A Dataset of Word-Complexity Judgements from Deaf and Hard-of-Hearing Adults for Text Simplification

Oliver Alonzo
Rochester Institute of Technology
Rochester, NY 14623
oa7652@rit.edu

Sooyeon Lee
New Jersey Institute of Technology
Newark, NJ 07102
sooyeon.lee@njit.edu

Mounica Maddela and **Wei Xu**
Georgia Institute of Technology
Atlanta, GA 30332
mmaddela3@gatech.edu, wei.xu@cc.gatech.edu

Matt Huenerfauth
Rochester Institute of Technology
Rochester, NY 14623
matt.huenerfauth@rit.edu

Abstract

Research has explored the use of automatic text simplification (ATS), which consists of techniques to make text simpler to read, to provide reading assistance to Deaf and Hard-of-hearing (DHH) adults with various literacy levels. Prior work in this area has identified interest in and benefits from ATS-based reading assistance tools. However, no prior work on ATS has gathered judgements from DHH adults as to what constitutes complex text. Thus, following approaches in prior NLP work, this paper contributes new word-complexity judgements from 11 DHH adults on a dataset of 15,000 English words that had been previously annotated by L2 speakers, which we also augmented to include automatic annotations of linguistic characteristics of the words. Additionally, we conduct a supplementary analysis of the interaction effect between the linguistic characteristics of the words and the groups of annotators. This analysis highlights the importance of collecting judgements from DHH adults for training ATS systems, as it revealed statistically significant interaction effects for nearly all of the linguistic characteristics of the words.

1 Introduction

Automatic text simplification (ATS) consists of computing techniques that make text simpler to read, while preserving the meaning of the original text (Shardlow, 2014; Siddharthan, 2014; Al-Thanyyan and Azmi, 2021). ATS can be applied at the lexical level by replacing complex words with simpler synonyms, at the syntactic level by rewriting sentences to reduce their syntactic complexity, or by doing both at the same time (Shardlow, 2014; Siddharthan, 2014; Al-Thanyyan and Azmi, 2021). Prior work has explored the use of ATS to provide reading assistance to different user groups, including non-native speakers (henceforth referred to as *L2 speakers*) and Deaf and Hard-of-hearing (DHH) adults because of their diversity in literacy skill (e.g., Azab et al., 2015; Alonzo et al., 2020; Kushalnagar et al., 2018; Ehara et al., 2010).

While there have been many efforts into the use of ATS for assistive applications, most ATS research from a natural language processing (NLP) perspective focuses on improving the machine learning models supporting those applications. However, training data for these models is scarce as texts are not usually written at various levels of linguistic complexity and thus the access to the needed corpora is limited (e.g., Simple Wikipedia or Newsela) (Coster and Kauchak, 2011; Xu et al., 2015).

In recent work, researchers created a dataset of 15,000 English words from a general lexicon, obtaining word-complexity judgement on all 15,000 words from adult L2 speakers (Maddela and Xu, 2018). Using this dataset to train simplification models provided promising results (Maddela and Xu, 2018). However, considering that research has identified that various linguistic characteristics may affect text complexity differently for different reader groups (Paetzold and Specia, 2016b), researchers have called for the creation of datasets with judgements from intended target audiences (Gooding, 2022; Maddela and Xu, 2018). Prior work identified benefits from lexical simplification among DHH adults (Alonzo et al., 2020); thus, we collect judgements from DHH adults on the lexicon previously created in Maddela and Xu (2018). Then, to understand whether there is indeed value in gathering these judgements from annotators from target reader groups, we conduct an analysis of complexity judgements, to determine

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 119 - 124
December 8, 2022

whether there was an interaction effect between the annotator groups and various linguistic characteristics of words, which had been identified as relevant for word complexity in prior work.

