different domains for which documents have been collected and the statistics for the documents in those domains.

Reading Level (L)	# Documents	# Sentences
4	1	23
5	0	0
6	4	200
7	1	695
8	6	1,869
9	30	36,783
10	54	83,123
11	4	1,457
12	1	461

Table 2: Corpus statistics by lexical reading level

5 Description of an Extended Corpus of Everyday Documents

To meet the needs described in Section 3 the basic corpus will be extended significantly. We are starting to collect more everyday documents from each of the domains in the basic corpus and to extend the corpus to other everyday document domains including prescription instructions, advertising materials, mandatory educational testing, and operating manuals for common products. We are also collecting human-generated simplifications for these documents. We will open up the corpus for outside contributions of more documents, readability statistics and simplifications generated by various human and automated methods. This section describes what the extended corpus will contain and the preliminary work to generate simplified versions of the documents we presently have.

5.1 Document Fields

The extended corpus includes both original documents and their simplified versions. The original documents will include all the same information as the basic corpus, listed in Section 4.1. Novel readability measures for each document can be contributed. For each readability measure that is contributed, the name of the measure, document score, date added, as well as relevant references to the system used to calculate it will be included.

Multiple simplified versions of each original document can be contributed. The simplification for each sentence in the original document will be on the same line in the simplified document as the corresponding sentence in the original document. Each simplified version will include a brief description of how it was simplified and relevant references to the simplification method. As with the original documents, the date added, optional comments and the same document statistics and read-

Domain	# Documents	Avg. # Sentences	Avg. # Words	Avg. # words/sent	Total # Sentences	Total # Words	Avg. F-K Grade Level	Avg. Readability (L)	Avg. Readability (G)
Driver's Licensing	60	1927.6	30,352.6	16.1	115,657	1,821,155	9.54	9.6	7.9
Vehicles	3	46.7	1,118.3	22.5	140	3355	13.3	8.2	7.9
Government Documents	11	150	2,242.8	16.4	1650	24,671	10.5	8.6	8.4
Utilities	5	412.8	8,447.2	21.5	2,064	42,236	13.4	9.8	8.9
Banking	3	158	2,900	17.6	474	8,700	11.4	10.5	8.9
Leasing	4	101	2,386.8	23.8	404	9,547	13.7	9.0	8.7
Government Aid	10	317.4	5,197.5	17.4	3,174	51,975	10.7	9.2	8.8
Shopping	3	281	5,266.7	19.7	843	15,800	12.2	9.9	9.0
Other	2	102.5	1,634	16.0	205	3268	9.7	8.8	8.2
All	101	1,233.8	19,611.0	17.2	124,611	1,980,707	10.4	9.4	8.2

Table 3: Corpus statistics for the basic corpus documents

ability metrics will be included. Additional readability metrics can also be contributed and documented.

5.2 Generating Simplifications Using Crowdsourcing

We conducted a preliminary study to determine the feasibility of collecting simplifications using crowdsourcing. We used AMT as the crowdsourcing platform to collect sentence-level simplification annotations for sentences randomly selected from the basic corpus of everyday documents.

5.2.1 AMT Task Details

We collected 10 simplification annotations for each of the 200 sentences which we posted in two sets of Human Intelligence Tasks (HITs) to AMT. Each HIT included up to four sentences and included an optional comment box that allowed workers to submit comments or suggestions about the HIT. Workers were paid $0.25 for each HIT, and 11 workers were given a $0.05 bonus for submitting comments which helped improve the task design and remove design errors in the first iteration of the HIT design. The first set of HITs was completed in 20.5 hours and the second set in only 6.25 hours. The total cost for all 2000 simplification annotations was $163.51 for 592 HITs, each with up to four simplifications. The breakdown of this cost is shown in Table 4.

Item	Cost
592 HITs	$148.00
11 bonuses	$0.55
AMT fees	$14.96
Total	$163.51

Table 4: Breakdown of AMT costs

5.2.2 Quality Control Measures

To ensure quality, we provided a training session which shows workers explanations, examples, and counter-examples of multiple simplification techniques. These include lexical simplification, reordering, sentence splitting, removing unnecessary information, adding additional explanations, and making no change for sentences that are already simple enough. One of the training examples is the following:

> Original Sentence: ”Do not use only parking lights, day or night, when vehicle is in motion.”
>
> Simplification: ”When your vehicle is moving do not use only the parking lights. This applies both at night and during the day.”

The explanations demonstrated how lexical simplification, sentence splitting, and reordering techniques were used.

The training session also tested workers' abilities to apply these techniques. Workers were given four test sentences to simplify. Test 1 required lexical simplification. Test 2 was a counter-example of a sentence which did not require simplification. Test 3 required sentence splitting. Test 4 required either moving or deleting an unclear modifying clause. We chose the test sentences directly from the corpus and modified them where necessary to ensure that they contained the features being tested. Workers could take the training session and submit answers as many times as they wanted, but could not work on a task without first successfully completing the entire session. After completing the training session once, workers could complete as many HITs as were available to them.

