Assessing sentence readability for German language learners with broad linguistic modeling or readability formulas: When do linguistic insights make a difference?

Zarah Weiss
Department of linguistics
University of Tübingen
Germany
zweiss@sfs.uni-tuebingen.de

Detmar Meurers
Department of linguistics
University of Tübingen
Germany
dm@sfs.uni-tuebingen.de

Abstract

The paper presents a new state-of-the-art sentence-wise readability assessment model for German L2 readers. We build a linguistically broadly informed machine learning model and compare its performance against four commonly used readability formulas. To understand when the linguistic insights used to inform our model make a difference for readability assessment and when simple readability formulas suffice, we compare their performance based on two common automatic readability assessment tasks: predictive regression and sentence pair ranking. We find that leveraging linguistic insights yields top performances across tasks, but that for the identification of simplified sentences also readability formulas – which are easier to compute and more accessible – can be sufficiently precise. Linguistically informed modeling, however, is the only viable option for high quality outcomes in fine-grained prediction tasks.

We then explore the sentence-wise readability profile of leveled texts written for language learners at a beginning, intermediate, and advanced level of German. Our findings highlight that a texts' readability is driven by the maximum rather than the overall readability of sentences. This has direct implications for the adaptation of learning materials and showcases the importance of studying readability also below the document level.

1 Introduction

Comprehensible input is key to foster language learning (Swain, 1985), especially when it challenges learners by falling slightly above their individual level of language competence (Vygotsky, 1978; Krashen, 1985). Also in content-matter education, input comprehensibility has been linked to learning success (e.g., O'Reilly and McNamara, 2007). Thus, automatic readability assessment (ARA) is a crucial tool to support education. ARA seeks to align language input with readers' comprehension skills (Vajjala, 2021; Collins-Thompson, 2014). It can not only identify suitable reading materials, but can also ensure learner-input alignment in applications such as tutoring systems or educational conversational agents or as a validation tool for publishers of educational materials. Yet, most work on ARA focuses on English native speakers, leaving much potential for other languages and approaches specifically tailored to the needs of second or foreign language (L2) learners who experience language barriers differently than native speakers (Greenfield, 2004; Collins-Thompson, 2014).

Although most work on ARA has focused on estimating the readability of entire documents, there are many application scenarios in which sentence-level readability assessment is more suitable. Beyond the identification of text simplification targets (Vajjala and Meurers, 2014), they are also more suitable for very short text types including social media language (e.g., tweets and chats), questionnaire or test items used in assessment and empirical education research, or shorter text units in traditional learning materials (e.g., captions or tasks in schoolbooks). Furthermore, there has been little research on the link between sentence and document readability (but see Vajjala and Meurers, 2014) which is immediately relevant for the targeted design and adaptation of educational materials.

There is a startling gap between the methods proposed in ARA research and those used in practice. While for the last two decades, research on ARA has favored machine learning approaches over traditional readability formulas (Vajjala, 2021) due to their generally better performance (e.g., François and Miltsakaki, 2012), simple formulas continue to be used extensively in practice due to their ease of use and low computation demands (Benjamin, 2012). This discrepancy raises the practical question when simple approximations of readability through formulas suffice, and when the use of more

Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022), pages 141 - 153

elaborate systems is necessary.

This paper addresses these issues with four major contributions: First, we present a new state-of-the-art (SOTA) sentence-level readability model for L2 German readers which is based on broad linguistic complexity assessment. Its performance on a 7-point Likert scale is comparable to human raters when it comes to estimating the readability of sentences for German L2 readers. Second, we make this model accessible online to enhance the impact of our work outside academic discourse. Users can extract features from their texts using the publicly available web platform CTAP (Chen and Meurers, 2016; Weiss et al., 2021) and use the results as input for a pre-written R script that applies the model to users' input files in the format that is returned by CTAP.[1] Third, we compare our SOTA machine learning-based approach with commonly used readability formulas for the two common ARA tasks predictive regression and ranking to answer the question when using linguistic insights indeed makes a difference and for which tasks simple readability formulas suffice. Finally, we leverage our SOTA model to explore sentence profiles of leveled L2 articles to provide new insights into the role of sentence readability for document difficulty that can help inform input adaptation strategies for educational materials.

The remainder of this paper is structured as follows: after a brief literature review (Section 2), we introduce the data (Section 3) and linguistic features (Section 4) used for our studies. We then report on the model training and evaluation for predictive regression and sentence ranking (Section 5). Finally, we explore the readability profile of German L2 articles on a document level (Section 6) and discuss our overall findings (Section 7). We conclude with finals remarks on the impact of our findings and an outlook on future work (Section 8).

2 Related work

Early approaches to ARA date back to the last century when traditional readability formulas (e.g., Flesch, 1948; Dale and Chall, 1948) were developed, see DuBay (2004, 2006) for a comprehensive overview. Readability formulas estimate text readability solely based on surface level proxies of text characteristics (e.g., sentence and word length or word frequency). They have been heavily criticized for their lack of linguistic insight and robustness, and have been shown to yield inferior results to statistical approaches to ARA on authentic data (François and Miltsakaki, 2012; Collins-Thompson, 2014; Benjamin, 2012; Vajjala, 2021). Yet, they are still the most widely distributed form of ARA in practice due to their low computational demands, ease of use, and availability for a variety of languages (Benjamin, 2012). Common use cases include work on health literacy (Kiwanuka et al., 2017; Grootens-Wiegers et al., 2015; Esfahani et al., 2016) and as evaluation metrics in computational linguistic work on machine translation (Agrawal and Carpuat, 2019; Marchisio et al., 2019; Stymne et al., 2013) or conversational agents (Langevin et al., 2021; Gnewuch et al., 2018; Santhanam et al., 2020).

