Comparative Analysis of Evaluation Measures for Scientific Text Simplification

Dennis Davari[1], Liana Ermakova[2(✉)][0000−0002−7598−7474], and
Ralf Krestel[1,3][0000−0002−5036−8589]

[1] Kiel University, Kiel, Germany
stu218009@mail.uni-kiel.de
[2] Université de Bretagne Occidentale, HCTI, Brest, France
liana.ermakova@univ-brest.fr
[3] ZBW – Leibniz Information Centre for Economics, Kiel, Germany
rkr@informatik.uni-kiel.de

Abstract. Access to reliable scientific knowledge is crucial to making informed decisions for both policymakers and citizens. However, scientific documents are inherently difficult due to their complex terminology and vernacular. Automatic text simplification aims to remove some of these barriers. Evaluation frameworks, which include collections and evaluation measures, are designed to assess the generated text simplifications. In this paper, we perform a comparative analysis of current text simplification evaluation measures on both scientific text and a generic corpus based on Wikipedia. Our main finding is that the currently existing measures tend to perform worse on scientific texts and on longer texts consisting of several sentences. More generally, our analysis informs the development of suitable text simplification evaluation measures for scientific texts.

Keywords: Scientific documents · Automatic text simplification· Evaluation measures · Test collections · Scholarly Communication · User study

1 Introduction

Accessing reliable scientific knowledge is essential for informed decision-making among policymakers and the general public. However, scientific documents can be challenging to understand due to their complex terminology and specialized language. The goal of automatic text simplification is to simplify a text to make it easily understandable while maintaining its original meaning [36]. While text simplification in other domains, such as medical [8,22] and legal [15], has been widely explored, little research has been done on text simplification with a focus on scientific texts. Especially when it comes to the evaluation of text simplification methods for scientific texts there is a lack of focused test collections. Further, it is unclear how well-existing evaluation measures developed for other domains are suited for judging the simplification of scientific texts.

In general, there are two ways to assess the quality of generated text simplifications: one can rely on either manual human evaluation or automatic measures and test collections. The main drawback of manual evaluation is its lack of scalability and reproducibility as it is expensive and time-consuming. On the other hand, automatic evaluation frameworks are hard to create and they might be biased toward specific use cases. Hence, on the road towards reliable and reproducible evaluation results, more research on evaluation measures for text simplification is needed [16]. Text simplification measures are typically evaluated based on Pearson correlation [2,41,43,28,32,18,7,23,4] or Spearman's rank correlation [38,35] between the scores the measures generate and human judgment. It is standard practice to assess the correlations with the human ratings along three dimensions [2,41,43,28,32,38,18,7,23,35,4]:

Simplicity: How simple is the simplification?
Meaning preservation: How well was the information preserved?
Fluency: How readable and grammatically correct is the simplification?

While existing studies on text simplification evaluation take these three dimensions into account, e.g., by computing the correlations separately and optionally aggregating them into an overall score, other dimensions, such as the domain of the test collection are often not varied. Thus, the difference in measuring performance across different domains is a research gap we aim to fill. For example, technical vocabulary and abstract concepts in scientific texts often necessitate supplementary explanations to facilitate comprehension for readers when they can not be dropped or replaced. In other domains, replacing a complicated term with a more common term might be enough to simplify a sentence [10]. Besides, most text simplification evaluation frameworks are designed to assess individual sentences, while paragraph- and document-level simplification might follow the strategy to delete entire sentences [6] to simplify a paragraph or document. Thus, we also need to distinguish sentence level from paragraph/document level when assessing the performance of different measures.

The main goal of this paper is to conduct a comparable analysis of text simplification evaluation measures on both scientific text and non-scientific text. Specifically, this paper aims to answer the following two research questions:

1. How do existing text simplification measures perform when applied to scientific texts compared to non-scientific texts?
2. How do existing text simplification measures perform when applied to entire paragraphs rather than the standard sentence-level text simplification?

To answer these questions and assess the quality of evaluation measures, we calculated the correlations between the scores of the measures and the human annotations for the different corpora (scientific/non-scientific, sentence-level/paragraph-level). This is in line with previous work [2], which analyzed the correlations of text simplification measures with human annotations across three dimensions: the perceived simplicity level, the system type, and the set of reference simplifications used. Other work [4] evaluated correlations between measures and human-annotated corpora for meaning preservation. They used simple test cases to achieve either minimum scores (e.g., two unrelated sentences) or

Table 1. Properties of the text simplification measures. Measures marked by * are based on LMs. “Ref-based” indicates whether the measure requires reference texts.

