Beyond Vocabulary: Capturing Readability from Children’s Difficulty

Arif Ahmed
Department of Computer Science
Boise State University
arifahmed@u.boisestate.edu

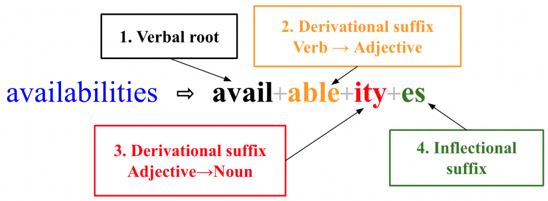
Abstract

Readability formulae targeting children have been developed, but their appropriateness can still be improved, for example by taking into account *suffixation*. Literacy research has identified the suffixation phenomenon makes children’s reading difficult, so we analyze the effectiveness of suffixation within the context of readability. Our analysis finds that suffixation is potentially effective for readability assessment. Moreover, we find that existing readability formulae fail to discern lower grade levels for texts from different existing corpora.

1 Introduction

Readability is employed as a tool for various audiences, including children and second-language users, as well as diverse tasks such as web search, recommendation, selecting textbook materials, calibrating books, text summarization, machine translation, automatic text simplification, and more (Bilal and Huang, 2019; Alharthi and Inkpen, 2019; Stenner, 1996; Paul and Sumita, 2011; Štajner and Saggion, 2013). Importantly, the use of readability for such tasks becomes critical when the target users are children (grades K-6). Unlike adults, they do not (yet) have all the necessary reading skills, so children require more appropriate text according to their grade level (Rahman et al., 2020).

However, Allen et al. (2022) highlighted that the performance of traditional readability formulae greatly varies across different grade levels while estimating the readability of children’s resources. Also, they proposed a lexicon-based formula named *Spache-Allen*, which could capture readability better than other traditional formulae. Generally, lexicon-based readability formulae consider sentence length and static vocabulary to determine text readability (Spache, 1968). Over the years, researchers augmented these static vocabularies (from 1064 to 65,669 words) to increase lexicon-based formula’s performance (Spache, 1968; Madrazo Azpiazu et al., 2018; Allen et al., 2022). While looking up a word within the vocabulary, such formulae do not consider words’ complex properties, such as inflectional endings and derivational suffixes. More recently, Allen et al. (2022) included the Age-of-Acquisition dataset (Kuperman et al., 2012) to the original Spache (1968) vocabulary in their Spache-Allen formula, because they considered children are taught these words over the years. Importantly, children learn these words in a staircased fashion from lower to more complex words across grade levels. Even though vocabulary augmentation has increased their formula’s performance, it does not capture the children’s staircased word learning process. Researchers on literacy identified *suffixation* as an influential factor that affects children’s reading experience (Nagy et al., 1985, 1991). In Figure 1, we show how suffixation makes a word more complex. To the best of our knowledge, no readability research has taken into account the factor of suffixation carefully, which makes children’s reading difficult. Instead of increasing the size of static vocabulary to push the formulas’ performance digits, we should carefully understand children’s vocabulary acquisition process from literacy research for the readability assessment task.


1. Verbal root
2. Derivational suffix Verb → Adjective
availabilities ⇒ avail+able+ity+es
3. Derivational suffix Adjective→Noun
4. Inflectional suffix

Here, derivational suffixes increase the complexity of the word ‘availabilities’, changing both its syntactic category and meaning.

Figure 1: Suffixation in ‘availabilities’

In this paper, we investigate how suffixes indicate the readability level of English text with the research question **RQ:** ***How effective are ranked suffixes from literacy research for readability assessment?*** To answer this research question, we take advantage of prior work of Jarmulowicz (2002), where they identified 43 derivational suffixes and ranked them in 25 discrete levels based on frequency. We posit that these ranks will help us capture the staircased word complexity that children learn over the grade levels. Furthermore, we have made our suffixation approach implementation publicly available on GitHub.[1]