Since the early 2000s (cf. Vajjala, 2021), statistical approaches became dominant in research on ARA. This includes feature-based approaches leveraging rich linguistic information for their predictions as well as neural approaches without prior feature engineering. While either method has been shown to yield SOTA performances (e.g., Vajjala and Lučić, 2018; Weiss et al., 2021; Martinc et al., 2021; Bengoetxea et al., 2020) on the OneStopEnglish corpus by Vajjala and Lučić (2018), neural approaches have been argued to be more easily applicable for cross-linguistic readability assessment (Martinc et al., 2021; Madrazo Azpiazu and Pera, 2019), but see Weiss et al. (2021); De Clercq and Hoste (2016). Feature-based approaches, instead, are more appropriate when little data is available or when users need an explanation for the obtained readability score, as is commonly the case in education contexts and for publishers of leveled reading materials who might want to revise their texts after obtaining a readability score (Collins-Thompson, 2014). Established features measure aspects of syntax and lexicon (Collins-Thompson, 2014), morphology (Gonzalez-Dios et al., 2014; Hancke et al., 2012; Weiss et al., 2021), and discourse features. They intersect with common features from automatic writing quality assessment (Crossley, 2020) and Second Language Acquisition research (Vajjala and Meurers, 2012).

Only limited progress has been made on ARA for German, after early work on readability formulas (e.g., Amstad, 1978; Björnsson, 1983; Bamberger and Vanecek, 1984). The now unavailable

[1] Both, the complexity-based model and the R script can be accessed at https://osf.io/jg6kc/?view_only=2d62778d592642a4a210eb4c7cc61f87

DeLite system has been used to predict readability for German municipal texts (Vor der Brück and Hartrumpf, 2007; Vor der Brück et al., 2008a,b). Hancke et al. (2012) and Weiss and Meurers (2018) focused on the binary distinction of texts for adult versus young native speaking readers. However, binary ARA is of limited use in practice. Weiss et al. (2021) present to our knowledge the first and only multi-level classification approach for German documents after introducing the first multi-level readability corpus for German, which is part of a larger multi-lingual readability corpus for language learners. For sentence-wise readability assessment, Naderi et al. (2019a) compiled a German corpus of rated sentences and sentence simplification pairs. Naderi et al. (2019b) used this corpus to train a feature-based regression model yielding a root mean squared error (RMSE) of 0.847 which is to our knowledge the current SOTA on this data.

Little research has investigated the relationship between sentence and document readability, even though there has been some work testing the reliability of readability assessment for very short texts (Collins-Thompson and Callan, 2004) and sentences (Dell'Orletta et al., 2011; Vajjala and Meurers, 2014; Pilán et al., 2014). Vajjala and Meurers (2014) inspect readability differences between sentences from Wikipedia and Simple Wikipedia to investigate the poor performance of document-level ARA models for the identification of sentences from simple and regular texts. They find that sentences from Wikipedia are not systematically more complex than sentences from Simple Wikipedia. This raises several questions for further inquiry. The lack of observable differences might be caused by an insufficient sensitivity of the document-level model for sentence-level readability differences. Also, Simple Wikipedia has criticized as not systematically simpler than Wikipedia (e.g., Štajner et al., 2012; Xu et al., 2015; Yaneva et al., 2016). More research is needed to confirm or refute their finding that harder texts are not simply characterized by containing generally less readable sentences which would have direct implications for work on targeted document adaptation seeking to identify language barriers in educational materials.

3 Data

3.1 TextComplexityDE

The TextComplexityDE corpus (Naderi et al., 2019a) consists of 1,119 sentences. 1,019 sentences were extracted from 23 Wikipedia articles related to history, society, or science and 100 sentences from two articles in *Leichte Sprache* (engl. "simple language"). All were rated by 267 German L2 learners along three separate dimensions defined by Naderi et al. (2019a): readability, understandability, and lexical difficulty. For each dimension, sentences were rated by up to ten learners on a 7-point Likert scale. These ratings were aggregated into a single Mean Opinion Score (MOS). For this article, we focus on sentences' readability score (MOS-R).

	Mean	Std.	Min.	Max.
MOS-R	3.02	1.18	1.00	6.33
Words / sent.	20.08	10.62	4.00	63.00
Syll. / word	2.07	0.35	0.96	4.00

Table 1: Summary statistics for the TextComplexityDE sentences including number of words per sentence (sent.), number of syllables (syll.) per word, and the Mean Opinion Score for readability (MOS-R)

Table 1 contains summary statistics for the number of words per sentence sentence, the number of syllables per word, and MOS-R. It shows that MOS-R not quite uses the full range of the scale and that sentences are on average quite long (around 20 words) whereas words are relatively short (around two syllables). Sentence length has a strong Spearman rank correlation with MOS-R score ($r_s = 0.70; p < 0.01$). Word length only exhibits a weak correlation with MOS-R ($r_s = 0.26; p < 0.01$). The current SOTA performance for a ARA model lies at RMSE $= 0.847$ (Naderi et al., 2019b).

Sentence simplification pairs The corpus contains 250 sentence pairs of sentences with MOS-R > 4 sampled from all 23 Wikipedia articles and their simplifications. The texts were manually simplified by 75 native speakers and contain additional meta information on whether the simplification is only slightly or considerably simpler than the original. One sentence could not be successfully simplified and was excluded by us, resulting in 249 sentence pairs with valid simplifications.

3.2 Spotlight-DE

The Spotlight-DE corpus (Weiss et al., 2021) consists of 1.447 leveled articles by the Spotlight publisher. Articles' topics are connected to German politics, culture, and every-day life. The texts tar-

get L2 learners at a beginning ($N = 763$), medium ($N = 509$), or advanced ($N = 175$) level. The publisher aligns these three levels with the levels A2, B1/B2, and C1 of the Common European Framework of Reference (Council of Europe).