In addition to the training session, we blocked submissions with empty or garbage answers (defined as those with more than 15% of the words not in a dictionary). We also blocked copy-paste functions to discourage worker laziness. Workers who submitted multiple answers that were either very close to or very far from the original sentence were flagged and their submissions were manually reviewed to determine whether to approve them. Similarity was measured using the ratio of Levenshtein distance to alignment length; Levenshtein distance is a common, simple metric for measuring the edit distance between two strings. The Levenshtein ratio $\left(1 - \frac{Levenshtein\ dist.}{alignment\ length}\right)$ provides a normalised similarity measure which is robust to length differences in the inputs. We also asked workers to rate their confidence in each simplification they submitted on a five point scale ranging from “Not at all” to “Very confident”.

5.2.3 Effectiveness of Quality Control Measures

To determine the quality of the AMT simplifications, we examine the effectiveness of the quality control measures described in the previous section.

Training: In addition to providing training and simplification experience to workers who worked on the task, the training session effectively blocked workers who were not able to complete it and spammers. Of the 358 workers who looked at the training session only 184 completed it (51%) and we found that no bots or spammers had completed the training session. Tables 5 and 6 show the performance on the four tests in the training session for workers who completed the training session and for those who did not, respectively.

# of workers	181
Avg. # Attempts Test 1	1.1
Avg. # Attempts Test 2	1.5
Avg. # Attempts Test 3	1.6
Avg. # Attempts Test 4	1.4

Table 5: Training statistics for workers who completed training

# of workers	174
# Completed Test 1	82
# Completed Test 2	47
# Completed Test 3	1

Table 6: Training statistics for workers who did not complete training

Blocking empty and garbage submissions: Empty simplifications and cut-paste functions were blocked using client-side scripts and we did not collect statistics of how many workers attempted either of these actions. One worker submitted a comment requesting that we do not block copy-paste functions. In total only 0.6% of submissions were detected as garbage and blocked.

Manual reviews: We (the first author) reviewed workers who were automatically flagged five or more times. We found that this was not an effective way to detect work to be rejected since there were many false positives and workers who did more HITs were more likely to get flagged. None of the workers flagged for review had submitted simplifications that were rejected.

5.2.4 Evaluating Simplification Quality

To determine whether it is feasible to use crowdsourced simplifications to simplify documents for the extended corpus, we examine the quality of the simplifications submitted. The quality control measures described in the previous sections are designed to ensure that workers know what is meant by simplification and how to apply some simplification techniques, to block spammers, and to limit worker laziness. However, workers were free to simplify sentences creatively and encouraged to use their judgement in applying any techniques that seem best to them.

It is difficult to verify the quality of the simplification annotations that were submitted or to determine how to decide what simplification to chose as the “correct” one for the corpus. For any given sentence there is no “right” answer for what the simplification should be; there are many different possible simplifications, each of which could be valid. For example, below is an original sentence taken from a driving manual with two of the simplifications that were submitted for it.

Original Sentence: ”Vehicles in any lane, except the right lane used for slower traffic, should be prepared to move to another lane to allow faster traffic to pass.”

Simplification 1: ”Vehicles that are not in the right lane should be prepared to move to another lane in order to allow faster traffic to pass.”

Simplification 2: ”Vehicles not in the right lane should be ready to move to another lane so faster traffic can pass them. The right lane is for slower traffic.”

There are a number of heuristics that could be used to detect which simplifications are most likely to be the best choice to use in the corpus.

The average time for workers to complete one HIT of up to four simplifications was 3.85 minutes. This includes the time to complete the training session during a worker's first HIT; excluding this, we estimate the average time per HIT is approximately 2.75 minutes. Simplifications which are completed in significantly less time, especially when the original sentence is long, can be flagged for review or simply thrown out if there are enough other simplifications for the sentence.

Workers' confidence in their simplifications can also be used to exclude simplifications which were submitted with low confidence (using worker confidence as a quality control filter was explored by Parent and Eskenazi (2010)). Table 7 shows the statistics for the worker-submitted confidences. Again, simplifications with very low confidence can either be reviewed or thrown out if there are enough other simplifications for the sentence.