As the main contribution of this paper, we collect and publicly release[1] word-complexity judgements from DHH adults (and automatically-computed linguistic characteristics) on a set of 15,000 words, which had previously been annotated by L2 speakers in prior work. As a supplementary contribution, we provide an additional analysis of the interaction effects between the groups of annotators and the linguistic characteristics identified, which highlight the importance of collecting word-complexity judgements with annotators from different reader groups.

2 Related Work

Advances in ATS have motivated research into its use to support the reading tasks of various groups of people, including people with disabilities such as dyslexia or aphasia (e.g., Rello et al., 2013a; Devlin and Unthank, 2006) , or people who are DHH (e.g., Alonzo et al., 2020), as well as children (e.g., De Belder and Moens, 2010; Xu et al., 2015) or foreign language learners (e.g., Azab et al., 2015; Ehara et al., 2010.) A key challenge in the field, however, is obtaining access to datasets (Xu et al., 2015) and there have been calls in the community for the collection of datasets from people from the intended audiences for the systems (Paetzold and Specia, 2016b; Maddela and Xu, 2018; Gooding, 2022). In the next section, we summarize work on obtaining datasets for ATS and motivate our approach.

2.1 Datasets for Automatic Text Simplification

As mentioned in the introduction, there are various approaches to automatic text simplification, including lexical and syntactic approaches (Shardlow, 2014; Al-Thanyyan and Azmi, 2021), and there have been efforts to create datasets for both of these tasks (Xu et al., 2015; Al-Thanyyan and Azmi, 2021). When it comes to syntactic simplification, most approaches require sentence-aligned training data. Thus, researchers have typically created datasets based on aligning the sentences of existing resources that provide texts at different levels of complexity. These include the articles provided in Simple English in Wikipedia (Kauchak, 2013; Jiang et al., 2020), as well as news articles from Newsela, a website that provides news articles with human-produced simplifications (Xu et al., 2015).

When it comes to lexical simplification, the two main tasks that require the use of datasets are the complex word identification (CWI) stage, where systems identify potential words to simplify, and the substitution generation (SG) stage, where systems identify potential synonyms to replace a complex word (Shardlow, 2014; Paetzold and Specia, 2016b). While sentence-aligned datasets can also be used for lexical simplification by identifying complex words that have been replaced, most datasets created specifically for lexical simplification are obtained by having readers judge individual isolated word forms (e.g., Maddela and Xu, 2018; Gooding and Tragut, 2022) or identify complex words in sentences (e.g., Paetzold and Specia, 2016b). There are trade-offs with these approaches, including the fact that judging individual words is less time consuming, but identifying complex words in sentences may also provide insights into how a reader may judge a particular word in context, which is especially relevant for polysemous words.

As prior work has highlighted, many of the datasets presented in the literature are not targeted to any specific group (Xu et al., 2015; Gooding, 2022). However, evidence supports the need for collecting datasets with people from specific target audiences for ATS, including the fact that what makes text complex may vary depending on various characteristics of a reader group (Paetzold and Specia, 2016b). While prior work has identified benefits from both syntactic and lexical simplification for people who are DHH (Kushalnagar et al., 2018; Alonzo et al., 2020), to the best of our knowledge no prior work has gathered datasets of judgements from DHH adults.

3 The Dataset

Our dataset was originally gathered by researchers in Maddela and Xu (2018) by selecting the 15,000 most frequent words in Google's IT Ngram Corpus. Word-complexity judgements on all 15,000 words were also obtained from 11 L2 English speakers using a 6-point scale, going from "very simple" (1) to "very complex" (6), and using 6 points to avoid a neutral choice (Maddela and Xu, 2018). In this new work, we expand this dataset by obtaining word-

[1] https://github.com/oliveralonzo/DHH-lexical-dataset

complexity judgements from 11 DHH annotators following that prior approach, and we also compute several linguistic characteristics of the words. The selection of these linguistic characteristics was based on prior work, which had identified linguistic characteristics (e.g., word length or number of syllables) that had affected text complexity for various reader groups (Paetzold and Specia, 2016b).