The reading levels in this corpus are assigned at the document level rather than at the sentence level. To obtain sentence-wise estimates, we split each article into individual sentences. Table 2 characterizes the resulting sentence-wise corpus. Compared

	Mean	Std.	Min.	Max.
Easy ($n = 16,694$)				
Words / sent.	11.00	5.09	1.00	73.00
Syll. / word	1.71	0.35	0.50	5.00
Medium ($n = 27,522$)				
Words / sent.	12.50	6.26	1.00	60.00
Syll. / word	1.73	0.35	0.33	6.00
Advanced ($n = 11,952$)				
Words / sent.	13.30	6.99	1.00	63.00
Syll. / word	1.78	0.37	0.50	5.50

Table 2: Summary statistics for the Spotlight-DE sentences across document reading levels (easy, medium, advanced) including number of number words per sentence (sent.), number of syllables (syll.) per word

to the TextComplexityDE corpus, sentences are much shorter. Also, there are no systematic differences in either sentence or work length across reading levels and no meaningful Spearman rank correlation between sentence length and article reading level ($r_s = 0.12; p < 0.001$) or word length and article reading level ($r_s = 0.06; p < 0.001$). Thus, unlike many other learner corpora, the SpotlightDE corpus does not rely on surface level simplifications to differentiate between proficiency levels.

4 Feature extraction and selection

We extracted 543 features of linguistic complexity from the linguistic domains of syntax, lexicon, and morphology as well as psycho-linguistic features of text cohesion, language use, and human language processing and surface level text features inspired by traditional readability formulas. All features have a long standing tradition in ARA research (Collins-Thompson, 2014) or in related work on automatic text scoring (Crossley, 2020) and Second Language Acquisition complexity research (Wolfe-Quintero et al., 1998; Housen et al., 2012).

For feature extraction, we used the CTAP system (Chen and Meurers, 2016, `http://ctapweb.com`) which has been extended to facilitate the analysis of German by Weiss et al. (2021). We chose this system, because it is to our knowledge the most extensive available analysis system for German. The underlying feature extraction engine for German has proven highly successful and robust in a variety of education-related tasks including readability assessment (Weiss and Meurers, 2018; Weiss et al., 2021; Kühberger et al., 2019) and work linked to writing quality assessment (Weiss and Meurers, 2019a,b; Weiss et al., 2019; Bertram et al., 2021; Riemenschneider et al., 2021). Also, using a publicly available web-based system increases the re-usability of any model using these features in practice.

4.1 Feature description

The German pipeline used in CTAP is described in detail in Weiss et al. (2021) and Weiss and Meurers (2021). The latter also contains a comprehensive definition of all complexity measures. We will limit ourselves here to summarize the types of features used to represent the individual linguistic domains.

Syntax The system measures 75 syntactic features which can be further distinguished into measures of clausal elaboration (e.g., *dependent clauses per clause* or *sentence coordination ratio*) and measures of phrasal elaboration (e.g., *prenominal modifiers per noun phrase* or *mean length of prepositional phrases*), as well as measures of syntactic variance (e.g., *edit distances between constituency parses* or *coverage of nominal modifier types*). This set also includes measures of specific grammatical patterns that have been associated with comprehension difficulties for non-native speakers of German (e.g., *the percentage of non-subject prefields* which Ballestracci (2010) identified as language barriers for Italian learners of German) and raw counts of syntactic patterns, such as the number of dependent clauses.

Lexicon There are 146 features of lexical complexity which can be further divided into measures of lexical richness (e.g., *MTLD* by McCarthy (2005) as well as different mathematical transformations of the type-token ratio), measures of lexical variation (e.g., *verb variation*), and lexical density (e.g., *noun type-token ratio* and other parts-of-speech specific type-token ratios). This group also

contains also features measuring the overall occurrence of different parts-of-speech such as nouns, verbs, or puncuation marks.

Morphology CTAP measures 64 measures of morphological complexity for German. We extract features of nominal and verbal inflection (e.g., *genitive case per noun*), derivation (e.g., *derived nouns per noun*), and compounding (e.g., *average compound depth*). We also measure the variability of morphological exponents using different parametrizations of the Morphological Complexity Index (MCI; Brezina and Pallotti, 2019).

Cohesion We extract 46 measures of text cohesion and discourse for German. The features used here include explicit measures of cohesion (e.g., *causal connectives per sentence*) as well as implicit measures of cohesion linked to the use of pronouns and repetitions of subjects, objects, or nouns.

Language use The system offers 172 lexical language use features based on external German data bases. CTAP calculates average word frequencies and their standard deviations with and without log transformations and binned in log frequency bands for four frequency data bases that represent different types of language use: frequencies based on the Subtlex-DE data base consisting of movie and TV captions and Google Books 2000 (both Brysbaert et al., 2011), dlexDB frequencies (Heister et al., 2011) based on German newspaper articles, and frequencies and age of active use measures extracted from the Karlsruhe Children's Text corpus (Lavalley et al., 2015) consisting of essays written by German children in first to eighth grade.

Human sentence processing There are 21 measures of human processing that can be calculated for German. Weiss and Meurers (2018) and Weiss et al. (2021) have used features based on the Dependency Locality Theory (DLT; Gibson, 2000) for German readability classification using different weight configurations by Shain et al. (2016).

Surface length We extract 18 surface length features for German that solely rely on the identification of sentences, words, letters, and syllables. These features include the raw number of these constructs as well as means and standard deviations for sentence and word length based on these units, e.g., *mean sentence length in syllables*.

4.2 Feature selection

After extracting these features from the TextComplexityDE corpus, we removed all features with near-zero variance, i.e., all features for which at least 80% of the data exhibit the same value. This is the case for 31.3% of features ($N = 170$) due to near-exclusively zero values (i.e., not occurring in most data). This leaves 373 features for the analysis coming from all feature domains which were used for model training in Study 1 (Section 5).

This considerable reduction in the number of features is to be expected for data that is as short as the sentences in the TextComplexityDE corpus (e.g, Weiss and Meurers (2021) also report a reduction of 50% of complexity features for short texts). For example, only 7 of the 46 cohesion measures are sufficiently variable on this data, because most cohesion measures are calculated across sentence boundaries. Similarly, only 19 of 64 measures of morphological complexity are sufficiently variable, because there is not enough language material to produce a variety of inflectional properties. Conversely, nearly all language use and lexical features as well as most features of phrasal elaboration remain included in the reduced feature set.

