OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification

Sowmya Vajjala
Iowa State University, USA
sowmya@iastate.edu

Ivana Lučić
Iowa State University, USA
ilucic@iastate.edu

Abstract

This paper describes the collection and compilation of the OneStopEnglish corpus of texts written at three reading levels, and demonstrates its usefulness for through two applications - automatic readability assessment and automatic text simplification. The corpus consists of 189 texts, each in three versions (567 in total). The corpus is now freely available under a CC by-SA 4.0 license[1] and we hope that it would foster further research on the topics of readability assessment and text simplification.

1 Introduction

Automatic Readability Assessment (ARA), the task of assessing the reading difficulty of a text, is a well-studied problem in computational linguistics (cf. Collins-Thompson, 2014). A related problem is Automatic Text Simplification (cf. Siddharthan, 2014) which aims to generate simplified texts from complex versions. While most of the research on these problems focused on feature engineering and modeling, there is very little reported work about the creation of open access corpora that supports this research.

Corpora used in ARA were primarily derived from textbooks or news articles written for different target audiences. In most of the cases, the texts at different levels in these corpora are not comparable versions of each other, which would not help us develop fine-grained readability models which can identify what parts of texts are difficult compared to others, instead of having a single score for the whole text. Corpora of parallel texts simplified for different target reading levels can solve this problem, and support better ARA models. On the other hand, ATS systems by default need parallel corpora, and primarily relied on parallel sentence pairs from Wikipedia-Simple Wikipedia for training and evaluating the simplification models. While the availability and suitability of this corpus is definitely a positive aspect, the lack of additional corpora makes an evaluation of the generalizability of simplification approaches difficult.

In this background, we created a corpus aligned at text and sentence level, across three reading levels (beginner, intermediate, advanced), targeting English as Second Language (ESL) learners. To our knowledge, this is the first such free corpus in any language for readability assessment research. While a sentence aligned corpus from the same source was discussed in previous research, the current corpus is larger, and cleaner. In addition to describing the corpus, we demonstrate the usefulness of this corpus for automatic readability classification and text simplification. The corpus is freely available[2]. Its creation and relevance are described in the sections that follow: Section 2 describes other relevant corpus creation projects. Section 3 describes our corpus creation. Section 4 describes some preliminary experiments with readability assessment and text simplification using this corpus. Section 5 concludes the paper with pointers to future work.

2 Related Work

Washburne and Vogel (1926) and Vogel and Washburne (1928) can be considered one of the early works on corpora creation for readability research, where they collected a corpus of 700 books annotated by children in terms of reading difficulty. While there are other such efforts in the past century, corpora from those early projects are not available for current use. Contemporary approaches to readability assessment typically rely on compiling large corpora from the Web. The WeeklyReader magazine was used as a source

[1] https://creativecommons.org/licenses/by-sa/4.0/

[2] https://zenodo.org/record/1219041

Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications, pages 297–304
New Orleans, Louisiana, June 5, 2018. ©2018 Association for Computational Linguistics

for graded news texts in past ARA research (Petersen, 2007; Feng, 2010). Petersen and Ostendorf (2009) described a corpus of articles from Encyclopedia Britannica, where each article had a comparable ”Elementary Version”, which, however, is not freely available as far as we know. Vajjala and Meurers (2012) compiled WeeBit corpus, combining WeeklyReader with BBC BiteSize, and this corpus was used in several ARA approaches in the past few years. (Vajjala and Meurers, 2013) described a large corpus of age specific TV program transcripts from BBC, and (Napoles and Dredze, 2010) used a corpus of Wikipedia-Simple Wikipedia articles. (Hancke et al., 2012; Dell'Orletta et al., 2011; Gonzalez-Dios et al., 2014), describe such web-based corpus compilation efforts for German, Italian and Basque respectively.