Name		Year	Task	Measure	Ref-based
FKGL	[17]	1975	Readability analysis	Simplicity	No
BLEU	[25]	2002	Machine translation	Closeness to reference	Yes
iBLEU	[34]	2012	Paraphrase generation	Paraphrase quality	Yes
chrF	[26]	2015	Machine translation	Closeness to reference	Yes
SARI	[38]	2016	Text simplification	Overall quality	Yes
FKBLEU	[38]	2016	Text simplification	Overall quality	Yes
BERTScore*	[40]	2019	Text generation	Closeness to reference	Yes
BLEURT*	[29]	2020	Text generation	Closeness to reference	Yes
D-SARI	[35]	2021	Text simplification	Overall quality	Yes
SMART	[3]	2022	Text generation	Closeness to reference	Yes
ISiM	[24]	2022	Text simplification	Simplicity	No
LENS*	[23]	2022	Text simplification	Closeness to reference	Yes
BETS*	[41]	2023	Text simplification	Overall quality	No
SLE*	[7]	2023	Text simplification	Simplicity	No
ΔSLE*	[7]	2023	Text simplification	Simplicity	No
MeaningBERT*	[4]	2023	Text simplification	Meaning preservation	No

maximum scores (e.g., two identical sentences). They combined several corpora with similar characteristics to obtain an overall score. This is rather uncommon and we follow standard practice [2,41,43,28,7,23] and not merge different (similar) corpora but assess the results of every individual corpus.

The rest of the paper is organized as follows. Sections 2 and 3 provide an overview of the different measures and corpora used in this study. Sections 4 and 5 present and interpret the correlations between the measures and corpora and Section 6 summarizes the main findings of this paper.

2 Measures

The measure calculates a score that describes the quality of the simplified text. There are two types of measures which are reference-based and reference-free measures. Reference-based measures require reference simplification, which is a gold standard simplification created by humans, to calculate the score for the simplified text. More references lead to more accurate scores of the measures. Reference-free measures on the other hand don’t require references and are able to produce a score just based on the simplified text only (aka text statistics) or based on the comparison between the original and the simplified texts.

An overview of all measures considered in this study can be seen in Table 1.

The most widely used measures for text simplification are BLEU, FKGL, and SARI. BLEU (Bilingual Evaluation Understudy) [25] originates from machine translation and is based on n-gram matching between the system simplification and reference simplification. FKGL (Flesch-Kincaid Grade Level) [17] is a reference-free measure that scores text simplicity based on sentence length and word syllables, assuming longer sentences and words increase complexity. SARI (System output Against References and against the Input sentence) [38] assesses system simplification by comparing it to both the original text and reference simplifications, evaluating how well the simplification maintans, adds, or removes information.

These three measures were further extended in the literature. For example, iBLEU [34] is based on BLEU but penalizes conservativeness. FKBLEU [38] combines reference-based iBLEU and reference-free FKGL. As FKGL, ISiM (Independent Simplification measure)[24] measures sentence complexity based on its length and word commonness. D-SARI [35] is a modification of SARI for the document level.