5 Sentence-wise readability assessment

5.1 Set-up

We trained and compared several machine learning algorithms[2] using 10-folds cross-validation (10 CV) and the z-transformations of the 373 features selected in Section 4.2. We selected these algorithms based on their use in previous research or their robustness against large feature sets with multi-colinearity. The Bayesian Ridge Regression outperformed the other models and will be discussed in more detail in the following. To evaluate this complexity-based model's (henceforth: CBM) overall performance, we calculated its RMSE and Spearman rank correlation (r_s) during 10 CV (Section 5.2) and compared it against the current SOTA performance on the data (RMSE $= 0.847$, Naderi et al., 2019b). We also used the model to rank the pairs of regular and simplified sentences in TextComplexityDE (Section 5.3). We report the ranking accuracy in terms of the percentage of correctly ranked pairs for all i) pairs irrespec-

[2]Multiple linear regression with backward feature selection, linear support vector machine regression, random forests, Bayesian ridge regression (model averaged), Bayesian generalized linear model, quantile regression with LASSO penalty

tive of their degree of simplification ($N = 249$), ii) weakly simplified pairs ($N = 114$), and iii) strongly simplified pairs ($N = 135$).

In both evaluation steps, we compared the CBM's performance against five alternative models. We trained a Bayesian Ridge Regression model using only surface length measures as predictors as a baseline (henceforth: length-based model or LBM). We additionally use the following widely used readability formulas for both tasks:[3]

- the *Amstad Readability Index* (ARI; Amstad, 1978) which adapts the Flesch Reading Ease (Flesch, 1948) to German native speakers;
- the *Erste Wiener Sachtextformel* (WSF; Bamberger and Vanecek, 1984) designed for expository texts for German native speakers;
- The *LIX readability index* (Björnsson, 1983) which has been designed to align texts with adult native speakers' reading skills across a variety of languages including German; and
- the *Miyazaki EFL Readability Index* (MER; Greenfield, 1999, 2004) which was designed for English L2 readers.[4]

We calculated all formulas using a publicly available python-based readability calculator which we adjusted to use stanza (Qi et al., 2020) instead of NLTK (Bird and Loper, 2004) for segmentation.[5]

5.2 Results for regression with 10 CV

Table 3 shows the RMSE and Spearman rank correlation of the estimates with MOS-R in the TextComplexityDE data. Both, LBM and CBM outperform the current SOTA on the TextComplexityDE data ($RMSE = 0.847$; Naderi et al., 2019b). Our linguistically more informed CBM clearly outperforms the LBM in terms of both, RMSE and correlation. Due to the differences in the predicted scales, we cannot compute the RMSE for the four readability formulas, but the correlation shows that both, the CBM and LBM outperform the formulas.

	CBM	LBM	WSF	LIX	ARI	MER
RMSE	.685	.739	*n.a.*	*n.a.*	*n.a.*	*n.a.*
r_s	.806	.785	.681	.679	-.532	-.666

Table 3: RMSE and Spearman rank correlation between MOS-R and the predictions by CBM, LBM, and the readability formulas.

The correlation of ARI with MOS-R is much lower than for the other formulas. This is unexpected, because all formulas use only sentence and word length features. However, ARI assigns a much larger weight to word length than the other formulas which in turn correlates only weakly with MOS-R in TextComplexity-DE (see Section 3.1).

CBM's prediction error lies at RMSE $= 0.685$ points on the Likert scale. This is comparable to the variance between raters in the TextComplexityDE data. Averaged across all rated sentences the across-rater standard deviation for MOS-R is at $1.03 \pm 0.51; IQR = [0.71; 1.41]$. This shows that the error of our CBM lies even below the acceptable range of disagreement exhibited by human raters.

	CBM	LBM	WSF	LIX	ARI	MER
Acc.	96.0	93.0	93.6	93.6	95.6	96.8
$-$	95.6	92.1	91.1	91.1	95.6	96.5
$+$	96.5	94.1	96.5	96.5	95.6	97.0

Table 4: Overall ranking accuracy (Acc.), ranking accuracy for weakly simplified pairs ($-$), and ranking accuracy for strongly simplified pairs ($+$)

5.3 Results for ranking of sentence pairs

Table 4 shows the results of the sentence ranking experiment. The ranking accuracy for all ARA models lies above 90%. With an overall accuracy of 96%, CBM again outperforms LBM and the readability formulas WSF and LIX. However, ARI and MER perform comparably to CBM despite their weak performance on the previous regression experiment. It seems that word length (which is weighted higher for these two formulas than for the rest) is more informative than sentence length for distinguishing simplified and regular sentences.

To also estimate if the models reflect the degrees of simplification in the data (weak vs. strong), we compare the difference in the predicted readability score between each sentence and its simplified counterpart. The difference should be systematically larger for strongly than for weakly simplified sentences. We test this assumption using significance testing[6] ($\alpha < 0.05$) and by estimating

[3] All formula equations are defined in Appendix A.

[4] We added this formula to include an estimate tailored to L2 readers despite the lack of German L2 readability formulas.

[5] https://github.com/zweiss/RC_Readability_Calculator

[6] We used a two-sided t-test or Wilcoxon Rank Sum and Signed Rank Tests depending on the normality of predictions determined with a Shapiro-Wilk Normality Test ($\alpha < .05$).

the effect size with Cohen's d.[7] We see a significant, small effect for CBM ($p = 0.02; d = 0.31$), LBM ($p = 0.04; d = 0.25$), MER ($p < 0.01; d = -0.36$), ARI ($p < 0.01; d = -0.30$), LIX ($p = 0.02; d = 0.35$), and WSF ($p = 0.01; d = 0.35$), see Appendix B for a visualization of the findings.