2 Background and Related Work

2.1 Children's Reading Behaviour

As children learn to read and their vocabulary expands, derived words (e.g., inflectional morphology or compound formation) play a substantial role in text comprehension (Jarmulowicz, 2002). In fact, the knowledge of *vocabulary* children already have works as the best predictor for reading comprehension (Stahl and Nagy, 2007). Studies showed that children's knowledge of *morphology* has a significant impact on reading (Anglin et al., 1993; Carlisle, 2000, 2003). Whenever they encounter any unfamiliar morphologically complex words, they use their knowledge of root words and affixes to determine the meaning of that word. Children develop different facets of knowledge of morphology at different rates and times (Tyler and Nagy, 1989). Nagy et al. (1991) found that after the third grade, students gain knowledge of common English suffixes (e.g., '-es' in oxes), and some students face severe problems with understanding the function of suffixes. Children learn inflectional suffixes and compounding before derivational suffixation (e.g., '-able' in readable) (Derwing and Baker, 1979). Later, Nagy et al. (1993) identified one reason for that is the relative abstractness of the information conveyed in derivational suffixes.

2.2 Readability for Children

Over the past century, researchers proposed hundreds of readability assessment methods ranging from classic formulae to featureless models (Flesch, 1948; Madrazo Azpiazu et al., 2018; Filighera et al., 2019; Vajjala and Lučić, 2018; Deutsch et al., 2020; Huebner et al., 2021; Lee et al., 2021; Rao et al., 2021). Still today, traditional formulae from early periods are widely used, which consider word counts, sentence length, lexical, and syntactic features (Flesch, 1948; Dale and Chall, 1948; Flesch, 1950; Gunning et al., 1952). These readability formulae are widely used in real-world environments (Begeny and Greene, 2014; Crossley et al., 2019), as these formulae are easy to deploy. In real-world settings, children are becoming a large user group. So, it is crucial to investigate the appropriateness of the existing readability formula. Article no. 17 of United Nations' Convention on the Rights of the Child also encouraged so.[2] To support children to understand real-world text, we should develop an appropriate readability formula for them.

3 Method

3.1 Data Setup

3.1.1 Corpora

Targeting children (grades K-6), we consider the following datasets.

(a) Common Core State Standards (CCSS): We extract book excerpts from the appendices of the CCSS.[3] Targeting children (grades K-6), we consider 196 books from grades K-8, as texts from grades 6-8 are grouped under the same labeling.

(b) WeeBit: We consider this for web resources (Vajjala and Meurers, 2012). We apply the downsampling technique to the dataset and consider 629 samples from each class. This is a common approach researchers apply to this dataset (Deutsch et al., 2020; Lee et al., 2021).

(c) Science: This corpus has science-related text (i.e., informational text) for K-12 population (Nadeem and Ostendorf, 2018). However, only their publicly available test samples covering grades 3-12 are accessible. To ensure consistent comparison across the three corpora, we select 1035 samples from grades 3-8.

3.1.2 Corpus Analysis

Before we answer our research question, we conduct correlation analysis on the data (Sec. 3.1.1) to identify potential biases. For correlation analysis, we denote two variables– shallow factors (vocabulary size per text, number of words per text, average words per sentence, number of sentences per text) as X (continuous) and grade levels as Y (ordinal). Here, the Y variable is ordinal because each of the

[1] `https://github.com/arif09/arix`

[2] https://www.unicef.org/child-rights-convention/convention-text-childrens-version

[3] http://www.corestandards.org/assets/Appendix_B.pdf

grade levels is a discrete ordinal representing the degree of text complexity. Based on the best practices (Khamis, 2008), we choose Kendall τ as the correlation metric for CCSS, and Spearman's ρ for the WeeBit and Science corpora.