Confidence Level	# of answers
1 (Not at all)	9
2 (Somewhat confident)	143
3 (Neutral)	251
4 (Confident)	1030
5 (Very confident)	567

Table 7: Self-assessed worker confidences in their simplifications

Worker agreement can also be used to detect simplifications that are very different from those submitted by other workers. Using the similarity ratio of Levenshtein distance to alignment length, we calculated which simplifications had at most one other simplification with which they have a similarity ratio above a specific threshold (here referred to as 'outliers'). Table 8 reports how many simplifications are outliers while varying the similarity threshold. Since there are many different valid simplifications possible for any given sentence this is not necessarily the best way to detect poor quality submissions. For example, one of the outliers, using the 50% threshold, was a simplification of the sentence "When following a tractor-trailer, observe its turn signals before trying to pass" which simplified by using a negative - "Don't try to pass ... without ...". This outlier was the only simplification of this sentence which used the negative but it is not necessarily a poor one. However, the results in Table 7 do show that there are many simplifications which are similar to each other, indicating that multiple workers agree on one simplification. One of these similar simplifications could be used in the corpus, or multiple different possible simplifications could be included.

Threshold	90%	85%	75%	65%	50%
# Outliers	1251	927	500	174	12

Table 8: Number of outlier simplifications with similarity ratio above the threshold for at most one other simplification

To further verify that usable simplifications can be generated using AMT the first author manually reviewed the 1000 simplifications of 100 sentences submitted for the first set of HITs. We judged whether each simplification was grammatical and whether it was a valid simplification. This is a qualitative judgement, but simplifications were judged to be invalid simplifications if they had significant missing or added information compared to the original sentence or added significant extra grammatical or lexical complexity for no apparent reason. The remaining grammatical, valid simplifications were judged as more simple, neutral, or less simple than the original for each of the following features: length, vocabulary, and grammatical structure. The results of this review are shown in Table 9. These results show that approximately 15% of the simplifications were ungrammatical or invalid, further motivating the need to use the other features, such as worker agreement and confidence, to automatically remove poor simplifications.

5.2.5 Extending the Corpus Using Crowdsourcing

The preliminary work undertaken demonstrates that it is feasible to quickly collect multiple simplifications for each sentence relatively inexpensively. We have also presented an evaluation of the quality of the crowdsourced simplifications and several methods of determining which simplifications could be used in the extended corpus. More work is still needed to determine the most cost effective way of getting simplification results that are of sufficient quality to use without gathering overly redundant simplifications for each sentence. Additionally, simplifications of more sentences are needed to assess improvements in reading level since the reading level measures we use are not accurate for very short input texts.

Un-grammatical	Invalid (excludes ungrammatical)	Simpler vocabulary	Less simple vocabulary	Equivalent vocabulary	Grammatically simpler	Less grammatically simple	Equivalent grammar	Longer	Shorter	Same length
35	122	383	21	596	455	21	524	99	537	364

Table 9: Manual evaluation of 1000 AMT simplifications. Numbers of simplifications with each feature.

6 Contributing to & Accessing the Corpus

6.1 Contributing to the Extended Corpus

The following items can be contributed to the corpus: original everyday copyright-free documents, manual or automated simplifications of the original documents (or parts of the documents), and readability scores for original or simplified documents.

Original documents submitted to the corpus can be from any domain. Our working definition of an everyday document is any document which people may have a need to access in their everyday life. Examples include government and licensing forms and their instructions, banking forms, prescription instructions, mandatory educational testing, leasing and rental agreements, loyalty program sign-up forms and other similar documents. We excluded Wikipedia pages because we found that many article pairs actually had few parallel sentences. Documents should be in English and of North American origin to avoid dialect-specific issues.

Hand generated or automatically generated simplifications of everyday documents are also welcome. They should be accompanied the information detailed in Section 5.1. The document statistics listed in Sections 4 and 5 will be added for each simplified document.

Readability scores can be contributed for any of the documents.They should also include the information detailed in Section 5.1 and pertinent information about the system that generated the scores.

6.2 Accessing the Extended Corpus

The extended corpus will be made publicly accessible at the same location as the basic corpus. The names and statistics of each of the documents will be tabulated and both the original and simplified documents, and their statistics, will be available to download. Users will submit their name or organizational affiliation along with a very brief description of how they plan to use the data. This will allow us to keep track of how the corpus is being used and how it could be made more useful to those researching simplification.

The goal of this corpus is to make its contents as accessible as possible. However, many of the original documents from non-governmental sources may not be freely distributed and will instead be included under a data license, unlike the remainder of the corpus and the simplifications[1].

7 Conclusions & Future Work

In this paper we have given the motivation for creating a large and publicly accessible corpus of everyday documents and their simplifications. This corpus will advance research into automated simplification and evaluation for everyday documents. We have already collected a basic corpus of everyday documents and demonstrated the feasibility of collecting large numbers of simplifications using crowdsourcing. We have defined what information the extended corpus will contain and how contributions can be made to it.