6 Exploring text profiles in leveled articles

6.1 Set-up

We used CBM to explore the text profiles of easy, medium, and advanced articles in the Spotlight-DE corpus, because it was the most precise model in Study 1. With CTAP, we extracted the 373 features from the sentence-split Spotlight-DE data that are used by the model and calculated their z-scores. We inspected the distribution of sentence readability scores across article levels from several perspectives. We first compared the overall differences in sentence complexity per article level and the differences in maximum sentence complexity using significance testing, effect size estimation (parallel to Study 1) and data visualization. We then evaluated the proportions of sentences within a 0.5 point sentence readability interval across article levels. Finally, we visualized the sentence readability of the first ten sentences in a sample of Spotlight-DE articles in three heatmaps, one for each article levels annotated in the Spotlight-DE corpus. This way, we obtain a non-aggregated estimate of the text profiles. To keep the heatmaps comparable, we used all 175 advanced articles as well as a random sample of 175 easy and 175 medium articles containing at least ten sentences.

6.2 Results

Figure 1 combines different perspectives on the sentence-wise article profiles split by article level. We see that the prediction ranges from 1 to 5, a reasonable coverage of the empirically observed MOS-R scale ($1 - 6.33$) in the TextComplexityDE data given the corpus characteristics discussed in Section 3. Figure 1a summarizes the overall sentence readability grouped by article levels with notches indicating the 95% confidence interval. There are small significant differences between easy and medium ($p < 0.001; d = -0.259$) and easy and advanced ($p < 0.001; d = -0.435$) articles, but only negligible albeit significant differences medium and advanced ($p < 0.001; d = -0.178$) articles. The boxplot shows considerable overlap for the 50% range of the data even between easy and advanced sentences. In Figure 1b, which considers only articles' maximum sentence readability scores, this overlap is considerably reduced. Here, we observe large significant differences between easy and advanced ($p < 0.001; d = -2.05$) and medium and advanced ($p < 0.001; d = -1.24$) articles, and moderate significant differences medium and advanced ($p < 0.001; d = -0.689$) articles. This indicates that the maximum sentence readability is more indicative for overall readability level of a text than considering the readability of all its sentences.

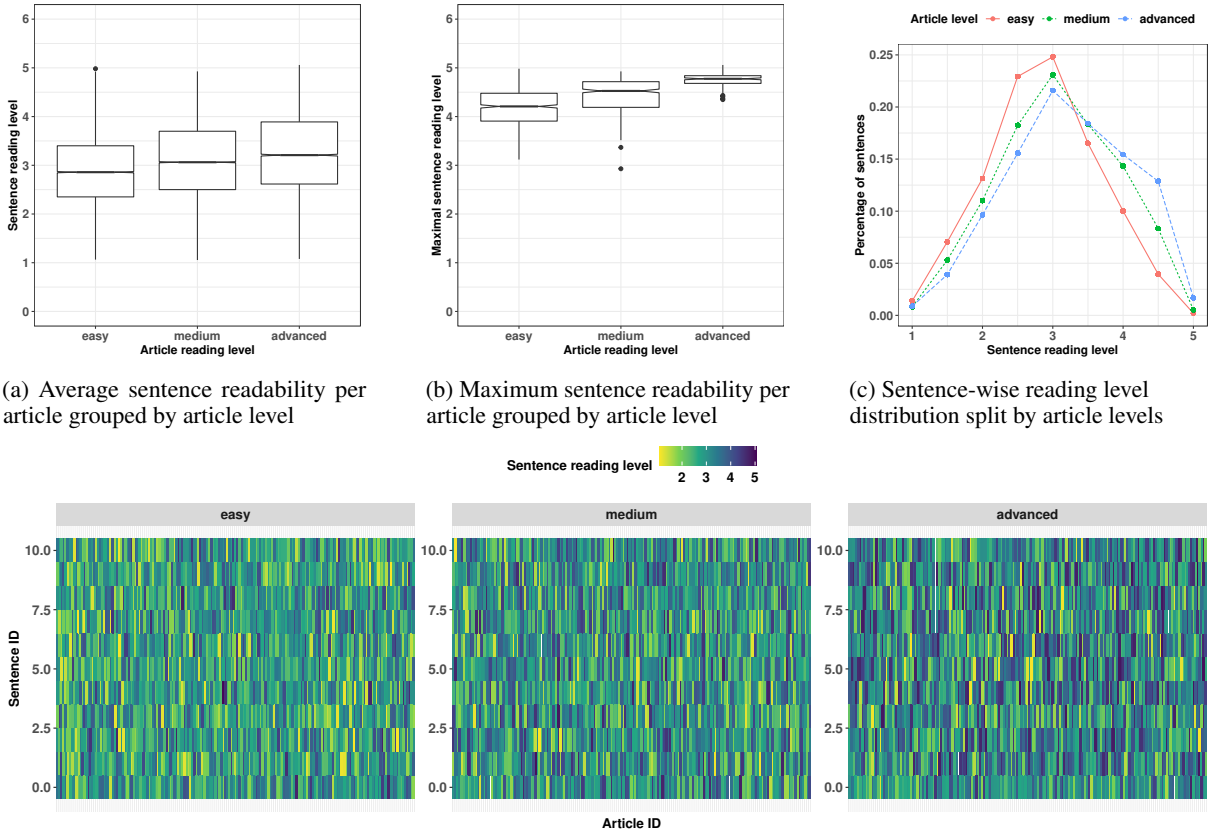
Figure 1c confirms this by comparing the percentage of sentences falling within a 0.5 point readability range across article levels. Sentences from articles at all levels are predominantly medium difficult (MOS-R$= 3$) and between 55.6% (advanced) to 64% (easy) of sentences fall in the range from $2.5 \leq$ MOS-R ≤ 3.5. Article levels differ mostly in the tails of the distribution. The difference is most pronounced for higher difficulty levels (MOS-R ≥ 4): 30% of sentences from advanced articles fall into this range, but only 23.1% of sentences from medium and 14.1% of sentences from easy articles. Even so, it is worth noting that the percentage of sentences with MOS-R ≤ 3 is systematically highest for easy articles and higher for medium than advanced articles. Inversely, the percentage of sentences with MOS-R > 3 is highest for advanced articles and higher for medium than easy articles.

Figure 1d visualizes the sentence readability scores of the first ten sentences of 175 articles per article level. The heatmap depicts the first ten sentences of each sampled article rather than summarizing across sentences and articles at the same article level to demonstrate the relative homogeneity of sentence reading scores for articles at the same article level and the systematic increase in the proportion of more demanding sentences across individual articles with higher article levels.

7 Discussion

Study 1 investigated the performance of linguistically informed readability models and readability formulas for sentence-wise readability assessment for two common ARA tasks: precise predictive regression (Section 5.2) and ranking to identify simplified sentences in sentence simplification pairs (Section 5.3). The results showcase the ver-

[7] We tested for unequal variance using an F test ($\alpha < .05$). In case of unequal variance, we used a Welch approximation for unequal variances to calculate Cohen's d.



(a) Average sentence readability per article grouped by article level

(b) Maximum sentence readability per article grouped by article level

(c) Sentence-wise reading level distribution split by article levels

(d) Predicted sentence readability for the first ten sentences of 175 randomly sampled easy, medium, and advanced articles. Each sentence is represented by a cell. Its readability is encoded with the cell color. The cell's position on the x-axis encodes the article it belongs to and its position on the y-axis its position in that article, e.g., the third sentence in each article is located at $y = 3$.

Figure 1: Sentence readability profiles predicted by our complexity-based model on the Spotlight-DE corpus grouped by article levels (easy, medium, advanced) to showcase differences in sentence readability across documents at different difficulty levels.

satile performance of linguistically informed readability models: only our complexity-based model achieved top performance for both tasks. For the more difficult and authentic task of precise predictive regression, we showed that our linguistically informed complexity-based model clearly outperforms simplistic formulas and set a new SOTA performance (RMSE=0.685) on the data set. The better performance cannot be exclusively attributed to the statistically stronger method, because on both tasks, the complexity-based model clearly outperformed the length-based model. This shows that broad linguistic modeling adds valuable insights beyond the powerful statistical training method.

For ranking, all ARA models achieved an accuracy well above 90% and two readability formulas performed at par with our complexity-based model. This shows that even simple ARA approaches can successfully distinguish relative differences in readability between content-wise equivalent sentences that are being introduced by text simplification. Despite being a rather artificial task, this has some limited applications, e.g., when evaluating machine translation and text simplification systems.

In Study 2, we used our complexity-based model to inspect the sentence-wise readability profiles of leveled texts for L2 readers. Our findings clearly show that while there is a tendency for easier texts to contain more sentence with lower difficulty scores, also medium and advanced texts contain mostly accessible sentences. It is really the presence of difficult sentences within documents that dictates an articles' overall readability. This has clear implications for the design and simplification of educational materials: to efficiently adjust the overall readability level of a text, we need to identify specific sentences that form language barriers rather than simplifying the entire text.

8 Conclusion

We have presented a new SOTA sentence-wise ARA model for German L2 readers which is pub-

licly available and accessible for users with minimal background in R. Leveraging broad linguistic insights, it predicts readability with a margin of error even below the acceptable disagreement range for humans raters. We showed that to flag simplified sentences also traditional readability formulas suffice, but that broad linguistic modeling is needed to obtain the precise predictive readability estimates that are often required in practice (e.g., to adapting learning and teaching materials).

We further explored leveled articles for German L2 readers to illustrate the practical benefits of sentence-level ARA and gain insights into text profiles of leveled documents. Our findings highlight that the readability of texts is driven by the maximum rather than the overall readability of sentences. This has direct implications for the adaptation of teaching materials, which should focus on identifying specific sentences posing language barriers rather than the simplification of all or any sentence in a text. To our knowledge, this is the first time detailed analysis of sentence profiles of leveled reading materials for German. Future work should further explore the implications of this for text simplification, for example using eye-tracking studies. Our work lays the foundation for further research on ARA for German and opens up numerous opportunities for educational applications, such as ARA for captions and task descriptions in school books or the analysis of social media and chat conversations with L2 learners.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1549–1564.

T. Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, University of Zurich.

Sabrina Ballestracci. 2010. Der erwerb von verbzweitsätzen mit subjekt im mittelfeld bei italophonen daf-studierenden. erwerbsphasen, lernschwierigkeiten und didaktische implikationen. *Linguistik online*, 41(1).

Richard Bamberger and Erich Vanecek. 1984. *Lesen – Verstehen – Lernen – Schreiben. Die Schwierigkeitsstufen von Texten deutscher Sprache*. Jugend und Volk, Vienna.

Kepa Bengoetxea, Itziar González-Dios, and Amaia Aguirregoitia. 2020. AzterTest: Open source linguistic and stylistic analysis tool. *Procesamiento del Lenguaje Natural*, 64:61–68.

Rebekah George Benjamin. 2012. Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educational Psychology Review*, 24:63–88.

Christiane Bertram, Zarah Weiss, Lisa Zachrich, and Ramon Ziai. 2021. Artificial intelligence in history education. linguistic content and complexity analyses of student writings in the cahist project (computational assessment of historical thinking). *Computers and Education: Artificial Intelligence*, page 100038.

Steven Bird and Edward Loper. 2004. Nltk: The natural language toolkit. In *Proceedings of the ACL demonstration session*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Carl-Hugo Björnsson. 1983. Readability of newspapers in 11 languages. *Reading Research Quarterly*, pages 480–497.

Vaclav Brezina and Gabriele Pallotti. 2019. Morphological complexity in written L2 texts. *Second language research*, 35(1):99–119.

Marc Brysbaert, Matthias Buchmeier, Markus Conrad, Arthur M. Jacobs, Jens Bölte, and Andrea Böhl. 2011. The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, 58:412–424.

Xiaobin Chen and Detmar Meurers. 2016. CTAP: A web-based tool supporting automatic complexity analysis. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity*, pages 113–119, Osaka, Japan. COLING.

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.

Kevyn Collins-Thompson and Jamie Callan. 2004. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL 2004*, Boston, USA.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

Scott A. Crossley. 2020. Linguistic features in writing quality and development: An overview. *Journal of Writing Research*, 11(3):415–443.

Edgar Dale and Jeanne S. Chall. 1948. A formula for predicting readability. *Educational research bulletin; organ of the College of Education*, 27(1):11–28.

Orphée De Clercq and Véronique Hoste. 2016. All mixed up? Finding the optimal feature set for general readability prediction and its application to English and Dutch. *Computational Linguistics*, 42(3):457–490.

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.

William H. DuBay. 2004. *The Principles of Readability*. Impact Information, Costa Mesa, California.

William H. DuBay. 2006. *The Classic Readability Studies*. Impact Information, Costa Mesa, California.

B. Janghorban Esfahani, A. Faron, K. S. Roth, P. P. Grimminger, and J. C. Luers. 2016. Systematic readability analysis of medical texts on websites of German university clinics for general and abdominal surgery. *Zentralblatt fur Chirurgie*, 141(6):639–644.

Rudolf Franz Flesch. 1948. A new readability yardstick. *Journal of Applied Psychology*, 32(3):221–233.

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

Edward Gibson. 2000. The dependency locality theory: A distance-based theory of linguistic complexity. In Alec Marantz, Yasushi Miyashita, and Wayne O'Neil, editors, *Image, language, brain: papers from the First Mind Articulation Project Symposium*, pages 95–126. MIT.

Ulrich Gnewuch, Stefan Morana, Carl Heckmann, and Alexander Maedche. 2018. Designing conversational agents for energy feedback. In *International Conference on Design Science Research in Information Systems and Technology*, pages 18–33. Springer.

Itziar Gonzalez-Dios, María Jesús Aranzabe, and Arantza Díaz de Ilarraza. 2014. Making biographical data in wikipedia readable: A pattern-based multilingual approach. In *Proceedings of the Workshop on Automatic Text Simplification - Methods and Applications in the Multilingual Society (ATS-MA 2014)*, pages 11–20, Dublin, Ireland. Association for Computational Linguistics and Dublin City University.

Jerry Greenfield. 1999. *Classic readability formulas in an EFL context: Are they valid for Japanese speakers?* Ph.D. thesis, Temple Univesity.

Jerry Greenfield. 2004. Readability formulas for efl. *JALT Journal*, 26(1):5–24.

Petronella Grootens-Wiegers, Martine C. De Vries, Tessa E. Vossen, and Jos M. Van den Broek. 2015. Readability and visuals in medical research information forms for children and adolescents. *Science Communication*, 37(1):89–117.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 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, Mumbai, India. http://aclweb.org/anthology-new/C/C12/C12-1065.pdf.

Julian Heister, Kay-Michael Würzner, Johannes Bubenzer, Edmund Pohl, Thomas Hanneforth, Alexander Geyken, and Reinhold Kliegl. 2011. dlexDB - eine lexikalische Datenbank für die psychologische und linguistische Forschung. *Psychologische Rundschau*, 62:10–20.

Alexis Housen, Folkert Kuiken, and Ineke Vedder. 2012. Complexity, accuracy and fluency: Definitions, measurement and research. In Alex Housen, Folkert Kuiken, and Ineke Vedder, editors, *Dimensions of L2 Performance and Proficiency*, Language Learning & Language Teaching, pages 1–20. John Benjamins.

Elizabeth Kiwanuka, Raman Mehrzad, Adnan Prsic, and Daniel Kwan. 2017. Online patient resources for gender affirmation surgery: an analysis of readability. *Annals of Plastic Surgery*, 79(4):329–333.

Stephen D Krashen. 1985. *The input hypothesis: Issues and implications*. Longman, New York.

Christoph Kühberger, Christoph Bramann, Zarah Weiss, and Detmar Meurers. 2019. Task complexity in history textbooks. a multidisciplinary case study on triangulation in history education research. *History Education International Research Journal (HEIRJ)*, 16(1). Special Issue on Mixed Methods and Triangulation in History Education Research.

Raina Langevin, Ross J Lordon, Thi Avrahami, Benjamin R Cowan, Tad Hirsch, and Gary Hsieh. 2021. Heuristic evaluation of conversational agents. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

Rémi Lavalley, Kay Berkling, and Sebastian Stüker. 2015. Preparing children's writing database for automated processing. In *LTLT@ SLaTE*, pages 9–15.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Kelly Marchisio, Jialiang Guo, Cheng-I Lai, and Philipp Koehn. 2019. Controlling the reading level of machine translation output. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 193–203.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Philip M. McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, University of Memphis.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019a. Subjective assessment of text complexity: A dataset for german language. *arXiv preprint arXiv:1904.07733*.

Babak Naderi, Salar Mohtaj, Karan Karan, and Sebastian Möller. 2019b. Automated text readability assessment for german language: a quality of experience approach. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE.

Tenaha O'Reilly and Danielle S McNamara. 2007. The impact of science knowledge, reading skill, and reading strategy knowledge on more traditional "high-stakes" measures of high school students' science achievement. *American educational research journal*, 44(1):161–196.

Ildikó Pilán, Elena Volodina, and Richard Johansson. 2014. Rule-based and machine learning approaches for second language sentence-level readability. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 174–184, Baltimore, Maryland, USA. ACL.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Anja Riemenschneider, Zarah Weiss, Pauline Schröter, and Detmar Meurers. 2021. Linguistic complexity in teachers' assessment of german essays in high stakes testing. *Assessing Writing*, 50:100561.

Sashank Santhanam, Alireza Karduni, and Samira Shaikh. 2020. Studying the effects of cognitive biases in evaluation of conversational agents. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Cory Shain, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler. 2016. Memory access during incremental sentence processing causes reading time latency. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 49–58, Osaka.

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

Sara Stymne, Jörg Tiedemann, Christian Hardmeier, and Joakim Nivre. 2013. Statistical machine translation with readability constraints. In *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, pages 375–386.

Merrill Swain. 1985. Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In Susan M. Gass and Carolyn G. Madden, editors, *Input in second language acquisition*, pages 235–253. Newbury House, Rowley, MA.

Sowmya Vajjala. 2021. Trends, limitations and open challenges in automatic readability assessment research. *arXiv preprint arXiv:2105.00973*.

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.

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. http://aclweb.org/anthology/W12-2019.pdf.

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.

Tim Vor der Brück and Sven Hartrumpf. 2007. A semantically oriented readability checker for German. In *Proceedings of the 3rd Language & Technology Conference*, pages 270–274, Poznań, Poland. Wydawnictwo Poznańskie.

Tim Vor der Brück, Sven Hartrumpf, and Hermann Helbig. 2008a. A readability checker with supervised learning using deep syntactic and semantic indicators. *Informatica*, 32(4):429–435.

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

Lev S. Vygotsky. 1978. *Mind in society: The development of higher psychological processes*. Harvard University Press, Cambridge, MA.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the Joint 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54.

Zarah Weiss and Detmar Meurers. 2018. Modeling the readability of German targeting adults and children: An empirically broad analysis and its cross-corpus validation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, Santa Fe, New Mexico, USA. https://www.aclweb.org/anthology/C18-1026.

Zarah Weiss and Detmar Meurers. 2019a. Analyzing linguistic complexity and accuracy in academic language development of German across elementary and secondary school. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy. Association for Computational Linguistics.

Zarah Weiss and Detmar Meurers. 2019b. Broad linguistic modeling is beneficial for German L2 proficiency assessment. In *Widening the Scope of Learner Corpus Research. Selected Papers from the Fourth Learner Corpus Research Conference*, Louvain-La-Neuve. Presses Universitaires de Louvain.

Zarah Weiss and Detmar Meurers. 2021. Analyzing the linguistic complexity of German learner language in a reading comprehension task: Using proficiency classification to investigate short answer data, cross-data generalizability, and the impact of linguistic analysis quality. *International Journal of Learner Corpus Research*, 7(1):84–131.

Zarah Weiss, Anja Riemenschneider, Pauline Schröter, and Detmar Meurers. 2019. Computationally modeling the impact of task-appropriate language complexity and accuracy on human grading of German essays. In *Proceedings of the 14th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, Florence, Italy.

Kate Wolfe-Quintero, Shunji Inagaki, and Hae-Young Kim. 1998. *Second Language Development in Writing: Measures of Fluency, Accuracy & Complexity*. Second Language Teaching & Curriculum Center, University of Hawaii at Manoa, Honolulu.

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.

Victoria Yaneva, Irina P. Temnikova, and Ruslan Mitkov. 2016. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the 10h International Conference on Language Resources and Evaluation (LREC)*, pages 293–299.

A Definition of readability formulas

Equation 1 shows the general form of all four readability formulas consisting of an intercept (β_0), a weighted sentence length estimate ($\beta_1 \times SL$), and a weighted word length estimate ($\beta_2 \times WL$).

$$y = \beta_0 + \beta_1 \times SL + \beta_2 \times WL \quad (1)$$

Table 5 shows the respective weights ($\beta_0, \beta_1, \beta_2$) and measurement units for sentence length (SL) and word length (WL). Equation 2 specifies the definition of the composite score for word length used in the *Erste Wiener Sachtextformel*.

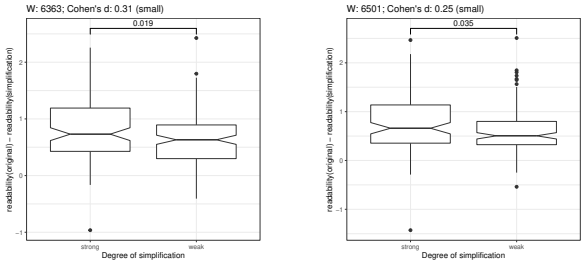
y	β_0	β_1	β_2	SL	WL
LIX	0.0	1.0	1.0	words	syll.
ARI	180.0	−1.0	−58.6	words	syll.
MER	164.9	−1.9	−18.8	words	char.
WSF	0.0	0.2	1.0	words	Eq. 2

Table 5: Weights and measurement units across readability formulas (syll. = syllables, char. = characters)

$$WL_{WSF} = 0.19 \times 3SW + 0.13 \times 6CW - 0.03 \times 1SW - 0.88, \quad (2)$$

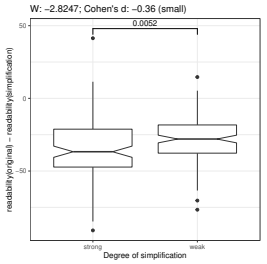
with $3SW$ being the percentage of three or more syllable words, $6CW$ being the percentage of six or more character words, and $1SW$ being the percentage of monosyllabic words. All weights in Table 5 and Equation 1 have been rounded to one decimal point for simplicity.

B Prediction differences between different degrees of simplification

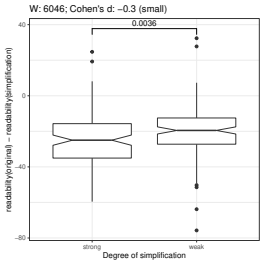


(a) CBM

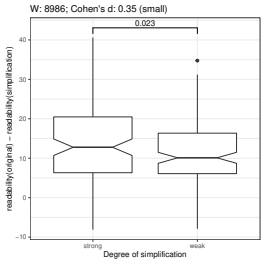
(b) LBM



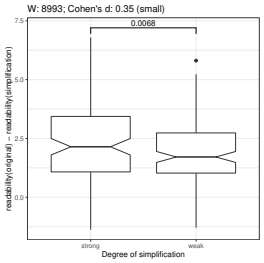
(c) MER



(d) ARI



(e) LIX



(f) EWS

Figure 2: Predicted readability difference between regular and simplified sentences by degree of simplification