3.2 Suffixation Based Text Complexity

3.2.1 Suffix Ranking for Words

The text simplification research shows that text containing a few complex words or sentences can increase overall text difficulty (Glavaš and Štajner, 2015). In Sec. 2.1, we explain that derivational suffixes make a word more complex than other affixes. To explore this direction, we take advantage of the prior work of Jarmulowicz (2002), which identified 43 derivational suffixes and ranked them from 1 to 25 based on the frequency of a child-directed corpus. We mark this *'derivational suffix rank'* as a complexity indicator that can capture children's cognitive processing effort. In this paper, we represent these 43 ranked derivational suffixes with S_{der}. To the best of our knowledge, this is the first attempt that uses the rank of derivational suffixes as a way to capture a word's complexity.

It is certain that lower grade text (e.g., K-2) may not have any or few derivational suffixes as children start learning the function of suffixes in grade 3 (Nagy et al., 1991). To thoroughly cover a broad spectrum of words, it is appropriate to consider all the derivational and inflectional suffixes (e.g., '-s', '-es') in addition to the 43 S_{der}s. Therefore, we find unique 556 inflectional suffixes and 452 derivational suffixes from UniMorph.[4] Among these 1008 suffixes, it is necessary to assign a rank to these 965 (1008−43) unranked suffixes, $S_{der+inf}$. To achieve this, we follow these three steps:

1. Create character's positional vectors $\vec{S}_{der+inf}$ and $\vec{S}_{der}$ from $S_{der+inf}$ and S_{der} respectively.

2. Derive cosine similarity, $cos(\vec{S}_{der+inf}, \vec{S}_{der})$.

3. For each *candidate* suffix in $S_{der+inf}$, identify the most similar suffix from S_{der} and assign the corresponding rank to the *candidate* suffix.

In Figure 2, we illustrate the process of ranking the suffix '-sion' through the three aforementioned steps. First, we generate positional vectors for the characters of the ranked S_{der} suffixes and the unranked '-sion' suffix. Second, we compute the cosine similarity scores between the vector of the unranked suffix and all other ranked suffix vectors. Third, we determine that the unranked '-sion' suffix shows the highest similarity score of 0.75 with the '-tion' suffix. Since the '-tion' suffix holds a rank value of 1, we assign the same rank value of 1 to the unranked '-sion' suffix.

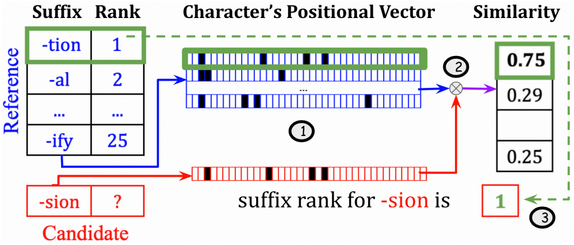


Figure 2: Suffix Ranking Example

3.2.2 Measuring Text Complexity

Although we have assigned rank values (1 to 25) to all the inflectional and derivational suffixes (Sec. 3.2.1), we must put more weight on derivational suffixes. This is because, derivational suffixes are more complex (e.g., changes both syntactic category and meaning of a word) than inflectional suffixes. Considering this fact, we first define word-level complexity. Using these complexity scores, we define text-level complexity.

(a) Word Level: We check a word's derivational suffix by looking it up in UniMorph and verifying if the word is in its derived form. Next, if the derived word, along with its suffix, alters the base word's syntactic category (parts of speech), we categorize that suffix as derivational; otherwise, we classify it as inflectional. We compute C_w, the complexity score of the given word w following the equation:

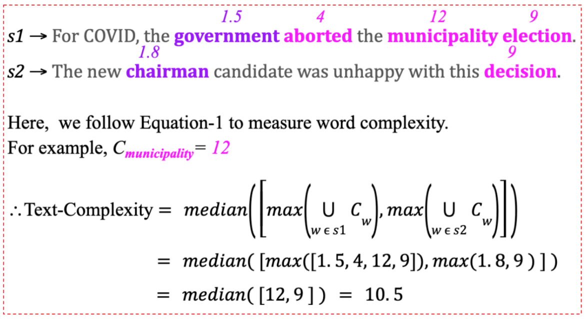
$$C_w = \begin{cases} rank & : w \text{ has derivational suffix} \\ 1 + \frac{rank}{n} & : w \text{ has inflectional suffix} \\ 0 & : w \text{ has no suffix} \end{cases} \qquad (1)$$