There is significantly more work which must be completed in the future to create an extended corpus which meets the needs described in this paper. There are three tasks that we plan to undertake in order to complete this corpus: we will collect significantly more everyday documents; we will manage a large crowdsourcing task to generate simplifications for thousands of the sentences in these documents; and we will create a website to enable access and contribution to the extended simplification corpus. By making this work accessible we hope to motivate others to contribute to the corpus and to use it to advance automated text simplification and evaluation techniques for the domain of everyday documents.

Acknowledgments

The authors would like to thank the anonymous reviewers for their detailed and helpful feedback and comments on the paper.

[1] Thanks to Professor Jamie Callan for explaining some of the issues with including these types of documents in our dataset.

References

Sandra Aluisio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: The porsimples project for simplification of portuguese texts. In *Proc. of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53. Association for Computational Linguistics.

Vamshi Ambati and Stephan Vogel. 2010. Can crowds build parallel corpora for machine translation systems? In *Proc. of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 62–65. Association for Computational Linguistics.

Regina Barzilay and Noemie Elhadad. 2003. Sentence alignment for monolingual comparable corpora. In *Proc. of the 2003 Conference on Empirical Methods in Natural Language Processing*, EMNLP '03, pages 25–32. Association for Computational Linguistics.

Stefan Bott, Horacio Saggion, and David Figueroa. 2012. A hybrid system for spanish text simplification. In *Proc. of the Third Workshop on Speech and Language Processing for Assistive Technologies*, pages 75–84. Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proc. of the 16th Conference on Computational Linguistics - Volume 2*, COLING '96, pages 1041–1044. Association for Computational Linguistics.

Orphée De Clercq, Veronique Hoste, Bart Desmet, Philip van Oosten, Martine De Cock, and Lieve Macken. 2014. Using the crowd for readability prediction. *Natural Language Engineering*, FirstView:1–33.

William H. DuBay. 2004. *The Principles of Readability*. Costa Mesa, CA: Impact Information, http://www.impact-information.com/impactinfo/readability02.pdf.

Dan Feblowitz and David Kauchak. 2013. Sentence simplification as tree transduction. In *Proc. of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 1–10. Association for Computational Linguistics.

Lijun Feng, Martin Jansche, Matt Huenerfauth, and Noémie Elhadad. 2010. A comparison of features for automatic readability assessment. In *Coling 2010: Posters*, pages 276–284. Coling 2010 Organizing Committee.

Thomas François and Eleni Miltsakaki. 2012. Do nlp and machine learning improve traditional readability formulas? In *Proc. of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 49–57. Association for Computational Linguistics.

Michael Heilman, Kevyn Collins-Thompson, Jamie Callan, and Maxine Eskenazi. 2007. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *HLT-NAACL 2007: Main Proceedings*, pages 460–467. Association for Computational Linguistics.

J. Peter Kincaid, Robert P. Fishburne Jr., Richard L. Rogers, and Brad S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command, Millington Tn.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a german/simple german parallel corpus for automatic text simplification. In *Proc. of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19. Association for Computational Linguistics.

Sigrid Klerke and Anders Søgaard. 2012. Dsim, a danish parallel corpus for text simplification. In *Proc. of the Eighth Language Resources and Evaluation Conference (LREC 2012)*, pages 4015–4018. European Language Resources Association (ELRA).

Gabriel Parent and Maxine Eskenazi. 2010. Toward better crowdsourced transcription: Transcription of a year of the let's go bus information system data. In *SLT*, pages 312–317. IEEE.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Proc. of Workshop on Speech and Language Technology for Education*, pages 69–72.

Matt Post, Chris Callison-Burch, and Miles Osborne. 2012. Constructing parallel corpora for six indian languages via crowdsourcing. In *Proc. of the Seventh Workshop on Statistical Machine Translation*, pages 401–409. Association for Computational Linguistics.

Advaith Siddharthan and Napoleon Katsos. 2012. Offline sentence processing measures for testing readability with users. In *Proc. of the First Workshop on Predicting and Improving Text Readability for target reader populations*, pages 17–24. Association for Computational Linguistics.

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

Irina Temnikova and Galina Maneva. 2013. The c-score – proposing a reading comprehension metrics as a common evaluation measure for text simplification. In *Proc. of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 20–29. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proc. of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173. Association for Computational Linguistics.

Sanja Štajner, Richard Evans, Constantin Orasan, , and Ruslan Mitkov. 2012. What can readability measures really tell us about text complexity? In *Proc. of the Workshop on Natural Language Processing for Improving Textual Accessibility (NLP4ITA)*.

Kristian Woodsend and Mirella Lapata. 2011. Wikisimple: Automatic simplification of wikipedia articles. In *Proc. of the Twenty-Fifth AAAI Conference on Artificial Intelligence*, pages 927–932.

Omar F. Zaidan and Chris Callison-Burch. 2011. Crowdsourcing translation: Professional quality from non-professionals. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1220–1229. Association for Computational Linguistics.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proc. of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361.