Textbooks from school curricula were also used as training corpora for readability assessment models in the past (e.g., Heilman et al. (2008) for English, Berendes et al. (2017) for German, (Islam et al., 2012) for Bangla). In all these cases, the grade level of the text was decided based on the target reader group (according to the website/textbook) which was decided by either publishers or authors. Another way of creating such corpora is through human annotations. DeLite corpus Vor der Brück et al. (2008) for German legal texts, and van Oosten and Hoste (2011); Clercq et al. (2014) for Dutch texts describe crowd annotated resources whereas the common core standards corpus described in Nelson et al. (2012) is annotated by experts according to the common core guidelines on text complexity. Corpora created with such human annotations are expensive to obtain and hence, are generally smaller in size. Therefore, such corpora may not be sufficient to build new models, although they can serve as good evaluation datasets.

Primary concern with all these corpora is that the articles in different reading levels are not comparable versions of each other (except Encyclopedia Britannica). The only other publicly and/or freely accessible readability corpus that potentially has parallel and comparable texts in multiple reading levels is the NewsEla[3] corpus which is a corpus of manually simplified news texts. While the corpus is available for research under some license restrictions, it also addresses a different target audience, young L1 English learners. In this background, we release an openly accessible corpus of texts with text and sentence level mapping across three reading levels, targeting L2 learners of English.

In terms of sentence aligned corpora for text simplification, different versions of aligned Wiki-Simple Wikipedia sentences have been used in NLP research (Zhu et al., 2010; Coster and Kauchak, 2011; Hwang et al., 2015). Different supervised and unsupervised approaches were proposed to construct such corpora (Bott and Saggion, 2011; Klerke and Søgaard, 2012; Klaper et al., 2013; Brunato et al., 2016). Our corpus adds a new resource for the English text simplification task.

3 Corpus

Our corpus was compiled from onestopenglish.com over the period 2013–2016. onestopenglish.com is an English language learning resources website run by MacMillan Education, with over 700,000 users across 100 countries. One of the features of the website is a weekly news lessons section, which contains articles sourced from The Guardian newspaper, and rewritten by teachers to suit three levels of adult ESL learners (elementary, intermediate, and advanced). That is, content from the same original article is rewritten in three versions, to suit three reading levels. The advanced version is close to the original article, although not with exact same content. Texts from this source were previously used for training sentence level readability models (Vajjala and Meurers, 2016; Ambati et al., 2016; Howcroft and Demberg, 2017), for performing corpus analysis about the characteristics of simplified text (Allen, 2009), and in user studies about the relationship between text complexity and reading comprehension (Crossley et al., 2014; Vajjala et al., 2016), although the corpus was not publicly available in the past.

Original articles from the website consisted of pdf files containing the article text, some pre/post test questions, and other additional material. So, the first step in the corpus creation process involved removing the irrelevant content. We first explored off-the-shelf pdf to text converters, and while they worked, they did not always result in a clean text, sometimes missing entire pages of content. While this may not be a significant issue for

[3] https://newsela.com/

Reading Level	**Example**
Advanced (Adv)	*Amsterdam still looks liberal to tourists, who were recently assured by the Labour Mayor that the city's marijuana-selling coffee shops would stay open despite a new national law tackling drug tourism. But the Dutch capital may lose its reputation for tolerance over plans to dispatch nuisance neighbours to scum villages made from shipping containers.*
Intermediate (Int)	*To tourists, Amsterdam still seems very liberal. Recently the city's Mayor assured them that the city's marijuana-selling coffee shops would stay open despite a new national law to prevent drug tourism. But the Dutch capitals plans to send nuisance neighbours to scum villages made from shipping containers may damage its reputation for tolerance.*
Elementary (Ele)	*To tourists, Amsterdam still seems very liberal. Recently the city's Mayor told them that the coffee shops that sell marijuana would stay open, although there is a new national law to stop drug tourism. But the Dutch capital has a plan to send antisocial neighbours to scum villages made from shipping containers, and so maybe now people wont think it is a liberal city any more.*

Table 1: Example sentences for three reading levels

doing text level classification, it becomes important when we try to align sentences or use this corpus for any qualitatiokave analyses. Hence, one of the authors manually went through all the files, comparing with the pdf version, to ensure there are no missing pages/content, resulting in a clean corpus[4]. An example of the degree of simplification performed is shown in Table 1.