Here, **Case 1**: if a word contains a derivational suffix, we directly assign its suffix rank. **Case 2**: in the case of words with an inflectional suffix, we divide the rank by $n = 10$ (randomly chosen) and then add 1. This approach limits the complexity score advancement of inflectional suffixes, thereby emphasizing the contribution of derivational suffixes to the overall complexity score. So, the complexity score of inflectional suffixes would range from 1.1 to 3.5, a considerably lower range compared to the values obtained for derivational suffixes, which span from 1 to 25. **Case 3**: if a word

[4] https://github.com/unimorph/eng/

is either in its base form or bears only prefixes, we consider a 0 (zero) complexity score for it.

(b) Text Level: After measuring the word complexity within a sentence, we take the maximum complexity score from a sentence. Taking the mean value from all the sentences could potentially affect the overall score, so we take the median value from all the sentences within a text.



1.5
4
12
9
s1 → For COVID, the government aborted the municipality election.
1.8
9
s2 → The new chairman candidate was unhappy with this decision.

Here, we follow Equation-1 to measure word complexity.
For example, $C_{municipality} = 12$

$$\therefore \text{Text-Complexity} = median\left(\left[max\left(\bigcup_{w \in s1} C_w\right), max\left(\bigcup_{w \in s2} C_w\right)\right]\right)$$
$$= median(\,[max([1.5, 4, 12, 9]), max(1.8, 9)\,]\,)$$
$$= median(\,[12, 9]\,) \;=\; 10.5$$

Figure 3: Suffixation-Based Text Complexity Scoring

Figure 3 illustrates our process for measuring text complexity using our novel suffixation approach for a sample text comprising two sentences.

3.2.3 Readability Analysis

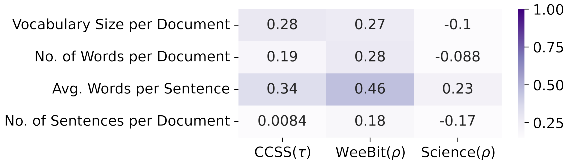
To answer our research question, we first compute text complexity using our suffixation-based approach (Sec 3.2). To see how these scores indicate different reading levels for our selected corpora, we carefully conduct this analysis. Since outliers can significantly influence correlation analysis, it is important to analyze the relationship between actual scores and grade levels visually. Visual inspection can reveal unusual circumstances (e.g., flat or rise to specific grades) that might not be apparent from correlation scores alone. In order to illustrate the effectiveness of our novel suffixation-based approach, we employ a visual technique (i.e., box-and-whisker plot) as opposed to reporting only numeric correlation scores.

Now, we estimate readability levels for Spache-Allen (Allen et al., 2022) and employ the same visual technique to gain insight and compare with our proposed approach.[5] However, Allen et al. (2022) showed the performance of formulae using the Mean Error Rate (MER) metric where the error was computed by taking the absolute difference between actual grade level and predicted grade level. Thus, MER does not indicate if the formula is estimating a grade level above or below the actual grade level. So, we use the raw scores (grades) of Spache-Allen (Allen et al., 2022) for our visual inspection. To gain further insight, we also consider eight other traditional readability formulae and estimate readability levels using TextStat.[6] We are considering nine formulae: Flesch-Kincaid Grade Level (FKGL), Dale-Chall (DC), Gunning Fog Index (FOG), SMOG, Spache Readability Formula (Spache), Spache-Allen (SA), Coleman-Liau Index (COLE), RIX, and LIX (Flesch, 1950; Chall and Dale, 1995; Albright et al., 1996; Mc Laughlin, 1969; Spache, 1968; Allen et al., 2022; Coleman and Liau, 1975; Anderson, 1983).

4 Results and Discussion

4.1 Corpus Analysis

From Figure 4, we find that shallow factors of texts are not highly correlated with the grade levels for all three corpora. This finding confirms that no confounding factors impact our analysis and result.



Here, ρ: Spearman's ρ correlation, τ: Kendall τ correlation.