3.1 Annotators and Annotation Process

Our 11 annotators were hired as part-time research assistants over 3 academic years at the Rochester Institute of Technology. All annotators identified as DHH, and their reported first languages included: ASL alone, English alone, ASL and English, and Chinese. At the beginning of their employment, our annotators completed the sentence-comprehension sub-test from the Wide Range Achievement Test 4 (WRAT-4), which had previously been validated as a measure of DHH's adults literacy skill. Their average WRAT-4 scores were 81 (SD = 11.62, range = 73 - 111), which is slightly below the U.S. average of 100.

Following the approach of Maddela and Xu (2018), we provided our annotators with the list of individual words, and asked them to provide a complexity judgement for each word using a 6-point scale where 1 meant "very simple" and 6 meant "very complex." A 6-point scale was employed to avoid a neutral choice. Furthermore, participants were instructed to rate a word as -1 if they considered that it was not a word.

3.2 Linguistic Characteristics

Prior work (Paetzold and Specia, 2016b) had identified that the relationship between various linguistic characteristics of words and their perceived complexity for a reader may vary depending upon the reader group. Thus, to investigate whether perceptions of word complexity among DHH annotators differed from those of non-DHH annotators in prior work (Maddela and Xu, 2018), we computed various linguistic characteristics for each word in the dataset. Notably, these characteristics were computed separately from the annotation process, so our annotators did not see those characteristics during the annotation process described above in section 3.1. Similar to linguistic properties investigated in prior work (e.g., Paetzold and Specia, 2016b), our characteristics were grouped into three categories: morphological, semantic and lexical features. These characteristics were computed using a Python script and employing publicly-available libraries as detailed below.

3.2.1 Morphological Features

The morphological features included **word length** and the **number of syllables**. Word length was computed using Python's built-in function for string variables, while the number of syllables was computed using the 'pronouncing' Python module[2], which provides an interface for the CMU Pronouncing Dictionary.

3.2.2 Semantic Features

The semantic features included the number of **senses** (possible meanings for a word), **synonyms** (words with the same meaning), **hypernyms** (words that a specific word is a type of, e.g., 'number' is a hypernym of 'five') and **hyponyms** (words that are a type of a specific word, e.g., 'five' is a hyponym of 'number'). All of these semantic features were computed using the Natural Language ToolKit (NLTK) implementation of WordNet[3].

3.2.3 Lexical Features

These lexical features consisted of **unigram log-probabilities** based on their frequency on three corpora used in: **SubIMDB**, a dataset comprised of 38,102 subtitles obtained from OpenSubtitles and IMDB (Paetzold and Specia, 2016a); **Subtlex**, a dataset of 50 million English words containing their word frequencies based on American movies and TV shows (Brysbaert and New, 2009); and **Simple Wikipedia**, a dataset of articles from Wikipedia written in the Simple English language (Kauchak, 2013). The unigram log-probabilities were computed using the NLTK toolkit.

4 Dataset Analysis and Results

4.1 Descriptive Statistics

When combining all the data from all of the DHH annotators, their average word-complexity judgements were 2.2 (SD = 0.67), where 1 meant "very simple" and 6, "very complex." The average word-complexity judgements previously obtained from L2 speakers in Maddela and Xu (2018), in turn, were 2.7 (SD = 0.83). Table 1a provides descriptive statistics for each of the linguistic characteristics.

Following the approach of Maddela and Xu (2018) and Agirre et al. (2014), we computed the average of the Pearson correlation between each