Table 2 contains some descriptive statistics about the final corpus. As expected, advanced texts are much longer than elementary texts. However, the standard deviation for each level is also high, indicating that text length may not be the deciding factor in terms of reading level.

Reading level	Avg. Num. Words	Std. Dev
Elementary	533.17	103.79
Intermediate	676.59	117.15
Advanced	820.49	162.52

Table 2: Descriptive Statistics about the corpus

We performed some preliminary corpus analysis of the three reading levels in terms of some common features used in readability literature. Table 3 shows the summary of these results, using traditionally used features such as Flesch-Kincaid Grade Level (FKGL) (Kincaid et al., 1975) and Type-token ratio (TTR), and occurrences of different phrases, as given by Stanford Parser (Chen and Manning, 2014). In general, all feature values decrease from ADV to ELE, which is expected, if we assume all these features to be indicative of reading level of text.

Feature	**ADV**	**INT**	**ELE**
FKGL	9.5	8.2	6.4
TTR	0.56	0.432	0.42
avg num. NP	6.08	5.52	4.92
avg num. VP	4.49	4.03	3.49
avg num. PP	2.72	2.30	1.82

Table 3: Some of the features across reading levels

Sentence Alignment: A sentence aligned version was created using cosine similarity, taking one pair of reading levels at a time and performing a one-to-all comparison of sentences in both texts. We chose a similarity range of [0.7-0.95] for pairing sentences, after experimenting with several thresholds. The reason for choosing 0.95 instead of 1 is that there were some sentences with only a change of punctuation, which we did not want in our sentence aligned data. The final sentence aligned corpus had 1674 pairs for ELE-INT, 2166 pairs for ELE-ADV and 3154 pairs for INT-ADV. On an average, INT-ADV sentence pairs had a higher degree of similarity (0.9) than ELE-ADV (0.77) or ELE-INT (0.85).

[4] We acquired permission both from Onestopenglish.com and The Guardian to release this plain-text version of the corpus.

4 Experiments

We demonstrate the usefulness of this corpus for two applications: readability assessment and text simplification.

4.1 Readability Assessment

We modeled this as a classification problem using both generic text classification features such as word ngrams as well as features typically used in readability classification research[5]. Generic text classification features include:

1. Word n-grams: Uni, Bi, Trigram features
2. POS n-grams: Bi and Trigrams of POS tags from Stanford tagger (Toutanova et al., 2003)
3. Character n-grams: 2–5 character n-grams, considering word boundaries
4. Syntactic production rules: phrase structure production rules from Stanford parser (Klein and Manning, 2003)
5. Dependency relations: Dependency relation triplets of the form (relation, head, word) from Stanford dependency parser (Chen and Manning, 2014)

All n-gram features and grammar rules/relations that occurred at least 5 times in the entire corpus were retained for the final feature set. All these features were extracted using LightSide text mining workbench (Mayfield and Rosé, 2013). Table 4 shows the classification results with these features, using Sequential Minimal Optimization (SMO) classifier with linear kernel (with a random baseline of 33% as all classes are represented equally in the data).

Features	Accuracy
Word n-grams	61.38%
POS n-grams	67.37%
Char n-grams	**77.25**%
Syntactic Production Rules	54.67%
Dependency Relations	27.16%

Table 4: Text Classification Results with generic features

Character ngrams seem to be the best performing group of generic features, achieving 77% accuracy. Data-driven features that rely on deeper linguistic representations seem to perform poorly compared to these simple features. Particularly, dependency relations perform worse than the random baseline. Since we are working with parallel texts, there will be a lot of word level overlap across reading levels, and hence, it is not entirely surprising to see word n-grams not doing well. However, despite this, character n-grams seems to do well. We speculate they capture sub-word simplified text information such as usage of certain suffixes or prefixes, which has to be further explored in future.