Figure 4: Correlation of Shallow Factors with Grades

4.2 Readability Analysis

Figure 5(a) shows how suffix-based complexity measurements indicate reading levels of different corpora. We can see a gradual increase in complexity scores across grade levels for all corpora. Specifically, the median values for each boxplot gradually increase from lower grades to upper grades except for the Science corpus. We see more longer boxes and outliers toward the upper grades. We also see that the upper whiskers are much longer than the lower whiskers. These findings indicate that suffixation increases from lower to upper grades. In particular, almost no presence of suffixes in K-1 grade levels and a very low presence of suffixes from grades 2-3, which supports the findings from literacy education research (Nagy et al., 1991). As the Science corpus represents scientific text, it contains more derived words. We find the suffixes increase very slowly across grade levels 3-6.

[5] We follow the author-provided implementation.

[6] https://pypi.org/project/textstat/

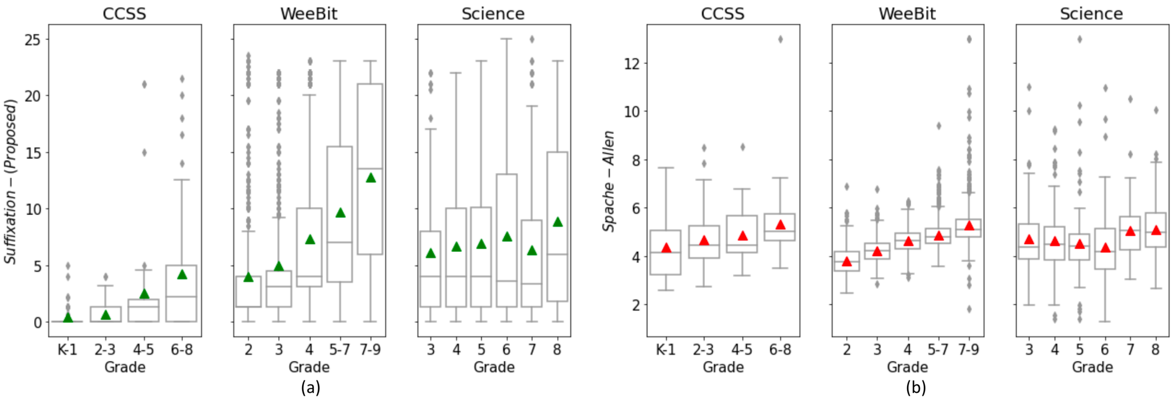


Figure 5: (a) Suffixation-Based Text Complexity Score and (b) Estimated Grades (raw) Using Spache-Allen Formula

On the contrary, Spache-Allen (Allen et al., 2022) readability formula estimated grade levels for all corpora around grade 4 [Figure 5(b)]. In the CCSS corpus, we discover nearly identical median values across grades K-5. It is a concern that this formula estimates higher complexity for texts from grades K-3 in CCSS. In fact, K-2 text contains mostly very simple words (e.g., cat, bat). In a recent study, Bettencourt et al. (2022) utilized Spache-Allen for assessing text complexity in web search results to study children's (grades 1-6) web search engagement. Here, a potential concern arises that using an appropriate readability formula might yield different results in their analysis. Hence, our analysis addresses our research question, confirming that suffixation effectively captures readability for children.

While our focus is on suffixation, we do not delve into the performance of other readability formulae; however, we provide their performance in Appendix A. Our findings indicate that traditional formulae estimate significantly higher grade levels for our chosen corpora, whereas Allen et al. (2022) discovered only an increase of 1-3 grades. For WeeBit, most of the formulae show an upward linear trend but estimated grade levels inaccurately. This observation stems from our corpus analysis, where we address that shallow factors correlate with grade levels of WeeBit corpus better than CCSS and Science corpora.