New advances in language models (LMs) have led to the development of various transformer-based metrics. BERT (Bidirectional Encoder Representations from Transformers) is the most popular masked LM used in simplification measures [9]. BERTScore [40] computes the cosine similarity between BERT [9] embeddings of the system and reference simplification. SMART (Sentence MAtching for Rating Text) [3] is a document-level measure that uses sentence-unit soft-matching between candidate and reference sentences. SMART offers two variants: SMART-L and SMART-N. SMART-L aggregates scores based on the longest common subsequence of sentences, prioritizing sequence and coherence while SMART-N considers all possible n-grams of sentences to evaluate content coverage and diversity. In this study, we consider only SMART-L with two variants of the matching functions: BLEURT (Bilingual Evaluation Understudy with Representations from Transformers) [29] with an LM-based matching function and chrF (character n-gram F-score) with an n-gram-based matching function [26] as they showed the best results [3]. LENS (Learnable Evaluation measure for Text Simplification) [23] is a reference-based measure that compares different edits, such as splitting, paraphrasing, and deletion, in simplified text using an adaptive ranking loss. SLE (Simplicity Level Estimate) [7] is a RoBERTa-based [21] reference-free measure to assess the simplicity of a sentence. ΔSLE (Δ Simplicity Level Estimate) [7] is a RoBERTa-based measure which calculates the simplicity gain relative to a source sentence. BETS (BERT Embedding-based evaluation for Text Simplification) [41] uses BERT siamese networks to assess the simplicity gain and meaning preservation without references. It compares the original and simplified sentences using contextual embeddings from BERT to evaluate their semantic similarity and the simplicity of the simplified text. Unfortunately, we couldn't recreate the reported simplicity results and therefore we only included the meaning preservation component in this study. MeaningBERT [4] assesses meaning preservation based on the BERT model trained on ASSET [1], Simplicity-DA [2], SIMPEVAL$_{2022}$ [23] and QuestEvalCorpus [28].

3 Corpora

Table 2 shows an overview of the properties of the corpora included in this study.

Table 2. Properties of the text simplification corpora. Paragraph-level corpora are in bold. Scientific corpora are underlined. # References=# of sentences/paragraphs × # of references per sentence/paragraph; # Annotations=# of sentences/paragraphs × # of simplifications per sentence/paragraph × # of annotations per simplification. "Overall simplicity" denotes a single score over all dimensions (fluency, meaning preservation, simplicity), while "Simplicity" refers specifically to the simplicity dimension only. The numbers for the annotation dimensions indicate the scale's granularity (i.e. 3 represents a 3-point Likert scale whereas 101 represents a score between 0 and 100).

				Annotation dimensions					
Name		#References	#Annotations	Fluency	Meaning preservation	Simplicity	Lexical simplicity	Syntactic simplicity	Overall simplicity
ASSET	[1]	$2,359 \times 10$	$100 \times 1 \times 15$	101	101	101			
D-Wikipedia	[35]	100×1	$100 \times 6 \times 3$	5	5		5	5	5
Medical-NE	[22]	$4,459 \times 1$	$200 \times 2 \times 3$	3	3	3			
Medical-EX	[22]	$4,459 \times 1$	$49 \times$ 1-2 $\times 2$		3				
QuestEvalCorpus	[28]	—	$100 \times$ 2-4 $\times 23.5$	101	101	101			
SIMPEVAL$_{2022}$	[23]	60×2	$60 \times 6 \times 3$	101	101	101			101
Simplicity-DA	[2]	—	$100 \times 6 \times 15$	101	101	101			
Simple-2018	[33]	—	$70 \times 28 \times 3$	5	5	5		5	
SimpleTextSent	[13]	359	2,048		11		11	11	
SimpleTextPara	[13]	53	271		11		11	11	

The corpora are different in terms of simplification units (sentence-, paragraph- or document-level), the origin of the source texts (Wikipedia, news articles, scientific abstracts etc.), and simplification methods (manual paraphrasing, automatic alignment from different sources, generated simplifications). In this paper, we do not distinguish paragraph- and document-level simplifications as the text size in many existing corpora is comparable for both categories. We contrast paragraph- and document-level simplifications to sentence-level simplifications. A large number of simplification corpora are based on the alignment of *sentences* from English Wikipedia[4] and Simple Wikipedia[5]. Thus, ASSET (Abstractive Sentence Simplification Evaluation and Tuning) [1], QuestEvalCorpus [28], Simplicitiy-DA [2] and Simple-2018 [33] are all based on the test set of TurkCorpus [38], which contains multiple crowd-sourced paraphrase-based reference simplifications of sentences from the Parallel Wikipedia Simplification (PWKP) corpus [42]. QuestEvalCorpus and ASSET are based on the same 100 unsimplified sentences with the difference that the ASSET simplifications were generated by simplification systems while QuestEvalCorpus consists of human simplifications [28]. However, in contrast to the description in [28], the actual QuestEval-