In addition to the generic features, we also trained classifiers with features that are typically used in ARA research. These are:

1. Traditional features and formulae, that have been used in all the ARA models in the past
2. lexical variation, type token ratio, and POS tag ratio based features
3. Features based on psycholinguistic databases
4. Features based on constituent parse trees
5. Discourse features include:
    - overlap measures among sentences in a document as used in Coh-Metrix (Graesser et al., 2014)
    - usage of different kinds of connectives obtained from the discourse connectives tagger (Pitler and Nenkova, 2009)
    - coreference chains in the text from Stanford CoreNLP

Table 5 summarizes the results from these experiments[6].

Feature Group	Num. Feats.	Accuracy
Traditional	10	58.5%
Word	10	67.19&
Psycholinguistic	11	52.02%
LexVar, POS	29	72.48%
Syntactic Features	28	73.89%
Discourse Features	67	63.66%
Total	**155**	**78.13**%

Table 5: Text Classification Results with specific linguistic complexity features

[5] full feature file is provided in the supplementary material.

[6] Code for feature extraction is available at: https://bitbucket.org/nishkalavallabhi/complexity-features

Highest classification accuracy is achieved when all the features are put together, as shown in Table 5. However, this only results in a less than 1% improvement over character n-grams. Character n-grams as features for readability assessment were not explored in the past, and these results would lead us to explore that in future. In terms of comparison with existing work on ARA, highest accuracies reported are close to 90% on WeeBit dataset (Vajjala and Meurers, 2012). However, considering that we are comparing texts on the same topic, differing primarily in terms of style rather than content, this is perhaps a difficult dataset to model, compared to other existing readability datasets.

Since we now have a corpus with parallel versions of sentences and paragraphs at different reading levels, one idea to explore further is to model readability assessment as a sentence and paragraph level pair-wise ranking problem, and then use those ”local” readability assessments to infer ”global” text level readability (e.g., Chapter 5.5, Vajjala (2015)). Previous research also (Ma et al., 2012) showed that pair-wise ranking resulted in better readability models than classification. A combination of both these approaches would be an interesting dimension to explore in future.

4.2 Text Simplification

Automatic Text Simplification (ATS) has been commonly modeled as a Phrase Based Machine Translation (PBMT) problem in the literature. To demonstrate the usefulness of this corpus for ATS experiments, we used the *adv-ele* sentence aligned version of the OSE corpus and treated it as a phrase based machine translation problem. We split the dataset with 2166 sentence pairs into - 1000 sentence pairs for training, 500 for development, and the remaining 666 pairs for testing. We did not explore a neural model, due to the size of the dataset considered. We used Moses (Hoang et al., 2007) to train the model, and evaluated the model performance on test data in terms of various evaluation metrics used in MT research, comparing machine generated and human translations.

This model resulted in a BLEU (Papineni et al., 2001) score of 54.45 and METEOR (Denkowski and Lavie, 2014) score of 46. While the scores are not interpretable by themselves, general guidelines by Lavie (2011) suggest that BLEU and METEOR scores above 50 indicate understandable translations. Comparing with existing results on ATS, Zhang and Lapata (2017) trained a neural network based MT model with 300K sentence pairs as training data, and reported a much higher BLEU score of 88.85. The results on current dataset (with 1000 sentence training data and PBMT) cannot be compared with this result though, especially considering the size of the dataset. However, previous research showed that a high BLEU score with one corpus did not generalize when the test set came from another source (Chapter 6 in Vajjala, 2015). While our dataset may not be sufficient to build robust text simplification models, it can be used to test the generalizability of such state of the art text simplification approaches, or to be combined with a larger dataset while training a simplification model.

5 Conclusion

In this paper, we described the creation of a new corpus for readability assessment and text simplification research, and demonstrated its usefulness for readability assessment and text simplification research. The corpus is released with this paper, and we hope it will foster further research into readability assessment and text simplification systems aimed at ESL learners.

Beyond researchers interested in computational modeling, this corpus is also useful for other groups such as: a) researchers interested in conducting user studies about the relationship between text simplification and reader comprehension, or between expert annotated readability labels and target reader comprehension of texts (e.g., Vajjala et al. (2016)) and b) researchers interested in doing corpus studies with simplified and unsimplified texts, which can give insights into creating both manual and automatically simplified texts (e.g., (Allen, 2009)).