We could not access NewsELA and Reading A-Z corpora which were merged with WeeBit and CCSS in Allen et al.'s experiments. It is possible that these unavailable datasets contributed to increasing formulas' performance in their conducted experiment. Typically, children's books might not be ideal for automatic readability assessment. For example, easy words are repeated more frequently in lower-grade text. Particularly, educators and teachers increase the amount of text across grade levels, which is a very common confounding factor that can deceive readability assessment. In fact, many complex instruction texts in books are not intended for children. While working on these children's books, we must carefully consider such factors that might affect our experiment.

5 Conclusion

Our investigation shows that findings from literacy research can help us develop the appropriate readability formula for children. We also show the current state-of-the-art readability formula for children fails to discern words with complex morphological properties. Moreover, our work shows that we should consider the findings from other disciplines (e.g., Education, Literacy) to better capture readability to suggest appropriate text for children, a rapidly increasing user group accessing digital platforms. Our word-level complexity scoring can directly support lexical simplification tasks and text-level complexity scoring can enhance text accessibility for diverse user groups (e.g.,second-language learners or marginalized populations). Besides, our novel suffixation approach can serve as a versatile feature for feature-based models across various Natural Language Processing tasks, encompassing various domains such as Information Retrieval or Human-Computer Interaction.

Acknowledgements

We extend our gratitude to Dr. Sowmya Vajjala for generously sharing their WeeBit corpus. We thank the anonymous reviewers for their suggestions and Sharif Ahmed for helpful discussions.

6 Lay Summary

Children in grades K-6 are becoming a large group in online platforms, they use various applications for educational and learning purposes. Specifically, their use of such platforms becomes useful when they can understand the information, mostly text. To serve their purpose, researchers from many disciplines working towards measuring the appropriateness of the text targeting children. To measure the difficulty of any given text, researchers have proposed many methods over the last hundred years. The term 'readability' measures how easy (i.e., text from specific grade levels) any text is for a reader group (i.e., preschool, school, or college).

Most readability research introduced new datasets or increased vocabulary (i.e., word lists) size to show their formula's performance better. Instead of proposing a new readability formula, we try to understand what factors make children's (grades K-6) reading difficult by exploring literacy education research. From that exploration, we identify that 'suffixation' makes children's reading difficult. So, we fit this theory for the readability problem and propose a new approach to compute text difficulty.

Our paper uncovers the effectiveness of 'suffixation' for determining the reading level of any text. Compared to the existing readability formula, it can discern lower-grade text effectively.

References

Judith Albright, Carol de Guzman, Patrick Acebo, Dorothy Paiva, Mary Faulkner, and Janice Swanson. 1996. Readability of patient education materials: implications for clinical practice. *Applied Nursing Research*, 9(3):139–143.

Haifa Alharthi and Diana Inkpen. 2019. Study of linguistic features incorporated in a literary book recommender system. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, pages 1027–1034.

Garrett Allen, Ashlee Milton, Katherine Landau Wright, Jerry Alan Fails, Casey Kennington, and Maria Soledad Pera. 2022. Supercalifragilisticexpialidocious: Why Using the "Right" Readability Formula in Children's Web Search Matters. In *Advances in Information Retrieval*, pages 3–18, Cham. Springer International Publishing.

Jonathan Anderson. 1983. Lix and rix: Variations on a little-known readability index. *Journal of Reading*, 26(6):490–496.

Jeremy M Anglin, George A Miller, and Pamela C Wakefield. 1993. Vocabulary development: A morphological analysis. *Monographs of the society for research in child development*, pages i–186.

John C Begeny and Diana J Greene. 2014. Can readability formulas be used to successfully gauge difficulty of reading materials? *Psychology in the Schools*, 51(2):198–215.

Benjamin Bettencourt, Arif Ahmed, Nic Way, Casey Kennington, Katherine Landau Wright, and Jerry Alan Fails. 2022. Searching for engagement: Child engagement and search engine result pages. In *Interaction Design and Children*, IDC '22, page 479–484, New York, NY, USA. Association for Computing Machinery.

Dania Bilal and Li-Min Huang. 2019. Readability and word complexity of serps snippets and web pages on children's search queries: Google vs bing. *Aslib Journal of Information Management*.