[4] https://www.wikipedia.org/

[5] https://simple.wikipedia.org/

Corpus also contains annotations for system simplifications. The aformentioned TurkCorpus-based corpora mostly employ more traditional simplification systems not based on LMs such as PBMT-R [37], SBMT-SARI [38] or HYBRID [30]. For all of these TurkCorpus-based corpora, the ASSET references were used when calculating the scores of the measures because these are the highest quality TurkCorpus references [1]. SIMPEVAL$_{2022}$ [23] is based on recent and more complex Wikipedia sentences and employs modern simplification systems such as GPT-3 [5] and T5 [27]. There are three different variants of this corpus all using different human rating methods. We used the variant which was rated using the Direct Assessment methodology [23]. Few document- and paragraph-level simplification corpora exist. For example, D-Wikipedia consists of document-level alignments between Wikipedia and Simple Wikipedia articles [35]. However, as opposed to the description in [35], there are actually six different simplifications per paragraph. Transformer-based simplification systems such as BART [19] or BERTSUMEXTABS [20] were employed. The existence of one reference simplification per paragraph allowed us to run reference-based measures on this corpus. The Medical Corpus contains annotations of plain language summaries of medical abstracts [8]. For 200 paragraphs there are two simplifications where every simplification is rated by three non-expert annotators (Medical-NE). Moreover, for 49 paragraphs there are one to two simplifications per paragraph and every simplification is rated by two expert annotators (Medical-EX) [22]. The two simplification systems employed (UL-BART [8] and NapSS [22]) are modern transformer-based systems that were created for the medical domain.

The CLEF (Conference and Labs of the Evaluation Forum) SimpleText 2023 track consisted of several shared tasks which all aim to facilitate the comprehension of scientific texts for laypersons [13,14]. In this paper, we used the source texts, references, and automatic simplifications of the participants of the shared tasks. More specifically, we created two corpora: the sentence-level SimpleTextSent corpus and the abstract-level SimpleTextPara corpus. In contrast to the majority of the simplification corpora, these two corpora are direct simplifications of source texts and not automatically aligned sentences. The source sentences of the SimpleText corpus were taken from the abstracts of scientific papers from the DBLP Citation Network Dataset for *Computer Science* and Google Scholar and PubMed articles on *Health and Medicine*. The reference simplifications were manually simplified by either master's students studying Technical Writing and Translation or by a collaboration between a domain expert (a computer scientist) and a professional translator. The system simplifications are submissions from CLEF 2023 participants which are mostly outputs from LM-based systems [13]. The annotations to the simplifications were performed by the linguists and the master students in translation and technical writing from the University of Brest (Université de Bretagne Occidentale). The SimpleTextPara corpus concatenates the sentences and corresponding ratings from the SimpleTextSent corpus to form paragraphs. Note that the annotators rated the information distortion severity instead of meaning preservation. However, to have better comparability with the other corpora, these ratings were treated as ratings for the meaning preservation dimension. This is reasonable because a high (or low) meaning preservation is characterized by a low (or high) information distortion severity.

4 Results

To evaluate the different measures, we computed the correlation of measures with human judgments. The correlations between the measures and humans were only analyzed for the dimensions of simplicity and meaning preservation as the state-of-the-art LLM systems do not suffer from the lack of fluency in the generated content. There are ratings for the lexical simplicity and syntactic simplicity for D-Wikipedia and all SimpleText corpora. In order to have a single simplicity value the mean of both values was calculated for each of these corpora. For better comparability, we report negative FKGL scores. This way, for every measure a higher score means more simplicity or better meaning preservation.

4.1 Simplicity

Table 3 shows the correlations for all measures for simplicity ratings. In general, the correlation between automatic measures and human scores of simplicity is low, especially for scientific corpora.

Several measures correlate with human judgments on QuestEvalCorpus, ASSET, Simplicity-DA, and D-Wikipedia while showing negative correlations with SIMPEVAL$_{2022}$ and Simple-2018. This could be observed for all BERTScore variants with significant correlations for the F-score and recall score as well as for BLEU, iBLEU, and all variants of SMART-L chrF. Furthermore, this pattern could also be observed for all variants of SMART-L BLEURT although with the p-values not being significant for the aforementioned corpora. What makes these findings interesting is that these corpora, which are mostly non-scientific and at the sentence level, are both counted among the corpora with the best and worst correlations. As for the scientific corpora, SimpleTextSent and SimpleTextPara stand in the middle between the corpora with the highest and lowest correlations with only weak and significant or insignificant correlations.