Acknowledgements:

We thank the onestopenglish.com team, and The Guardian, for allowing us to release the corpus for research use, and the anonymous reviewers for their useful comments. We also thank Detmar Meurers, University of Tübingen, for supporting the corpus creation in early stages.

References

David Allen. 2009. A study of the role of relative clauses in the simplification of news texts for learners of English. *System*, 37(4):58–599.

Bharat Ram Ambati, Siva Reddy, and Mark Steedman. 2016. Assessing relative sentence complexity using an incremental ccg parser. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1051–1057, San Diego, California. Association for Computational Linguistics.

Karin Berendes, Sowmya Vajjala, Detmar Meurers, Doreen Bryant, Wolfgang Wagner, Maria Chinkina, and Ulrich Trautwein. 2017. Reading demands in secondary school: Does the linguistic complexity of textbooks increase with grade level and the academic orientation of the school track? *Journal of Educational Psychology*.

S Bott and H Saggion. 2011. An unsupervised alignment algorithm for text simplification corpus construction. In *ACL Workshop on Monolingual Text-to-Text Generation*.

Tim Vor der Brück, Hermann Helbig, and Johannes Leveling. 2008. The readability checker delite. Technical Report Technical Report 345-5/2008, Fakultät für Mathematik und Informatik, FernUniversität in Hagen.

Dominique Brunato, Andrea Cimino, Felice Dell'Orletta, and Giulia Venturi. 2016. Paccss-it: A parallel corpus of complex-simple sentences for automatic text simplification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 351–361.

Danqi Chen and Christopher Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 740–750, Doha, Qatar. Association for Computational Linguistics.

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

Kevyn Collins-Thompson. 2014. Computational assessment of text readability: A survey of past, present, and future research. *International Journal of Applied Linguistics*, 165(2):97–135.

William Coster and David Kauchak. 2011. Simple English wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Scott A. Crossley, Hae Sung Yang, and Danielle S. McNamara. 2014. What's so simple about simplified texts? a computational and psycholinguistic investigation of text comprehension and text processing. *Reading in a Foreign Language*, 26(1):92–113.

Felice Dell'Orletta, Simonetta Montemagni, and Giulia Venturi. 2011. Read-it: Assessing readability of Italian texts with a view to text simplification. In *Proceedings of the 2nd Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380.

Lijun Feng. 2010. *Automatic Readability Assessment*. Ph.D. thesis, City University of New York (CUNY).

Itziar Gonzalez-Dios, María Jesús Aranzabe, Arantza Díaz de Ilarraza, and Haritz Salaberri. 2014. Simple or complex? assessing the readability of basque texts. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 334–344, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Arthur C. Graesser, Danielle S. McNamara, Zhiqang Cai, Mark Conley, Haiying Li, and James Pennebaker. 2014. Coh-metrix measures text characteristics at multiple levels of language and discourse. *The Elementary School Journal*, 115(2):pp. 210–229.

Julia Hancke, Detmar Meurers, and Sowmya Vajjala. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1063–1080, Mumbay, India.

Michael Heilman, Kevyn Collins-Thompson, and Maxine Eskenazi. 2008. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the 3rd Workshop on Innovative Use of NLP for Building Educational Applications at ACL-08*, Columbus, Ohio.

Hieu Hoang, Alexandra Birch, Chris Callison-burch, Richard Zens, Rwth Aachen, Alexandra Constantin, Marcello Federico, Nicola Bertoldi, Chris Dyer, Brooke Cowan, Wade Shen, Christine Moran, and Ondřej Bojar. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45nd Annual Meeting of the Association for Computational Linguistics (ACL '07)*.

David M Howcroft and Vera Demberg. 2017. Psycholinguistic models of sentence processing improve sentence readability ranking. In *Proceedings of the 15th Conference of the European Chapter of* *the Association for Computational Linguistics: Volume 1, Long Papers*, volume 1, pages 958–968.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning sentences from standard wikipedia to simple wikipedia. In *Human Language Technologies: The 2015 Annual Conference of the North American Chapter of the ACL*, pages 211–217.