Joanne F Carlisle. 2000. Awareness of the structure and meaning of morphologically complex words: Impact on reading. *Reading and writing*, 12(3):169–190.

Joanne F Carlisle. 2003. Morphology matters in learning to read: A commentary. *Reading Psychology*, 24(3-4):291–322.

Jeanne Sternlicht Chall and Edgar Dale. 1995. *Manual for use of the new Dale-Chall readability formula*. Brookline Books.

Meri Coleman and Ta Lin Liau. 1975. A computer readability formula designed for machine scoring. *Journal of Applied Psychology*, 60(2):283.

Scott A Crossley, Stephen Skalicky, and Mihai Dascalu. 2019. Moving beyond classic readability formulas: New methods and new models. *Journal of Research in Reading*, 42(3-4):541–561.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Bruce L Derwing and William J Baker. 1979. Recent research on the acquisition of english morphology. *Language acquisition*, pages 209–223.

Tovly Deutsch, Masoud Jasbi, and Stuart Shieber. 2020. Linguistic features for readability assessment. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–17, Seattle, WA, USA → Online. Association for Computational Linguistics.

Anna Filighera, Tim Steuer, and Christoph Rensing. 2019. Automatic text difficulty estimation using embeddings and neural networks. In *European Conference on Technology Enhanced Learning*, pages 335–348. Springer.

Rudolf Flesch. 1950. Measuring the level of abstraction. *Journal of Applied Psychology*, 34(6):384.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 63–68, Beijing, China. Association for Computational Linguistics.

Robert Gunning et al. 1952. Technique of clear writing.

Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. 2021. BabyBERTa: Learning more grammar with small-scale child-directed language. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 624–646, Online. Association for Computational Linguistics.

Linda D. Jarmulowicz. 2002. English derivational suffix frequency and children's stress judgments. *Brain and Language*, 81(1):192–204.

Harry Khamis. 2008. Measures of association: How to choose? *Journal of Diagnostic Medical Sonography*, 24(3):155–162.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ion Madrazo Azpiazu, Nevena Dragovic, Oghenemaro Anuyah, and Maria Soledad Pera. 2018. Looking for the movie seven or sven from the movie frozen? a multi-perspective strategy for recommending queries for children. In *Proceedings of the 2018 Conference on Human Information Interaction amp; Retrieval*, CHIIR '18, page 92–101, New York, NY, USA. Association for Computing Machinery.

G Harry Mc Laughlin. 1969. Smog grading-a new readability formula. *Journal of reading*, 12(8):639–646.

Farah Nadeem and Mari Ostendorf. 2018. Estimating linguistic complexity for science texts. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 45–55, New Orleans, Louisiana. Association for Computational Linguistics.

William Nagy, Irene-Anna Diakidoy, and Richard C. Anderson. 1991. The development of knowledge of derivational suffixes.

William E Nagy, Irene-Anna N Diakidoy, and Richard C Anderson. 1993. The acquisition of morphology: Learning the contribution of suffixes to the meanings of derivatives. *Journal of reading Behavior*, 25(2):155–170.

William E. Nagy, Patricia A. Herman, and Richard C. Anderson. 1985. Learning words from context. *Reading Research Quarterly*, 20(2):233–253.

Michael Paul and Eiichiro Sumita. 2011. Translation quality indicators for pivot-based statistical MT. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 811–818, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

Rashedur Rahman, Gwénolé Lecorvé, Aline Étienne, Delphine Battistelli, Nicolas Béchet, and Jonathan Chevelu. 2020. Mama/papa, is this text for me? In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6296–6301, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Simin Rao, Hua Zheng, and Sujian Li. 2021. Cross-lingual leveled reading based on language-invariant features. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2677–2682, Punta Cana, Dominican Republic. Association for Computational Linguistics.

George D Spache. 1968. Good reading for poor readers.

Steven A Stahl and William E Nagy. 2007. *Teaching word meanings*. Routledge.