[2]https://pronouncing.readthedocs.io/en/latest/
[3]https://www.nltk.org/howto/wordnet.html

	a) Descriptive Statistics			b) Interaction Effect	
Linguistic Characteristics	**Average**	**SD**	**Range**	**F Value**	**Statistical Significance**
Length	7.1	2.4	1 to 18	15.42	Yes; $p < 0.001$
Syllables	2.3	1	0 to 7	26.64	Yes; $p < 0.001$
Senses	4.5	5.5	0 to 75	1.84	Yes; $p < 0.001$
Synonyms	6.9	8.9	0 to 100	0.82	No; $p = 0.87$
Hyponyms	3.4	4.6	0 to 59	0.55	No; $p = 1$
Hypernyms	11.5	28.7	0 to 693	1.81	Yes; $p < 0.001$
SubIMDB Unigram Log-probability	-17.5	3.2	-26.4 to -6.5	3.01	Yes; $p < 0.001$
Subtlex Unigram Log-probability	-17.6	3.5	-24.4 to -6.2	2.85	Yes; $p < 0.001$
Simple Wikipedia Unigram Log-probability	-16.5	2.6	-22.2 to -7.5	2.18	Yes; $p < 0.001$

Table 1: A summary of a) the descriptive statistics for each of the linguistic characteristics for all words in the dataset of the 15,000 most frequent English words (Maddela and Xu, 2018), and b) the results of the interaction effect between the group of annotators (DHH or L2 speakers) and each linguistic characteristic.

annotator's annotations and the average of the rest of the annotators, to assess the quality of the annotations. The average inter-annotator agreement for the annotations was 0.53, which is in line with the agreement observed in Maddela and Xu (2018) before removing outliers. Following their approach to identify outliers (i.e. defining an outlier as an annotation that had an absolute difference ≥ 2 from the average of the rest of the annotators) resulted in 11.2% of the annotations being identified as outliers. After removing those, the average agreement was 0.55. However, we release our dataset and present our results *without* removing any outliers as the definition of an outlier may vary depending on the application.

4.2 Interaction Effects

While it is possible to conduct significant difference testing between the overall judgements of DHH and L2 annotators (which, in fact, revealed significant differences), it may not be meaningful as it may simply suggest that the way the two groups of annotators calibrated to the scale was different. Furthermore, we were not concerned with identifying exactly what features *correlate* with word complexity in our dataset as those would be better identified through machine-learning models trained using this dataset. Instead, we were interested in whether, when conducting a two-factor analysis of the average judgements obtained from both groups, there is an interaction effect between the group and the linguistic characteristics outlined above. An interaction effect occurs when the effect of one independent variable on a dependent variable depends on another independent variable. Thus, an interaction effect would suggest that the way these various linguistic characteristics affect word complexity may be different for these two groups of annotators, thereby further motivating the need to collect datasets from specific groups of annotators who, in our case, were DHH adults.

Thus, we conducted two-way analyses of variance (ANOVA) where the dependent variable was the average judgements from annotators, and the independent variables were each of the linguistic characteristics of the words and the group of annotators. Overall, we observed interaction effects between the group of annotators and nearly all of the linguistic characteristics, with the exception of the number of synonyms and hyponyms. Table 1b provides the detailed results for these analyses.

5 Discussion and Conclusion

Our results provide further evidence for the importance of gathering judgments from intended audiences to train systems using datasets based on those judgements. Prior work had suggested that the linguistic characteristics that affect text complexity for different user groups may vary (e.g. word length may affect word complexity for people with dyslexia but less so for L2 speakers) (Rello et al., 2013b; Paetzold and Specia, 2016b). Through our analysis of interaction effects, we observed that the way *several* linguistic characteristics affect word-complexity judgements depends indeed on the group of annotators that provide those judgements. Thus, it is important to gather judgements from target audiences and build models based on those judgements, which may better capture the nuanced relations between how these different features impact word complexity for target audiences.

Limitations and Future Work

Our work presented in this paper had various limitations, and opens several avenues for future work:

1. There are different approaches to gather word-complexity judgements. Our dataset is limited in that it provides out-of-context judgements from DHH annotators. Thus, it may miss the influence of context on word complexity. Future work should gather additional datasets that obtain in-context judgements.