For every measure, there are either corpora with negative correlations with the measure or with insignificant correlations. The exceptions to this rule are LENS and SARI. Both measures show significant and positive correlations with all corpora ranging from 0.11 to 0.77 for LENS and 0.17 to 0.55 for SARI. The correlations are above 0.7 for LENS for QuestEvalCorpus, Simplicity-DA, and ASSET. Overall, what can be seen is that LENS tends to perform better on non-scientific corpora at the sentence level and worse at the scientific context and paragraph level respectively. The findings for SARI indicate that this measure is more robust and less domain-dependent than LENS. Even though this measure is biased towards non-scientific corpora it also holds up fairly well for scientific texts. Moreover, it can be observed that this measure doesn't discriminate between the sentence and paragraph level.

Table 3. Pearson correlation values of the simplicity ratings; values in italics have a p-value of less than 0.05; The highest value for every corpus is underlined; If a score has different components, they are marked as follows: F=F-score, P=precision, R=recall, MP=meaning preservation. Paragraph-level corpora are marked in bold. Scientific corpora are underlined. LM-based measures marked by *.

Measure	ASSET	QuestEvalCorpus	SIMPEVAL$_{2022}$	SimpleTextSent	Simplicity-DA	Simple-2018	**D-Wikipedia**	**Medical-NE**	**SimpleTextPara**
BERTScore F*	.56	.68	-.20	*-.01*	.56	-.11	.29	*-.00*	*.07*
BERTScore P*	.63	.69	*-.05*	*.03*	.61	*.01*	.49	.25	*.10*
BERTScore R*	.43	.62	-.33	*-.05*	.48	-.21	*-.00*	-.21	*.01*
BETS MP*	.32	.14	-.59	-.33	.39	-.27	-.56	-.44	-.20
BLEU	.36	.50	-.29	*-.05*	.49	-.10	.17	-.24	*-.03*
D-SARI	.37	.36	.29	.12	.24	.34	.51	.22	.14
D-SARI (add)	*.17*	.48	.15	.10	.31	.05	.22	*-.10*	.12
D-SARI (delete)	.26	.26	.33	.13	.11	.36	.50	.36	.14
D-SARI (keep)	.37	.15	.15	*.04*	.26	.16	.40	*-.06*	*.09*
FKBLEU	.29	.42	-.13	*.00*	.43	.14	.16	-.22	*.09*
FKGL	*-.09*	*.03*	.39	.31	-.11	.08	.35	*.09*	*-.10*
iBLEU	.36	.50	-.29	*-.05*	.49	-.10	.17	-.24	*-.03*
ISiM	*-.12*	*.04*	.35	.45	*-.01*	.13	.42	.16	*.07*
LENS*	.73	.77	.35	.19	.75	.11	.47	.28	.19
MeaningBERT*	.51	.26	-.17	-.33	.63	-.29	-.50	-.24	.13
SARI	.25	.50	.36	.23	.21	.21	.55	.25	.17
SLE*	*-.11*	.16	.41	.34	*-.03*	.11	.47	*-.09*	*.02*
ΔSLE*	*.06*	.23	.47	.21	*-.01*	.31	.32	*-.12*	*.05*
SMART-L BLEURT F*	.57	.70	*-.07*	*.00*	.63	-.12	.28	*.04*	*.03*
SMART-L BLEURT P*	.55	.69	*-.04*	*-.00*	.70	*-.04*	.40	.17	*.04*
SMART-L BLEURT R*	.58	.71	*-.07*	*.01*	.59	-.18	.10	*-.09*	*.02*
SMART-L chrF F	.33	.58	-.28	-.10	.43	-.17	.18	*.02*	*.06*
SMART-L chrF P	.32	.56	-.25	-.10	.46	-.05	.29	*.10*	*.08*
SMART-L chrF R	.32	.57	-.28	-.10	.38	-.25	*.04*	*-.08*	*.03*

4.2 Meaning Preservation

Table 4 shows the correlations of all measures with the meaning preservation ratings.

Starting with BLEU, it can be seen that it has the highest correlations with all corpora that are based on TurkCorpus, which are ASSET, QuestEvalCorpus, Simplicity-DA, and Simple-2018. The correlations range from 0.44 to 0.62 for these corpora and they are significant. Overall, FKBLEU behaves similarly but has lower correlations with the corpora than BLEU. It can be seen for both the

Table 4. Pearson correlation values of the meaning preservation ratings; values in italics have a p-value of less than 0.05; The highest value for every corpus is underlined; If a score has different components, they are marked as follows: F=F-score, P=precision, R=recall, MP=meaning preservation. Paragraph-level corpora are marked in bold. Scientific corpora are underlined. LM-based measures marked by *.

Measure	ASSET	QuestEvalCorpus	SIMPEVAL$_{2022}$	SimpleTextSent	Simplicity-DA	Simple-2018	**D-Wikipedia**	**Medical-EX**	**Medical-NE**	**SimpleTextPara**
BERTScore F*	.82	.80	.27	.50	.78	.57	.15	.28	.14	.50
BERTScore P*	.77	.77	*.08*	.34	.74	.44	*-.04*	*.10*	*-.05*	.38
BERTScore R*	.77	.79	.38	.59	.76	.60	.30	.36	.27	.53
BETS MP*	.76	.59	.33	.47	.72	.58	.82	*.26*	.47	.20
BLEU	.60	.55	*.04*	.26	.61	.44	.12	*.04*	.20	.17
D-SARI	*.06*	*.11*	*-.07*	*.04*	*-.00*	-.30	-.29	*-.14*	-.21	.18
D-SARI (add)	.22	.45	*.08*	.20	.36	*.02*	*-.11*	*-.15*	*-.08*	.32
D-SARI (delete)	*-.12*	*-.07*	-.13	*-.03*	-.17	-.44	-.43	*-.10*	-.27	.18
D-SARI (keep)	.22	*.07*	*-.02*	.06	.12	*.02*	*-.05*	*-.08*	*.00*	*.11*
FKBLEU	.36	.31	*.02*	.17	.40	.10	.17	*-.15*	*-.02*	*.10*
FKGL	-.29	-.17	*-.08*	-.28	-.25	-.06	-.10	-.40	-.24	-.25
iBLEU	.60	.55	*.04*	.26	.61	.44	.12	*.04*	.20	.17
ISiM	-.30	*-.14*	*-.08*	-.24	-.20	-.13	-.35	-.31	-.32	-.19
LENS*	.66	.74	*-.02*	.24	.67	.32	-.12	*-.22*	-.36	.44
MeaningBERT*	.71	.46	*.09*	.38	.87	.44	.75	.28	.46	.17
SARI	*.15*	.32	*-.00*	*-.02*	.18	-.24	-.32	*-.22*	-.37	*.09*
SLE*	-.36	*-.11*	-.23	-.30	-.18	-.08	-.29	-.37	-.24	-.14
ΔSLE*	-.31	-.16	-.25	-.34	-.31	-.42	-.37	*-.08*	-.19	*-.10*
SMART-L BLEURT F*	.83	.80	.13	.56	.81	.56	*.08*	*.22*	.16	.38
SMART-L BLEURT P*	.81	.78	*.04*	.50	.77	.49	*-.04*	*.18*	*.05*	.33
SMART-L BLEURT R*	.84	.81	.20	.56	.80	.58	.17	*.24*	.22	.33
SMART-L chrF F	.68	.68	*.10*	.43	.67	.52	.14	*.19*	.27	.31
SMART-L chrF P	.65	.65	*.04*	.40	.63	.43	*.04*	*.22*	.19	.31
SMART-L chrF R	.67	.68	.14	.43	.66	.54	.21	*.15*	.29	.26

simplicity and meaning preservation dimension that the iBLEU correlations are identical to the BLEU correlations. Concerning SARI and D-SARI it can be seen that there are only either statistically insignificant positive or significant negative correlations with the corpora. The exceptions to this rule are only the correlations with QuestEvalCorpus (0.32) and Simplicity-DA (0.18) for SARI which are positive and significant. Just like SARI and LENS for the simplicity dimension, there is also a group of measures that almost only have significant and positive correlations with the corpora. This group consists of BERTScore, the meaning preservation component from BETS, MeaningBERT, SMART-L BLEURT (recall), and SMART-L chrF (recall).

In general, automatic measures correlate better with human judgments for meaning preservation than for simplicity, although this correlation remains low, especially for scientific corpora and a document level.

5 Discussion

5.1 Interpretation of Results

All in all, the findings suggest that the measures perform better for non-scientific texts and the sentence level respectively. One trend that can be observed is that measures based on LMs show higher correlations as well as in the case of both SMART-L variants. It can be seen that SMART-L BLEURT performs consistently better than SMART-L chrF. However, overall, the drops in correlations are higher for LM-based measures than for non-LM-based measures when being applied to scientific corpora and the paragraph level respectively. These findings are not surprising given the nature of LMs. It seems likely that LM-based measures perform worse at the scientific context and paragraph level because they have mostly been trained at the non-scientific context for the sentence level. Thus, in order to have an LM-based measure that performs well at the scientific context and paragraph level we would either need to train the currently existing LM-based systems on such data or create new models which have specifically been trained on such data.

Notably, SARI and D-SARI negatively correlate with many corpora for the meaning preservation dimension. Similarly, BLEU shows significant negative correlations with some corpora for the simplicity dimension. These findings confirm [31] that SARI is better for evaluating simplicity, while BLEU is more suitable for assessing meaning preservation. Moreover, the correlations with iBLEU are similar to those with BLEU, suggesting that iBLEU is not more suitable than BLEU for text simplification.

In general, there is a low correlation between automatic measures and human scores of simplicity, especially for scientific texts which implies that automatic measures do not fully capture this aspect. Traditional simplification methods focus on removing complex terms and structures to improve readability, but providing term definitions and background knowledge could enhance accessibility and comprehension of scientific texts [10]. Thus, the introduction of background knowledge might be a key factor in scientific text simplification, distinguishing it from general text simplification. Text statistics, like FKGL, based on word and sentence length showed low correlation, meaning that simplicity is not limited to these parameters. However, more research is needed to confirm this hypothesis.

Measures correlate better with human judgments for meaning preservation than for simplicity, though this correlation remains low, particularly for scientific corpora and at the document level. The low correlation between human-assigned meaning preservation scores and automatic measures suggests that high-scoring simplifications may still suffer from information distortion and spurious content. The prevalence of spurious content may be introduced gratuitously by the generative model and is informally referred to as "hallucinations." This finding confirms the results of the previous research attempting to quantify the prevalence of spurious content in text simplification both on sentence and paragraph levels [12]. In the generated output realigned with the original source sentences, entire output sentences that do not share a single token with the input are variable but remain notably frequent. Their results indicate that simplifications with significant spurious content can still achieve high text overlap with references, leading to very high-performance scores according to traditional automatic measures. Recent shared tasks provide evidence that simplification is primarily achieved using generative or foundational models which can suffer from "hallucinations" [12,11]. Thus, new simplification measures should more effectively account for the potential risk of hallucination, aka better considering meaning preservation. This is especially challenging for scientific texts, as introducing background knowledge, crucial for simplification, might be considered a "hallucination" if not present in the source text.

Another interesting finding is that for the simplicity dimension, non-LM-based measures appear unaffected by text length, showing similar correlations at both the sentence and paragraph levels. FKGL is unaffected by the input length since FKGL was designed to be used for whole documents and not only for sentences. As the source code of ISiM was modified to be able to be applied to multiple sentences, by averaging the scores of the individual sentences, it is also not surprising that the input length doesn't affect this measure's performance. D-SARI is also robust to the length of the input which is also not surprising given the fact that this measure was designed for assessing whole documents. SARI also seems to be robust to the length of the input. Overall, it can be seen that reference-free measures perform relatively better at the paragraph-level corpora than reference-based measures. This seems reasonable as longer references lead to more potential deviations between the simplification and reference.

Lastly, it can be observed that many reference-based measures show high correlations with TurkCorpus-based corpora (i.e. ASSET, Simplicity-DA, simplification-acl 2018 and QuestEvalCorpus). One explanation for this might be the existence of multiple high-quality reference simplifications leading to higher correlations for reference-based measures. However, this would not explain why LENS and SMART-L BLEURT perform considerably worse on the Simple-2018 corpus. This corpus differs the most from the other TurkCorpus-based corpora considering the simplification systems used as the systems used in Simple-2018 vastly differ from the systems used in the other TurkCorpus-based corpora. A possible reason why LENS and SMART-L BLEURT perform worse on the Simple-2018 corpus is that these systems are biased towards certain simplification systems. This assumption is supported by the findings in [3] which suggest that measures are biased concerning the different systems.