Sanja Štajner and Horacio Saggion. 2013. Readability indices for automatic evaluation of text simplification systems: A feasibility study for Spanish. In *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, pages 374–382, Nagoya, Japan. Asian Federation of Natural Language Processing.

A Jackson Stenner. 1996. Measuring reading comprehension with the lexile framework.

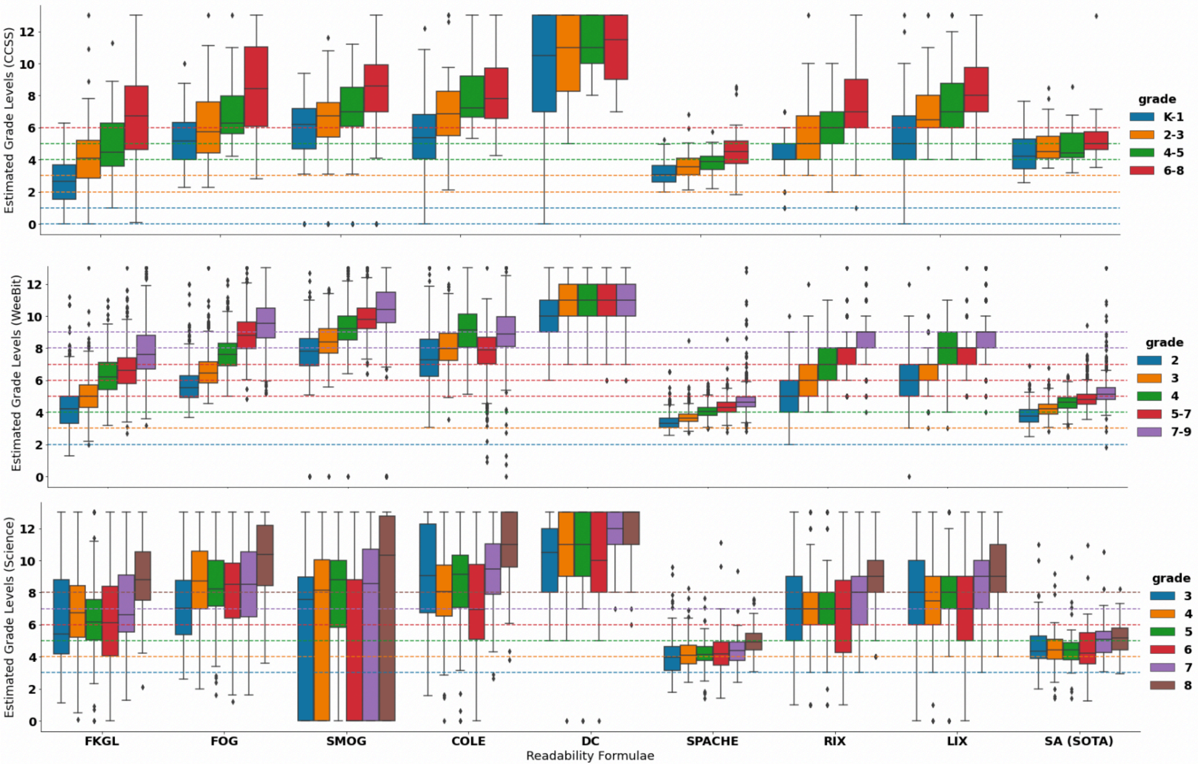
Andrea Tyler and William Nagy. 1989. The acquisition of english derivational morphology. *Journal of memory and language*, 28(6):649–667.

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics.

Sowmya Vajjala and Detmar Meurers. 2012. On Improving the Accuracy of Readability Classification using Insights from Second Language Acquisition. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 163–173, Montréal, Canada. Association for Computational Linguistics.

A Appendix

Because of limitations in scope and page count, we include Figure 6 in this section, illustrating the efficacy of readability formulae on our selected corpora. For better visualization, any estimated grade levels exceeding 13 were adjusted to 13.



Here, each colored dashed horizontal line represents the actual grade level for that corpus, with the boxes indicating the estimated grade levels.

Figure 6: Estimation of Text Readability Using Traditional Readability Formulae