2. Our supplementary analysis focused mainly on whether the group of annotators affected how the various linguistic characteristics affect word complexity, which served to validate the importance of our dataset. However, we do not discuss why or how those relationships work (e.g., why synonyms or hyponyms did not reveal significant differences) as our analysis did not provide insights into these aspects and thus our discussion would involve speculation. Future work based on our dataset can focus on providing further insights into these issues.

3. Our annotators were recruited from a university campus. While we still observed diversity in their literacy skills, as measured by their WRAT-4 scores, future work should expand our dataset by collecting judgements from a broader group of DHH adults with varying levels of education.

4. The main contribution of this paper consists of the release of the word-complexity dataset. However, future work should explore the utility of this dataset for the various stages of lexical simplification (e.g. CWI, or substitution generation and ranking). Furthermore, future work should explore how the use of this dataset to train ATS systems may impact the utility of these systems for DHH adults.

Ethics Statement

Our annotators were hired as research assistants for our study. However, we still followed the traditional considerations expected for an IRB-approved study. In addition, we took all the necessary steps to remove any personal identifiable information to preserve the privacy of our annotators.

Acknowledgements

We thank the anonymous reviewers for their thoughtful comments. We also thank Abraham Glasser and Jinlan Li for their contributions in managing the annotation process, as well as our colleagues who supported us in recruiting annotators.

This material is based upon work supported by the National Science Foundation under award No. 1822747.

References

Eneko Agirre, Carmen Banea, Claire Cardie, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Weiwei Guo, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2014. SemEval-2014 task 10: Multilingual semantic textual similarity. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 81–91, Dublin, Ireland. Association for Computational Linguistics.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Oliver Alonzo, Matthew Seita, Abraham Glasser, and Matt Huenerfauth. 2020. Automatic text simplification tools for deaf and hard of hearing adults: Benefits of lexical simplification and providing users with autonomy. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, CHI '20, page 1–13, New York, NY, USA. Association for Computing Machinery.

Mahmoud Azab, Chris Hokamp, and Rada Mihalcea. 2015. Using word semantics to assist English as a second language learners. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 116–120, Denver, Colorado. Association for Computational Linguistics.

Marc Brysbaert and Boris New. 2009. Moving beyond kučera and francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for american english. *Behavior research methods*, 41(4):977–990.

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: Short Papers - Volume 2*, HLT '11, page 665–669, USA. Association for Computational Linguistics.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Proceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Siobhan Devlin and Gary Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, Assets '06, pages 225–226, New York, NY, USA. ACM.

Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. 2010. Personalized reading support for second-language web documents by collective intelligence. In *Proceedings of the 15th international conference on Intelligent user interfaces*, pages 51–60.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 50–57, Dublin, Ireland. Association for Computational Linguistics.

Sian Gooding and Manuel Tragut. 2022. One size does not fit all: The case for personalised word complexity models.

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural crf model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria. Association for Computational Linguistics.

Poorna Kushalnagar, Scott Smith, Melinda Hopper, Claire Ryan, Micah Rinkevich, and Raja Kushalnagar. 2018. Making cancer health text on the internet easier to read for deaf people who use american sign language. *Journal of Cancer Education*, 33(1):134–140.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3749–3760, Brussels, Belgium. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016a. Collecting and exploring everyday language for predicting psycholinguistic properties of words. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1669–1679, Osaka, Japan. The COLING 2016 Organizing Committee.

Gustavo Paetzold and Lucia Specia. 2016b. Understanding the lexical simplification needs of non-native speakers of English. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 717–727, Osaka, Japan. The COLING 2016 Organizing Committee.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013a. Simplify or help?: Text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, pages 15:1–15:10, New York, NY, USA. ACM.

Luz Rello, Ricardo Baeza-Yates, Laura Dempere-Marco, and Horacio Saggion. 2013b. Frequent words improve readability and short words improve understandability for people with dyslexia. In *IFIP Conference on Human-Computer Interaction*, INTERACT 2013, pages 203–219. Springer.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL - International Journal of Applied Linguistics*, 165(2):259–298.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.