Zahurul Islam, Alexander Mehler, and Rashedur Rahman. 2012. Text readability classification of textbooks of a low-resource language. In *26th Pacific Asia Conference on Language,Information and Computation pages*.

J. P. Kincaid, R. P. Jr. Fishburne, R. L. Rogers, and B. S Chissom. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count and Flesch Reading Ease formula) for Navy enlisted personnel. Research Branch Report 8-75, 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 *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Meeting of the Association for Computational Linguistics (ACL 2003)*, pages 423–430, Sapporo, Japan.

Sigrid Klerke and Anders Søgaard. 2012. Danish parallel corpus for text simplification. In *In Proceedings of Language Resources and Evaluation Conference (LREC), 2012*.

Alon Lavie. 2011. Evaluating the output of machine translation systems. *MT Summit Tutorial*, page 86.

Yi Ma, Eric Fosler-Lussier, and Robert Lofthus. 2012. Ranking-based readability assessment for early primary children's literature. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, NAACL HLT '12, pages 548–552, Stroudsburg, PA, USA. Association for Computational Linguistics.

Elijah Mayfield and Carolyn Penstein Rosé. 2013. Open source machine learning for text. *Handbook of automated essay evaluation: Current applications and new directions*.

Courtney Napoles and Mark Dredze. 2010. Learning simple wikipedia: a cogitation in ascertaining abecedarian language. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing: Writing Processes and Authoring Aids*, CL&W '10, pages 42–50, Stroudsburg, PA, USA. Association for Computational Linguistics.

J. Nelson, C. Perfetti, D. Liben, and M. Liben. 2012. Measures of text difficulty: Testing their predictive value for grade levels and student performance. Technical report, The Council of Chief State School Officers.

Philip van Oosten and Véronique Hoste. 2011. Readability annotation: replacing the expert by the crowd. In *Proceedings of the 6th Workshop on Innovative Use of NLP for Building Educational Applications*, IUNLPBEA '11, pages 120–129, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2001. Bleu: A method for automatic evaluation of machine translation. Technical report, IBM Research Division, Thomas J. Watson Research Center.

Sarah E. Petersen. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Ph.D. thesis, University of Washington.

Sarah E. Petersen and Mari Ostendorf. 2009. A machine learning approach to reading level assessment. *Computer Speech and Language*, 23:86–106.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics: Special issue on Recent Advances in Automatic Readability Assessment and Text Simplification*, 165:2:259–298.

K. Toutanova, D. Klein, C. Manning, and Y. Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *HLT-NAACL*, pages 252–259, Edmonton, Canada.

Sowmya Vajjala. 2015. *Analyzing Text Complexity and Text Simplification: Connecting Linguistics, Processing and Educational Applications*. Ph.D. thesis, University of Tübingen.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *In Proceedings of the 7th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2013. On the applicability of readability models to web texts. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 59–68.

Sowmya Vajjala and Detmar Meurers. 2014. Readability assessment for text simplification: From analyzing documents to identifying sentential simplifications. *International Journal of Applied Linguistics, Special Issue on Current Research in Readability and Text Simplification*, 165(2):142–222.

Sowmya Vajjala and Detmar Meurers. 2016. Readability-based sentence ranking for evaluating text simplification. *arXiv preprint arXiv:1603.06009*.

Sowmya Vajjala, Detmar Meurers, Alexander Eitel, and Katharina Scheiter. 2016. Towards grounding computational linguistic approaches to readability: Modeling reader-text interaction for easy and difficult texts. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 38–48, Osaka, Japan. The COLING 2016 Organizing Committee.

Mabel Vogel and Carleton Washburne. 1928. An objective method of determining grade placement of children's reading material. *Elementary School Journal*, 38:58–66.

Carleton Washburne and Mabel Vogel. 1926. *Winnetka graded book list*, volume 1. American Library Association.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594.

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

A Supplemental Material

The corpus, and some processed output files are available at: `https://github.com/nishkalavallabhi/OneStopEnglishCorpus`. An example is shown below: