Readability Measures in Automatic Text Simplification: Is Simplification Quality a Coherent Construct?

Rémi Cardon[♦*], **A. Seza Doğruöz**[♥]
♦Computer Science and Engineering Department, UC3M, Spain
♥LT3, IDLab, Universiteit Gent, Belgium
rcardon@inf.uc3m.es, as.dogruoz@ugent.be

Abstract

Readability is a central concept in automatic text simplification (ATS), yet the two fields have largely developed in parallel, with limited cross-fertilization. While prior work has studied correlations between automatic evaluation metrics and human judgment in ATS, the correlations between these two aspects and readability measures have not received systematic attention. We address this gap by investigating to what extent readability measures align with both human judgment and automatic metrics in ATS. Using two English datasets annotated with human judgments (SimplicityDA at the sentence level and D-Wikipedia at the document level), we compute 1,066 linguistic features (covering lexical diversity, lexical sophistication, syntactic sophistication, and cohesion) and eight traditional readability formulas, and correlate them against human scores and standard ATS metrics (BLEU, SARI, BERTScore, LENS, D-SARI). Our results show that readability measures correlate poorly with both human judgment and automatic metrics across both levels. The meaning preservation criterion consistently yields the highest correlation values, while simplicity and fluency criteria remain low. We also find systematic differences between sentence-level and document-level simplification in terms of which features are most informative: type-token ratio features are predictive at the sentence level but not at the document level, while corpus-frequency features show the opposite pattern. These findings point to a broader issue: ATS lacks a shared theoretical construct for simplification quality, and the three main approaches to its assessment (human judgment, readability measures, and automatic metrics) do not consistently converge.

Keywords: Automatic Text Simplification, Readability, Evaluation

1. Introduction

The accessibility of written information is an important question: outside natural language processing (NLP), domains like medicine (Gu et al., 2024) or business (Huong Dau et al., 2024) have been studying the readability of the documents they produce (e.g., medical reports or information for patients, business reports for shareholders). Usually, those studies are performed using traditional readability formulas, like the Flesch Reading Ease (Flesch, 1948) or Dale-Chall (Dale and Chall, 1948) formulas mainly developed for written English texts. Recently, the reliability of these formulas has been questioned (Alzaid et al., 2024). In NLP, automatic text simplification (ATS) aims at transforming texts to make them more accessible, while preserving their meaning (Saggion, 2017).

Since ATS aims at making texts more accessible, readability is a natural way to frame its goal. However, it is not clear to what extent readability measures and judgments on ATS systems (e.g., human judgment or automatic metrics) correlate with each other. In this paper, we investigate the position of readability measures in the ATS landscape. While readability is regularly mentioned in ATS studies, ATS and automatic readability assessment (ARA) have largely developed in parallel, with limited cross-fertilization (Vajjala, 2022).

While there have been studies on the correlations between ATS evaluation metrics and human judgment (Alva-Manchego et al., 2021; Cripwell et al., 2024), the correlations between these two aspects and commonly available readability measures have not been studied. Our research question is: *to what extent do readability measures correlate with (a) human judgment and (b) automatic evaluation metrics in ATS?*

We are interested in this research question because readability is used differently across ATS research. Some studies incorporate a formula (e.g., FKGL) directly into a training loss (Flores et al., 2023), others include CEFR (Common European Framework of Reference for Languages[1]) levels in LLM prompts (Imperial and Tayyar Madabushi, 2023; Maddela and Alva-Manchego, 2025), and others compute hundreds of linguistic features to characterize a corpus (Battisti et al., 2020; Vajjala and Lučić, 2018), with no shared rationale for why any of these choices should reflect simplification quality.

We argue that this inconsistency is symptomatic of a broader issue: ATS lacks a clear *construct* for simplification quality. The term construct (bor-

*Research partially done while employed at UGent.

[1] https://www.coe.int/en/web/common-european-framework-reference-languages/level-descriptions

Proceedings of the Joint Workshop on Readability and Text Simplification (READIxTSAR) @ LREC 2026, pages 210–226
11 May 2026.

rowed from psychometrics) refers to the theoretical concept that an evaluation instrument is supposed to measure. In ATS research, there is a need to define what "simplification quality" refers to before attempting to measure it. Lexical complexity, syntactic structure, meaning preservation and fluency are all dimensions that readability features and ATS metrics capture separately, but there is no shared definition of what their combination is supposed to measure. It is hard to tell whether two systems that score differently on BLEU and SARI metrics are actually different in quality, or they are just optimizing for different aspects of the same vague objective. ATS research relies on three main approaches to evaluate simplification quality: human judgment, readability measures, and automatic metrics. We empirically test whether these converge, as a necessary condition for treating them as proxies for the same underlying construct. Any divergence would not render them useless but would reinforce the need for a clearer construct definition.

More specifically, this paper makes the following contributions: (i) a correlation study between readability measures and human judgment on two English datasets (sentence-level and document-level); (ii) a correlation study between readability measures and standard ATS evaluation metrics; (iii) a discussion of implications for construct definition in ATS.

2. Related Work

In this section, we provide some background information about readability and ATS research (Sections 2.1, 2.2) before discussing how the two fields interact with each other (Section 2.3).

2.1. Readability

Readability research dates back to the 1920s, initially motivated by the need to assess the suitability of texts for school-age readers (Lively and Pressey, 1923). This first method relied on a list of word frequencies (Thorndike, 1921), based on the assumption that texts made of frequent words are more readable. For a more detailed historical overview, see François (2015). We briefly summarize the key periods below.

The early period of readability research consisted of identifying predictors and tuning coefficient weights out of corpus-based observations and annotations by humans. The most famous readability formulas for English are Flesch Reading Ease (Flesch, 1948, FRE) and Flesch-Kincaid Grade Level (Kincaid et al., 1975, FKGL), which rely on word count and number of syllables per word.

Early NLP-based approaches (1990s–2000s) relied on regression, latent semantic analysis and language modeling (Daoust et al., 1996; Foltz et al., 1998; Si and Callan, 2001). In the 2020s, ARA (automatic readability assessment) has developed into a lively line of research (Vajjala, 2022). ARA has been explored with distributional text representations and with linguistic features. The distributional text representations follow the advancements of research in machine learning, notably with the development of transformers (Vaswani et al., 2017). Regarding linguistic features, the way to select and leverage them is still an open question. Nonetheless, research on this question is facilitated by the appearance of tools that can be used to compute an increasingly high number of features, for example for English (Kyle et al., 2021, 2018; Lu, 2010; Crossley et al., 2019) or French (Wilkens et al., 2022). These tools produce raw analyses of texts with hundreds of features but there are no recommendations about how to select and use them. This gap has fueled new research aimed at combining linguistic feature vectors with distributional representations to improve automatic readability assessment, based on the hypothesis that the two types of information are complementary (Deutsch et al., 2020; Lee et al., 2021; Wilkens et al., 2024).

The readability features depend heavily on the language that is under study, and the aforementioned tools rely on language-dependent resources such as reference corpora, vocabulary lists, or pre-trained models (e.g., for POS-tagging or syntactic analysis). So far, most studies on readability focus on English, due to the availability of tools and resources.

2.2. Automatic Text Simplification

In this section, we briefly describe ATS to provide a background for the discussion of how it integrates considerations about readability, as covered in Section 2.3.

Methods ATS has traditionally been performed at the sentence level (Saggion, 2017) and continues to be an active area of research (Kew et al., 2023). In the early works, the goal was to make sentences simpler to handle as an input for other NLP systems such as syntactic parsers (Chandrasekar et al., 1996). It was only later explored as a means of simplifying texts to make them easier to understand by humans (Carroll et al., 1999). These initial computational methods were rule-based and targeted only specific operations in a sentence (Cardon and Bibal, 2023) (e.g., removing appositive clauses, changing the voice of a sentence from passive to active). The recent developments of generative models have shifted ATS research

towards document-level simplification (Sun et al., 2021), notably with multi-agent architectures (Mo and Hu, 2024; Fang et al., 2025).

Evaluation. Evaluation of ATS is an open question. Traditional readability formulas, mostly FKGL or adaptations of FRE for other languages, are often reported (Engelmann et al., 2024; Flores et al., 2023; Aluisio et al., 2010; Paula and Camilo-Junior, 2024; Štajner and Saggion, 2013). However, they correlate poorly with human judgment on ATS (Tanprasert and Kauchak, 2021; Alva-Manchego et al., 2021). The most common automatic metrics are BLEU (Papineni et al., 2002), SARI (Xu et al., 2016), BERTScore (Zhang et al., 2020) and LENS (**?**). D-SARI (Sun et al., 2021) extends SARI to the document level. BLEU and BERTScore compare the text output to one or more references, while (D-)SARI adds the input into the computation. Although BLEU is often interpreted as an indicator of meaning preservation, SARI of simplicity, and BERTScore of meaning preservation and fluency, prior work has shown that these metrics correlate poorly and inconsistently with human judgment (Alva-Manchego et al., 2021; Sulem et al., 2018a).

These three criteria (fluency, simplicity and meaning preservation) are also used by human annotators to evaluate sentence simplification, typically on 5-point Likert scales. For document-level simplification, methods for human evaluation are not stabilized yet. Cripwell et al. (2024) use the same criteria with binary questions instead of Likert scales. Sun et al. (2021) ask human judges to evaluate "overall simplicity" that they define as simplicity with other quality criteria (e.g., ease of reading and meaning preservation). Vásquez-Rodríguez et al. (2023) ask human judges to evaluate textual coherence and Agrawal and Carpuat (2024) evaluate meaning preservation by studying human accuracy on reading comprehension questions about the simplified text.

2.3. Readability and Text Simplification

In most research where readability and ATS interact, readability is leveraged through linguistic features to give information about the datasets (Battisti et al., 2020; Vajjala and Lučić, 2018; Yaneva et al., 2016; Štajner and Saggion, 2013; Dell'Orletta et al., 2011; Aluisio et al., 2010). Other studies rely on linguistic features for data selection instead (Jingshen et al., 2024). Jingshen et al. (2024) leverage readability features, in conjunction with similarity measures, to mine sentence pairs and produce a parallel corpus for Chinese idiom simplification. De Martino (2023) investigates the link between eye-tracking data and readability features on Italian data. Although preliminary, this study suggests that eye-tracking is a promising method for assessing the cognitive processing cost of simplified texts, and could complement annotation-based measures of simplification quality.

Some simplification studies use readability features or metrics in their evaluation protocol as well. Scholz and Wenzel (2025) evaluate 18 readability features (i.e., syntactic, POS-based, semantic and fluency features) for English and German text simplification. They find that some measures (e.g., semantic and fluency features) transfer across languages, while the behavior of statistical, POS-based and syntactic metrics appears to be strongly language-dependent. Paula and Camilo-Junior (2024) use a Portuguese adaptation of FRE as an evaluation metric for ATS. Engelmann et al. (2024) use the FRE and Dale-Chall formulas to perform pairwise comparisons to rank simplifications. They compare them to human judgments and GPT 3.5 and find that Dale-Chall has the highest correlation to human judgment, above GPT 3.5, while FRE obtains the lowest correlations.

Readability can also be incorporated in ATS methods. Flores et al. (2023) use a bounded FKGL as a component of their loss in a neural model for text simplification. Maddela and Alva-Manchego (2025) and Imperial and Tayyar Madabushi (2023) prompt LLMs for document-level simplification by including CEFR levels in the prompt. Using CEFR as a proxy for readability was introduced with the release of the CEFR-SP dataset (Arase et al., 2022), a corpus of 17,000 English sentences annotated according to CEFR levels.

Readability features have also been used at a more granular level. Lexical complexity features, in particular, have been leveraged for lexical simplification (North et al., 2025). Hazim et al. (2022) introduce a system that highlights complex words in a text editor to help humans manually simplify texts. Maddela and Xu (2018) use lexical features to rank candidates for substitution in a neural lexical simplification system. Grigonyte et al. (2014) rely on lexical and morphological features to perform complex word identification.

It is worth noting that readability and ATS studies address related but distinct objectives. Readability research primarily targets lexico-syntactic properties of texts and their effect on comprehension by humans regardless of the source text. On the other hand, ATS transforms a source text into a simplified version. Its evaluation typically involves comparing the output to the input (or to a reference). These different perspectives motivate our research question: whether readability measures are aligned with how simplification quality is assessed, both by humans and by automatic metrics.

Overall, we observe that various readability measurement approaches (e.g., linguistic features, for-

mulas, eye-tracking, CEFR levels) are also explored in ATS research. Table 1 provides a structured overview of ATS research and illustrates the diversity of practices and the lack of a shared theoretical framework.

The two approaches most widely present in ATS are traditional formulas (e.g., FRE and FKGL, used as evaluation metrics) and readability features (for providing information about datasets), both of which we examine in this study.

3. Readability Measures and ATS Metrics

Throughout this paper, we use *features* to refer to linguistic features, *formulas* to refer to traditional readability formulas (e.g., FRE, FKGL), *measures* as an umbrella term for features and formulas, and *metrics* for automatic ATS evaluation metrics (e.g., BLEU, SARI).

3.1. Data

We focus on English because it is easier to find (i) human-annotated simplification datasets and (ii) feature-extraction tools for our study of correlations between readability measures and ATS evaluation criteria.

To study how readability measures correlate with the evaluation protocols in ATS, we rely on two datasets: at the sentence level (Alva-Manchego et al., 2021) and at the document level (Maddela and Alva-Manchego, 2025). They are both labeled with human judgment. Both works studied the link between automatic metrics and human judgment. Building on their findings, we explore (a) the link between readability measures and human judgment, and (b) the link between readability measures and automatic metrics. Below is a description of the datasets.

SimplicityDA. For the sentence-level study, we use Simplicity-DA (Alva-Manchego et al., 2021)[2]. It consists of 600 sentence-level ATS system outputs in English, all annotated by 15 crowdworkers based on three criteria (fluency, simplicity and meaning preservation) on a 0-100 scale. Besides the human judgments, the dataset also includes automatic scores (e.g., BLEU, SARI, BERTScore and SAMSA (Sulem et al., 2018b)) for each sentence.

D-Wikipedia. For the document-level study, we use D-Wikipedia (Sun et al., 2021). D-Wikipedia is a corpus of aligned paragraph pairs extracted from the English Wikipedia for the complex side and Simple English Wikipedia for the simple side. Maddela and Alva-Manchego (2025) released a subset of 100 paragraph pairs from D-Wikipedia, each with 4 simplified versions produced by automatic systems, resulting in 500 paragraph pairs. These 500 pairs were rated by three human judges on a 5-point Likert scale on three criteria (fluency, simplicity and meaning preservation). We compute the automatic metrics values (e.g., BLEU, SARI, D-SARI, BERTScore and LENS) with the code provided with the dataset[3].

3.2. Readability Measures

Readability Features. As discussed in Section 2, readability assessment is mostly explored with two types of text representations (distributional embeddings and textual features). As distributional embeddings are already leveraged for simplification methods (Kew et al., 2023) and evaluation (Zhang et al., 2020), we focus on textual features. To compute these features, we use four tools that implement a total of 1,066 readability-related features for English[4].

TAALED (Kyle et al., 2021)[5] computes 38 features related to lexical diversity, such as different type-token ratios or MTLD (Measures of Textual Lexical Diversity).

TAALES (Kyle et al., 2018)[6] computes 484 features related to lexical sophistication, i.e., the degree to which the vocabulary used in a text is advanced, infrequent, or complex relative to a reference population. Many of these features are variations of word frequency (computed on various corpora such as BNC (Consortium, 2007, The British National Corpus) and COCA (Davies, 2008, The Corpus of Contemporary American English)). Other features are related to lexical neighborhood (i.e., the number of words that are orthographically or phonologically similar to a given word, a measure linked to word recognition difficulty), age of acquisition (i.e., the estimated age at which a word is typically learned by native speakers, a correlate of word difficulty), psycholinguistic norms (e.g., concreteness, imageability, meaningfulness).

TAASSC (Lu, 2010)[7] computes 376 features related to syntactic sophistication. These features rely on grammatical dependency analysis and part-of-speech tagging. Some examples of these fea-

[2] https://github.com/feralvam/metaeval-simplification

[3] https://github.com/cardiffnlp/document-simplification

[4] The complete list of features and their formulas is available in the documentation of each tool.

[5] https://www.linguisticanalysistools.org/taaled.html

[6] https://www.linguisticanalysistools.org/taales.html

[7] https://www.linguisticanalysistools.org/taassc.html

Usage	Reference	Language(s)	Measure type
Corpus description	Aluisio et al. (2010)	Portuguese	Features / Formulas
	Dell'Orletta et al. (2011)	Italian	Features
	Štajner and Saggion (2013)	Spanish	Features / Formulas
	Yaneva et al. (2016)	English	Features
	Vajjala and Lučić (2018)	English	Features
	Battisti et al. (2020)	German	Features
	De Martino (2023)	Italian	Features + eye-tracking
Corpus construction	Jingshen et al. (2024)	Chinese	Features
Evaluation	Engelmann et al. (2024)	English	Formulas (FRE, Dale-Chall)
	Scholz and Wenzel (2025)	German	Features
Method	Grigonyte et al. (2014)	Swedish	Lexical features (CWI)
	Maddela and Xu (2018)	English	Lexical features
	North et al. (2025)	English	Lexical features
	Hazim et al. (2022)	Arabic	Lexical features
	Flores et al. (2023)	English	Formula (FKGL in loss)
	Paula and Camilo-Junior (2024)	Portuguese (Braz.)	Formula (FRE as target)
	Imperial and Tayyar Madabushi (2023)	English	CEFR (in prompt)
	Maddela and Alva-Manchego (2025)	English	CEFR (in prompt)
	Barayan et al. (2025)	English	CEFR (in prompt)

Table 1: Overview of works at the intersection of readability and ATS, grouped by how readability measures are used. CWI = Complex Word Identification.

tures are conjunctions per clause, verbal modifiers per noun phrase, frequency of constructions compared to references coming from different corpora (e.g., BNC, COCA and others) or more traditional ones (e.g., average sentence length).

TAACO (Crossley et al., 2019)[8] computes 168 features related to cohesion. Some examples include semantic similarity between word2vec (Mikolov et al., 2013) embeddings of adjacent sentences, or token overlap between adjacent sentences or paragraphs.

These four tools cover complementary dimensions of text complexity. Specifically, TAALED and TAALES focus on lexical aspects (diversity and sophistication, respectively), TAASSC targets syntactic structure, and TAACO captures discourse-level cohesion. Together, they cover the main dimensions associated with readability in the literature, providing a broad and principled basis for our correlation study.

Readability Formulas. We also compute the following set of traditional readability formulas for English, using the `textstat` Python library: Flesch Reading Ease (Flesch, 1948), Dale-Chall (Dale and Chall, 1948), Gunning-Fog (Gunning, 1952), Linsear Write (O'hayre, 1966), ARI (Smith and Senter, 1967), SMOG (Mc Laughlin, 1969), Flesch-Kincaid Grade Level (Kincaid et al., 1975), and Coleman-Liau (Coleman and Liau, 1975).

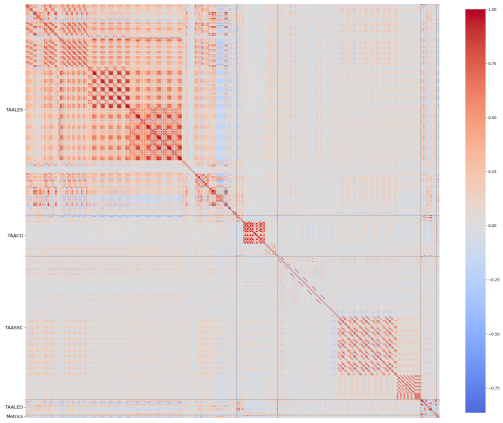
[8] https://www.linguisticanalysistools.org/taaco.html

4. Experiments

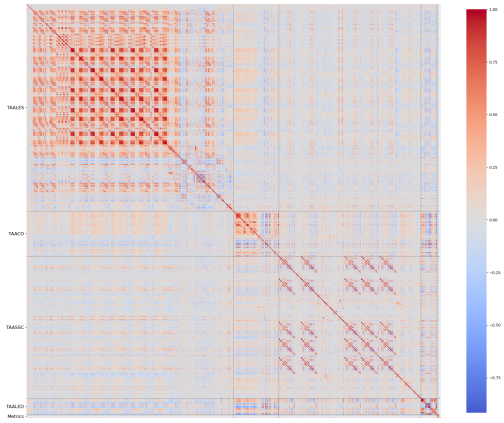
4.1. Readability Measures

First, we compute the correlations among the readability measures themselves (i.e., the 1,066 linguistic features from the four tools and the eight traditional readability formulas described in Section 3.2). Figures 1a and 1c show the correlation matrices computed on the SimplicityDA dataset (at the sentence level), respectively on the difference between the simplified and original sentences, and on the simplifications. Figures 1b and 1d show the correlation matrices computed on the D-Wikipedia dataset, respectively on the difference between the simplified and original sentences, and on the simplifications. Our findings include: (i) the measures mostly correlate with other measures of the same type, (ii) measures computed at the document level show higher absolute values and (iii) measures computed on the difference between original texts and simplifications exhibit lower absolute values.

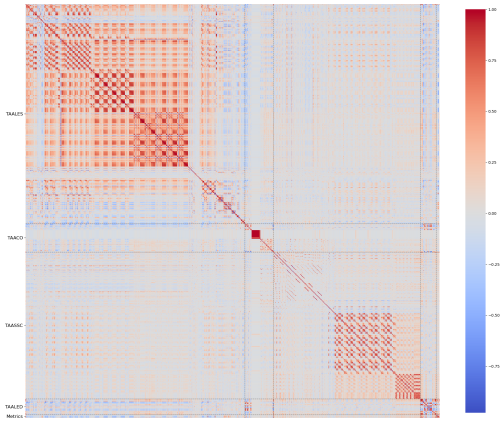
These observations are visible in Figure 1: in all four heatmaps, the block structure along the diagonal confirms that intra-group correlations dominate. The higher saturation in the document-level heatmaps (Figures 1b and 1d) visually reflects observation (ii). In the delta heatmaps (Figures 1a and 1b), the overall lighter coloration reflects observation (iii): subtracting original from simplified values reduces systematic co-variation between measures. Consequently, the few correlation clusters that remain visible in these heatmaps correspond to measures that respond differently to the



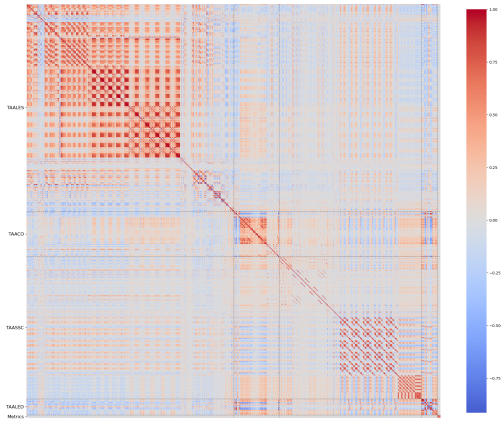
(a) Difference between original and simple, SimplicityDA (sentence-level).



(b) Difference between original and simple, D-Wikipedia (document-level).



(c) Simple side of SimplicityDA (sentence-level).



(d) Simple side of D-Wikipedia (document-level).

Figure 1: Pearson correlation matrices of readability measures and metrics. Dashed lines indicate the boundaries of feature groups (from top to bottom, and the same from left to right: TAALES, TAACO, TAASSC, TAALED, and Metrics).

Dataset	Mode	Criterion	Top $\|r\|$	10th $\|r\|$
SimplicityDA	simp	simplicity	.199	.159
	simp	fluency	.262	.195
	simp	meaning	.295	.231
	delta	simplicity	.215	.150
	delta	fluency	.300	.211
	delta	meaning	.426	.350
D-Wikipedia	simp	simplicity	.205	.091
	simp	fluency	.103	.096
	simp	meaning	.433	.192
	delta	simplicity	.298	.234
	delta	fluency	.291	n=4
	delta	meaning	.342	.211

(a) Human judgment correlations.

Dataset	Mode	Metric	Top $\|r\|$	10th $\|r\|$
SimplicityDA	simp	BERTScore	.339	.273
	simp	BLEU	.352	.290
	simp	SAMSA	.309	.253
	simp	SARI	.378	.322
	delta	BERTScore	.518	.413
	delta	BLEU	.403	.282
	delta	SAMSA	.199	.184
	delta	SARI	.499	.410
D-Wikipedia	simp	D-SARI	.220	.129
	simp	LENS	.347	.185
	simp	BERTScore	.143	.090
	delta	D-SARI	.305	.251
	delta	LENS	.279	.217
	delta	BERTScore	.219	.179

(b) Automatic metrics correlations.

Table 2: Top absolute $|r|$ values among significant Pearson correlations between readability measures and evaluation criteria. "simp" = simple version only; "delta" = difference simple−original; 10th $|r|$ = lowest of the top 10 values (n=4: fewer than 10 significant correlations found). Full tables per feature in Appendix A.1.

simplification operation itself. Therefore, they are considered to be the most informative ones for our study.

4.2. Measures and Human Judgment

To compare readability measures (the features with the four readability tools, and the readability formulas) and human judgment, we compute them all on both datasets: SimplicityDA for the sentence-level (100 original sentences and 600 simplifications including 100 human-written ones) and D-Wikipedia for the document-level (100 original paragraphs and 500 simplifications including 100 human-written ones). For each dataset, we compute the measures on both sides (original and simplified) separately. We compute the correlations with human judgment in two ways: (i) on the measures obtained on the simplified versions only, and (ii) on the difference between the measures obtained on the original texts and the ones obtained on the simplified versions. The first case focuses on simplicity, and the second on simplification by including a comparison with the original text.

For both datasets, we report the correlations on the three criteria (simplicity, fluency and meaning preservation) for human judgment.

4.3. Measures and Automatic Metrics

Metrics like BLEU and SARI measure string overlap rather than linguistic complexity. However, testing whether readability features correlate with them is also informative when researchers use these metrics as proxies for measuring simplification quality. Readability features can also be used for the same purpose; a low correlation would confirm that they capture fundamentally different aspects of the output.

To study the correlations between readability measures and automatic simplification metrics, we proceed in the same way as for the correlations between readability measures and human judgment. We report scores on the following automatic metrics: BLEU, SARI, BERTScore for SimplicityDA (sentence-level), and D-SARI, BERTScore, and LENS for D-Wikipedia (document-level).

Regarding the metrics that require references (BLEU, SARI), we use all the references that are available in Simplicity-DA (i.e., for each original sentence: 10 references from ASSET (Alva-Manchego et al., 2020), 1 from TurkCorpus (Xu et al., 2016) and 1 from HSplit (Sulem et al., 2018a)). For D-Wikipedia, we use the single reference simplification that is provided for each original text.

5. Results

5.1. Measures and Human Judgment

We report the top significant correlations between readability measures and human judgment in Table 2a (full tables per dataset and criterion in Appendix A.1).

For SimplicityDA, the highest absolute coefficient values are obtained with the meaning criterion computed on delta, with the top 10 ranging from -0.43 to -0.35. All of the other criteria have top absolute coefficient values between 0.15 and 0.30. In that regard, readability measures (features and formulas) and human judgment on simplification quality do not correlate well at the sentence level. We also observe that all absolute values are higher when computed on the delta rather than on simplifications only. As the human judges were asked to rate simplification instead of simplicity, the difference between simplicity and simplification has an effect on both humans and measures (while the coefficient values are low).

These results call for caution in using readability features as standalone indicators of simplification quality. They can help characterize corpora or contribute to composite metrics, but they should not be read as direct proxies for what human judges find simple.

Regarding the D-Wikipedia dataset, the observations are similar. Meaning exhibits the highest coefficient values, although with a higher discrepancy between the top 1 and 10 values (.433 vs .192 for simp-meaning and .342 vs .211 for delta-meaning). We found only 4 significant correlations for delta-fluency, suggesting limited correlation at best (the highest values being .291 for delta and .103 for simp). As for SimplicityDA, the values are generally higher for delta than for simp.

The observation that only 4 significant correlations were found for delta-fluency on D-Wikipedia is notable. It may reflect either that grammaticality is not captured by the features we use, or it is harder to detect it at the document level due to greater variability in the simplifications.

Regarding simplicity, the values are low for both datasets. The most correlated set of observations is delta-simplicity with top 10 absolute values ranging from .234 to .298.

Not many features are found in more than one set of highest correlating values. For SimplicityDA, we observe several kinds of type/token ratio (TTR). Root TTR and log TTR are the most recurrent, appearing in 3 and 4 sets of observations out of 6, respectively. For D-Wikipedia, we see that the word count appears in 4 sets of observations, and corpus-based metrics (especially calibrated on COCA but also on the BNC) appear in 5 out of 6 sets.

5.2. Measures and Automatic Metrics

We report the correlations between readability measures and automatic metrics in Table 2b (full tables in Appendix A.1).

For SimplicityDA, BERTScore has the highest correlation values, especially when the features are computed on delta (with a top 10 ranging from .413 to .518). SAMSA exhibits the lowest correlation values and is the only metric to correlate better when the features are computed on the simple texts only. SAMSA measures structural simplification (sentence splitting), which correlates with higher token counts on the simple side rather than with the *reduction* of linguistic features that other metrics reward. It is an example of how different metrics operationalize different sub-tasks of simplification, reinforcing the argument for a clearer construct definition.

Regarding D-Wikipedia, the correlations are generally lower. BERTScore has the lowest correlation values (from 0.09 to 0.219 across both computation modes), while LENS exhibits a slightly higher level of correlation than D-SARI.

Regarding the features themselves, COCA-based features are present in all criteria with D-Wikipedia, while they are only present for delta-SAMSA with SimplicityDA.

TTR measures are present in 5 out of 8 sets of observations for SimplicityDA, and are completely absent for D-Wikipedia. These observations suggest that sentence simplification and document simplification evaluation do not involve the same phenomena. It is worth noting that TTR correlates better with sentence-level simplification than with document-level simplification, even though TTR is commonly used as a rough measure of text complexity elsewhere.

5.3. Cross-level Observations

Comparing the two datasets reveals differences between sentence-level and document-level simplification that are not just a matter of scale. For human judgment, the overall pattern holds at both levels. The meaning criterion yields the highest correlation values, while the simplicity and fluency criteria remain low. This suggests that readability features are more sensitive to semantic distortion than to structural or fluency changes, regardless of granularity.

The divergence is more visible in which features matter. TTR measures are among the top correlations for SimplicityDA (present in 5 out of 8 sets), but completely absent for D-Wikipedia. This is consistent with a known limitation of TTR. It is sensitive to text length, since longer texts tend to accumulate repetitions that drive TTR down, irrespective of lexical richness. At the document level (where texts are longer and more varied), TTR loses its discriminative power. Conversely, frequency-based features calibrated on COCA appear across all criteria for D-Wikipedia, but only marginally for SimplicityDA. Corpus frequency measures may capture the lexical choices that distinguish document-level simplification outputs better.

These observations suggest that the readability features for sentence-level and document-level evaluation are not the same. Feature selection should be adapted to the granularity of the task rather than applying it uniformly.

6. Conclusion

This paper investigated how readability measures correlate with human judgment and automatic metrics in ATS. The question is relevant because the field already uses readability measures in evaluation protocols and training objectives, often without theoretical grounding, and because prior work has documented inconsistent correlations between human judgment and automatic metrics. We acknowledge that our findings do not go towards dissipating the uncertainty that the field has been experiencing. That said, our findings point to a lack of a well-defined construct in the ATS ecosystem. In the case of ATS, the question is whether "simplification quality" is a coherent, measurable construct, or whether it bundles together with other criteria (e.g., fluency, lexical simplicity, structural changes, meaning preservation) that different metrics and features capture in different ways. Our results show that readability measures, automatic metrics and human judgments do not consistently converge. When a system performs lexical simplification, lexical sophistication features will improve and FKGL may also improve, but SARI may penalize the output if the reference chose a different synonym. In practice, a simplified text results from dozens of such operations pulling in different directions, and each measure captures only one facet of the result. Without a shared construct, none of these scores can be read as unambiguous evidence of quality. This suggests that the field has not yet settled on a shared construct for simplification quality, and working towards such a definition would benefit both evaluation design and system development.

7. Limitations

The main limitations of our work concern the quantity and quality of data. We used the only data with human judgment that were available to us in English. Both datasets have limitations. SimplicityDA covers outputs from a limited number of systems, which may not reflect the full diversity of simplification approaches. D-Wikipedia is based on a

small annotated subset (100 paragraph pairs), with only three annotators per pair. These factors may contribute to the low correlations we observe independently of any conceptual mismatch between constructs. These findings may therefore vary on higher-quality or more recent corpora, on other languages, and with other human annotators. While this impairs the generalizability of our study, it reinforces our point that the field should, as a community, focus more on clearly defining the task and producing better-controlled evaluation data.

A further limitation is the univariate nature of our analysis. We compute each feature's correlation independently, which does not capture interactions between features. Methods such as principal component analysis or multivariate regression could reveal richer structure in the relationship between readability features and evaluation outcomes.

8. Acknowledgements

We would like to thank the reviewers for their valuable comments. Rémi Cardon is partially funded by grant PID2023-148577OB-C21 (Human-Centered AI: User-Driven Adapted Language Models-HUMAN_AI) by MICIU/AEI/10.13039/501100011033 and by FEDER/UE.

9. Plain Language Summary

Making written information easier to read is important, especially when using computers to rewrite complex texts. This process is called automatic text simplification (ATS). However, experts do not always agree on how to measure if a text has really become easier to read. There are three main ways to check: asking people to judge the text, using computer formulas that estimate readability, and using automatic scores that compare the original and simplified texts.

In our study, we wanted to see if these three ways of measuring simplification quality actually agree with each other. We used two collections of English texts that had already been rated by people for how simple, fluent, and meaningful they were. We also used computer programs to calculate over a thousand different features of the texts, as well as eight common readability formulas.

We found that readability formulas and features do not match well with what people think, or with the automatic scores. The only area where there was some agreement was in how well the meaning of the text was preserved. We also noticed that the best features for judging simplification are different for short sentences and for longer documents.

Overall, our research shows that there is no single, agreed-upon way to measure if a text has been successfully simplified. This means that more work is needed to find better ways to define and measure simplification quality.

10. Bibliographical References

Sweta Agrawal and Marine Carpuat. 2024. Do text simplification systems preserve meaning? a human evaluation via reading comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448.

Sandra Aluisio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2021. The (un)suitability of automatic evaluation metrics for text simplification. *Computational Linguistics*, 47(4):861–889.

Mohammad Alzaid, Faisal R Ali, and Emma Stapleton. 2024. Limitations of readability assessment tools. *European Archives of Oto-Rhino-Laryngology*, 281(9):5021–5022.

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Abdullah Barayan, Jose Camacho-Collados, and Fernando Alva-Manchego. 2025. Analysing zero-shot readability-controlled sentence simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6762–6781, Abu Dhabi, UAE. Association for Computational Linguistics.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and

text simplification of German. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3302–3311, Marseille, France. European Language Resources Association.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway. Association for Computational Linguistics.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

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

BNC Consortium. 2007. British national corpus 1994. Literary and Linguistic Data Service.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2024. Evaluating document simplification: On the importance of separately assessing simplicity and meaning preservation. *LREC-COLING 2024*, page 1.

Scott A Crossley, Kristopher Kyle, and Mihai Dascalu. 2019. The tool for the automatic analysis of cohesion 2.0: Integrating semantic similarity and text overlap. *Behavior research methods*, 51:14–27.

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

François Daoust, Léo Laroche, and Lise Ouellet. 1996. Sato-calibrage: Présentation d'un outil d'assistance au choix et à la rédaction de textes pour l'enseignement. *Revue québécoise de linguistique*, 25(1):205–234.

Mark Davies. 2008. The corpus of contemporary american english (coca). Available online at https://www.english-corpora.org/coca/.

Maria De Martino. 2023. Processing effort during reading texts in young adults: Text simplification, readability assessment and preliminary eye-tracking data. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 179–184, Venice, Italy. CEUR Workshop Proceedings.

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 Second Workshop on Speech and Language Processing for Assistive Technologies*, pages 73–83, Edinburgh, Scotland, UK. Association for Computational Linguistics.

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.

Björn Engelmann, Christin Katharina Kreutz, Fabian Haak, and Philipp Schaer. 2024. ARTS: Assessing readability & text simplicity. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14925–14942, Miami, Florida, USA. Association for Computational Linguistics.

Dengzhao Fang, Jipeng Qiang, Xiaoye Ouyang, Yi Zhu, Yunhao Yuan, and Yun Li. 2025. Collaborative document simplification using multi-agent systems. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 897–912, Abu Dhabi, UAE. Association for Computational Linguistics.

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

Lorenzo Jaime Flores, Heyuan Huang, Kejian Shi, Sophie Chheang, and Arman Cohan. 2023. Medical text simplification: Optimizing for readability with unlikelihood training and reranked beam search decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4859–4873, Singapore. Association for Computational Linguistics.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Thomas François. 2015. When readability meets computational linguistics: a new paradigm in readability. *Revue française de linguistique appliquée*, XX(2):79–97.

Gintarė Grigonyte, Maria Kvist, Sumithra Velupillai, and Mats Wirén. 2014. Improving readability of Swedish electronic health records through lexical simplification: First results. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 74–83, Gothenburg, Sweden. Association for Computational Linguistics.

Joey Z Gu, Grayson L Baird, Antonio Escamilla Guevara, Young-Jin Sohn, Melis Lydston, Christopher Doyle, Sarah EA Tevis, and Randy C Miles. 2024. A systematic review and meta-analysis of english language online patient education materials in breast cancer: Is readability the only story? *The Breast*, page 103722.

Robert Gunning. 1952. The technique of clear writing. *McGraw-Hill*.

Reem Hazim, Hind Saddiki, Bashar Alhafni, Muhamed Al Khalil, and Nizar Habash. 2022. Arabic word-level readability visualization for assisted text simplification. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 242–249, Abu Dhabi, UAE. Association for Computational Linguistics.

Nam Huong Dau, Duy Van Nguyen, and Hai Thi Thanh Diem. 2024. Annual report readability and firms' investment decisions. *Cogent Economics & Finance*, 12(1):2296230.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2023. Flesch or fumble? evaluating readability standard alignment of instruction-tuned language models. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 205–223, Singapore. Association for Computational Linguistics.

Zhang Jingshen, Chen Xinglu, Qiu Xinying, Wang Zhimin, and Feng Wenhe. 2024. Readability-guided idiom-aware sentence simplification (RISS) for Chinese. In *Proceedings of the 23rd Chinese National Conference on Computational Linguistics (Volume 1: Main Conference)*, pages 1183–1200, Taiyuan, China. Chinese Information Processing Society of China.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

J Peter Kincaid, RP Fishburne, RL Rogers, and BS Chissom. 1975. Derivation of new readability formulas (automated reliability index, fog count and flesch reading ease formula) for navy enlisted personnel (research branch report 8-75). memphis, tn: Naval air station; 1975. *Naval Technical Training, US Naval Air Station: Millington, TN*.

Kristopher Kyle, Scott Crossley, and Cynthia Berger. 2018. The tool for the automatic analysis of lexical sophistication (taales): Version 2.0. *Behavior research methods*, 50:1030–1046.

Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2):154–170.

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.

Bertha A Lively and Sidney L Pressey. 1923. A method for measuring the vocabulary burden of textbooks. *Educational administration and supervision*, 9(7):389–398.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Mounica Maddela and Fernando Alva-Manchego. 2025. Adapting sentence-level automatic metrics for document-level simplification evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6444–6459, Albuquerque, New Mexico. Association for Computational Linguistics.

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.

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

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

Kaijie Mo and Renfen Hu. 2024. ExpertEase: A multi-agent framework for grade-specific document simplification with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9080–9099, Miami, Florida, USA. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2025. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, 63(1):111–134.

John O'hayre. 1966. *Gobbledygook has gotta go*. US Department of the Interior, Bureau of Land Management.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Antonio Flavio Paula and Celso Camilo-Junior. 2024. Evaluating the simplification of Brazilian legal rulings in LLMs using readability scores as a target. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 117–125, Miami, Florida, USA. Association for Computational Linguistics.

Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool Publishers.

Karen Scholz and Markus Wenzel. 2025. Evaluating readability metrics for German medical text simplification. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6049–6062, Abu Dhabi, UAE. Association for Computational Linguistics.

Luo Si and Jamie Callan. 2001. A statistical model for scientific readability. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576.

E.A. Smith and R.J. Senter. 1967. *Automated Readability Index*. AMRL-TR. Aerospace Medical Research Laboratories, Aerospace Medical Division, Air Force Systems Command.

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.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

Renliang Sun, Hanqi Jin, and Xiaojun Wan. 2021. Document-level text simplification: Dataset, criteria and baseline. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7997–8013, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics.

Edward L Thorndike. 1921. Word knowledge in the elementary school. *Teachers College Record*, 22(4):1–27.

Sowmya Vajjala. 2022. Trends, limitations and open challenges in automatic readability assessment research. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5366–5377, Marseille, France. European Language Resources Association.

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.

Laura Vásquez-Rodríguez, Matthew Shardlow, Piotr Przybyła, and Sophia Ananiadou. 2023. Document-level text simplification with coherence evaluation. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 85–101, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez,

Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Rodrigo Wilkens, David Alfter, Xiaoou Wang, Alice Pintard, Anaïs Tack, Kevin P. Yancey, and Thomas François. 2022. FABRA: French aggregator-based readability assessment toolkit. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1217–1233, Marseille, France. European Language Resources Association.

Rodrigo Wilkens, Patrick Watrin, Rémi Cardon, Alice Pintard, Isabelle Gribomont, and Thomas François. 2024. Exploring hybrid approaches to readability: experiments on the complementarity between linguistic features and transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 2316–2331, St. Julian's, Malta. Association for Computational Linguistics.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Victoria Yaneva, Irina Temnikova, and Ruslan Mitkov. 2016. Evaluating the readability of text simplification output for readers with cognitive disabilities. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 293–299, Portorož, Slovenia. European Language Resources Association (ELRA).

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A. Appendices

A.1. Correlation Tables

The following tables report the top 10 readability features or metrics by absolute value of significant Pearson correlation coefficient, for each combination of dataset, computation mode (simp = simple side only; delta = difference simple−original) and evaluation criterion.

(a) simp – simplicity

Variable	r
log_ttr_aw	0.199
log_ttr_cw	0.193
McD_CD_FW	0.178
basic_connectives	-0.173
lemma_ttr	0.173
lemma_mattr	0.173
MRC_Familiarity_CW	0.168
msttr50_aw	0.164
mattr50_aw	0.164
bigram_lemma_ttr	0.159

(b) delta – simplicity

Variable	r
conjunctions	0.215
basic_connectives	0.208
av_pobj_deps_NN	-0.178
log_ttr_cw	-0.176
adv_ttr	-0.169
log_ttr_aw	-0.167
av_pobj_deps	-0.159
MRC_Familiarity_CW	-0.153
MRC_Imageability_CW	-0.152
hyper_verb_noun_Sav_P1	-0.150

(c) simp – fluency

Variable	r
COCA_magazine_bi_MI	0.262
log_ttr_aw	0.242
COCA_news_bi_MI	0.237
COCA_fiction_bi_MI	0.237
basic_connectives	-0.218
COCA_spoken_bi_MI	0.217
log_ttr_cw	0.217
conjunctions	-0.203
conj_per_cl	-0.195
acad_lemma_attested	0.195

(d) delta – fluency

Variable	r
root_ttr_aw	-0.300
root_ttr_cw	-0.283
basic_ntypes	-0.242
conjunctions	0.239
mtld_ma_wrap_aw	-0.237
basic_ncontent_types	-0.235
basic_connectives	0.226
av_pobj_deps_NN	-0.212
log_ttr_aw	-0.212
root_ttr_fw	-0.211

(e) simp – meaning

Variable	r
root_ttr_cw	0.295
root_ttr_aw	0.286
log_ttr_cw	0.269
basic_ncontent_types	0.254
hyper_verb_noun_Sav_P1	0.239
mtld_ma_wrap_aw	0.234
hyper_verb_noun_Sav_Pav	0.233
hyper_verb_noun_s1_p1	0.232
linsear	0.231
basic_ntypes	0.231

(f) delta – meaning

Variable	r
root_ttr_aw	-0.426
root_ttr_cw	-0.398
basic_ntypes	-0.392
nwords	-0.390
Word Count	-0.383
basic_ntokens	-0.373
basic_ncontent_tokens	-0.365
mtld_ma_wrap_aw	-0.363
basic_ncontent_types	-0.361
basic_nfunction_types	-0.350

Table 3: Top absolute values of significant correlation coefficients ($p < .001$) between human judgment on the SimplicityDA dataset and readability measures.

(a) simp – simplicity

Variable	r
Word Count	-0.205
Kuperman_AoA_FW	-0.138
Phono_N_FW	0.116
Phono_N_H_FW	0.113
hyper_noun_S1_P1	0.110
hyper_noun_Sav_Pav	0.102
MRC_Familiarity_FW	0.096
COCA_fiction_Range_CW	-0.095
BNC_Spoken_3gram_NF	-0.093
poly_adj	-0.091

(b) delta – simplicity

Variable	r
Word Count	0.298
TL_Freq_FW_Log	-0.267
COCA_fiction_Frequency_Log_FW	-0.258
KF_Freq_FW_Log	-0.252
BNC_Written_Freq_FW_Log	-0.249
COCA_news_Frequency_Log_FW	-0.245
COCA_magazine_Frequency_Log_FW	-0.239
AWL_Sublist_5_Normed	0.237
BNC_Spoken_Freq_FW_Log	-0.236
Brown_Freq_FW_Log	-0.234

(c) simp – fluency

Variable	r
COCA_spoken_Trigram_Frequency_Log	0.103
WN_SD_CW	-0.101
COCA_news_tri_2_DP	0.099
Brysbaert_CC_AW	0.098
COCA_magazine_tri_2_MI2	0.098
COCA_magazine_tri_2_DP	0.098
COCA_spoken_tri_prop_20k	0.098
OG_N_H	-0.098
Freq_N_OGH	-0.098
COCA_news_tri_prop_10k	0.096

(d) delta – fluency

Variable	r
AWL_Sublist_10_Normed	-0.291
COCA_fiction_Frequency_FW	-0.216
BNC_Spoken_3gram_NF_Log	0.214
TL_Freq_FW	-0.202

(e) simp – meaning

Variable	r
Word Count	-0.433
Kuperman_AoA_FW	-0.243
Brown_Freq_CW	0.240
TL_Freq_CW	0.239
KF_Freq_CW	0.219
OLDF_FW	0.213
Freq_N_OG_CW	0.211
OG_N_H_CW	0.200
Freq_N_OGH_CW	0.200
poly_adj	-0.192

(f) delta – meaning

Variable	r
Word Count	0.342
Kuperman_AoA_AW	0.270
COCA_spoken_Frequency_Log_CW	-0.255
PLDF_FW	-0.248
COCA_spoken_RL_CW	-0.228
SUBTLEXus_Range_FW	-0.225
COCA_news_RL_FW	-0.224
COCA_spoken_RL_FW	-0.221
KF_Ncats_FW	-0.220
Kuperman_AoA_CW	0.211

Table 4: Top absolute values of significant correlation coefficients ($p < .05$) between human judgment on the DWiki dataset and features.

(a) simp – bertscore_F1

Variable	r
root_ttr_aw	0.339
rootTTRCW	0.317
log_ttr_cw	0.311
mtld_ma_wrap_aw	0.294
hyper_verb_noun_Sav_P1	0.291
hyper_verb_noun_Sav_Pav	0.283
hyper_verb_noun_s1_p1	0.279
log_ttr_aw	0.277
hyper_noun_S1_P1	0.276
basic_ntypes	0.273

(b) delta – bertscore_F1

Variable	r
root_ttr_aw	-0.518
basic_ntypes	-0.466
root_ttr_cw	-0.464
nwords	-0.461
mtld_ma_wrap_aw	-0.457
Word Count	-0.445
basic_ncontent_tokens	-0.434
basic_ncontent_types	-0.432
basic_ntokens	-0.429
linsear	-0.413

(c) simp – bleu

Variable	r
root_ttr_aw	0.352
root_ttr_cw	0.341
linsear	0.329
mtld_ma_wrap_aw	0.324
basic_ntypes	0.315
Word Count	0.314
nwords	0.313
basic_ncontent_types	0.303
log_ttr_cw	0.295
fkgl	0.290

(d) delta – bleu

Variable	r
hyper_verb_noun_Sav_P1	-0.403
hyper_verb_noun_Sav_Pav	-0.384
hyper_verb_noun_s1_p1	-0.362
hyper_noun_Sav_P1	-0.305
av_pobj_deps_NN	-0.302
log_ttr_cw	-0.293
av_pobj_deps	-0.291
hyper_noun_S1_P1	-0.291
KF_Freq_CW	0.284
COCA_fiction_Freq_CW	0.282

(e) simp – samsa

Variable	r
cl_ndeps_std_dev	-0.309
basic_ntokens	-0.302
Word Count	-0.285
basic_ntypes	-0.284
basic_nfunction_tokens	-0.277
nwords	-0.277
basic_ncontent_tokens	-0.271
mtld_ma_wrap_aw	-0.265
basic_nfunction_types	-0.261
basic_ncontent_types	-0.253

(f) delta – samsa

Variable	r
COCA_news_RL_AW	0.199
fog	-0.198
COCA_news_RL_CW	0.197
basic_ncontent_types	-0.194
arindex	-0.193
basic_ncontent_tokens	-0.192
COCA_spoken_RL_AW	0.188
mtld_ma_wrap_aw	-0.185
fkgl	-0.184
poly_verb	0.184

(g) simp – sari

Variable	r
root_ttr_aw	0.378
Word Count	0.378
nwords	0.376
root_ttr_cw	0.365
basic_ntypes	0.363
mtld_ma_wrap_aw	0.357
basic_ntokens	0.355
basic_ncontent_types	0.337
basic_ncontent_tokens	0.327
mtld_ma_wrap_cw	0.322

(h) delta – sari

Variable	r
nwords	-0.499
Word Count	-0.496
root_ttr_aw	-0.477
basic_ntypes	-0.461
basic_ntokens	-0.459
mtld_ma_wrap_aw	-0.449
basic_nfunction_types	-0.447
basic_nfunction_tokens	-0.442
root_ttr_cw	-0.420
basic_ncontent_tokens	-0.410

Table 5: Top absolute values of significant correlation coefficients ($p < .05$) between human judgments and automatic metrics, on the SimplicityDA dataset.

(a) simp – D-SARI

Variable	r
Word Count	-0.220
MRC_Imageability_FW	0.172
Brysbaert_CC_FW	0.171
MRC_Concreteness_FW	0.164
MRC_Meaningfulness_FW	0.148
COCA_academic_tri_2_DP	-0.146
KF_Freq_CW	0.146
Kuperman_AoA_FW	-0.134
Brysbaert_CC_AW	0.133
eat_tokens	-0.129

(b) delta – D-SARI

Variable	r
WN_Zscore_CW	0.305
COCA_spoken_tri_MI2	-0.303
WN_Zscore	0.298
COCA_spoken_tri_MI	-0.283
COCA_spoken_Trigram_Range_Log	0.266
WN_Mean_RT_CW	0.263
COCA_spoken_tri_2_MI2	-0.258
Ortho_N_CW	-0.254
WN_Mean_RT	0.252
PLD	0.251

(c) simp – LENS

Variable	r
Word Count	-0.347
McD_CD_FW	0.199
Kuperman_AoA_FW	-0.196
lsa_average_all_cosine	0.191
Brown_Freq_CW	0.188
COCA_magazine_Range_Log_AW	-0.186
Brysbaert_CC_FW	0.186
COCA_academic_tri_2_DP	-0.185
OG_N_H_FW	0.185
Freq_N_OGH_FW	0.185

(d) delta – LENS

Variable	r
COCA_spoken_tri_2_MI	-0.279
COCA_spoken_tri_2_MI2	-0.267
LD_Mean_RT_SD	0.247
LD_Mean_RT_SD_CW	0.241
COCA_fiction_Frequency_AW	0.241
COCA_news_Frequency_AW	0.235
COCA_spoken_tri_MI2	-0.234
COCA_magazine_Frequency_AW	0.226
Brown_Freq_CW_Log	-0.221
poly_noun	0.217

(e) simp – BERTScore_F1

Variable	r
WN_Mean_Accuracy_CW	-0.143
WN_Mean_Accuracy	-0.134
COCA_Fiction_Trigram_Range_Log	0.133
LD_Mean_Accuracy_CW	-0.131
COCA_fiction_tri_2_MI	-0.127
COCA_fiction_tri_2_MI2	-0.112
LD_Mean_Accuracy	-0.112
COCA_spoken_Trigram_Range_Log	0.096
LD_Mean_RT_Zscore	0.092
BNC_Written_Trigram_Freq_Normed_Log	0.090

(f) delta – BERTScore_F1

Variable	r
AWL_Sublist_10_Normed	0.219
COCA_academic_tri_T	-0.202
COCA_magazine_tri_2_T	-0.194
COCA_academic_tri_2_T	-0.192
COCA_news_tri_T	-0.188
COCA_magazine_tri_T	-0.186
COCA_news_tri_2_DP	0.186
COCA_news_tri_2_T	-0.184
COCA_Academic_Trigram_Frequency_Log	-0.181
LD_Mean_RT_SD_CW	-0.179

Table 6: Top absolute values of significant correlation coefficients ($p < .05$) between human judgments and automatic metrics, on the D-WIKI dataset.