Summing up, a clear tendency of measures performing better for the non-scientific texts and sentence level could be observed. The reasons for this behavior are diverse but the main reason is that most measures were designed for the use case of non-scientific texts at the sentence level. That implies that different benchmarks are necessary for evaluating the performance of text simplification methods on scientific texts compared to non-scientific ones.

5.2 Limitations

SimpleTextSent annotations often had single raters per sentence, and no inter-annotator agreement analysis was conducted for multiple raters. In SimpleTextPara, paragraphs were formed by concatenating sentences from SimpleTextSent. This approach reduces the distinctiveness between the corpora and maintains the same sentence count in the simplified and the source paragraphs.

A limitation of LM-based measures is that the input length sometimes exceeds the maximum length supported by the language model. This puts LM-based models at a disadvantage compared to non-LM-based systems for scoring longer inputs, such as paragraph-level corpora. It was not explored how many paragraphs were affected by this problem.

Another limitation is the varying amount of references per simplification. While there is only one reference simplification for corpora such as D-Wikipedia or the Medical Corpus there are ten reference simplifications for all corpora which use TurkCorpus sentences. This skews the results by leading to higher correlations of reference-based measures with corpora with multiple references.

Lastly, the type of system simplification used influences the ratings of automatic measures as can be seen in the field of text summarization where the BARTScore [39] is biased towards outputs from the BART model [3]. This might skew the results but was not taken into consideration for this study.

6 Conclusion

Digital libraries aim to make scholarly information accessible to users with diverse backgrounds and expertise levels. While text simplification can improve understanding for non-experts, selecting appropriate metrics is crucial to ensure that simplification enhances accessibility without compromising accuracy or completeness. To determine how the scientific context affects simplification measures' effectiveness, we applied various measures to different human-annotated corpora and calculated the correlations between these measures and human ratings. Due to the limited availability of scientific corpora, two scientific corpora were created which are all based on sentences and human ratings from the SimpleText project. These are SimpleTextSent and SimpleTextPara for the sentence and paragraph level respectively. Access information for these corpora can be found on the SimpleText website[6].

The analysis of correlations between measures and various corpora for simplicity and meaning preservation reveals that measures perform better in non-scientific contexts and at the sentence level. This highlights a key issue in NLP:

[6] http://simpletext-project.com

the tendency to prioritize generic metrics over domain-specific optimization, emphasizing the need for specialized benchmarks. While state-of-the-art LLMs, including instruction-tuned autoregressive models like ChatGPT and Gemini, can handle inputs longer than a single sentence, evaluating their ability to simplify longer scientific texts is challenging due to the lack of test collections specifically designed for scientific texts at the paragraph and document levels.

Automatic measures poorly correlate with human scores of simplicity, especially in scientific texts, meaning that capturing this aspect is still challenging.

Measures correlate better for meaning preservation than simplicity, but this correlation remains low. The lack of correlation confirms previous findings that standard evaluation measures fail to detect a wide range of spurious text generation, aka "hallucinations", highlighting the need for more research in this area [12]. This is particularly challenging for scientific texts since adding essential background knowledge for simplification might be seen as a "hallucination" if it's not in the original text.

LM-based measures tend to be more sensitive to the context and length of the simplifications. Moreover, reference-free measures tend to be less sensitive to the input length than reference-based measures. Lastly, it could be observed that some measures perform especially well on the popular TurkCorpus-based corpora indicating the problem that measures tend to overfit these corpora. In general, our analysis informs the application and the development of appropriate evaluation measures for simplifying scientific texts.

Acknowledgments

This research was funded in part by the French National Research Agency (ANR) under the project ANR-22-CE23-0019-01. We would like to thank our colleagues and the students from the University of Brest (France) who participated in data construction and evaluation.

References

1. Alva-Manchego, F., Martin, L., Bordes, A., Scarton, C., Sagot, B., Specia, L.: 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. pp. 4668–4679. ACL (2020). https://doi.org/10.18653/v1/2020.acl-main.424
2. Alva-Manchego, F., Scarton, C., Specia, L.: The (un)suitability of automatic evaluation metrics for text simplification. Computational Linguistics **47**(4), 861–889 (Dec 2021). https://doi.org/10.1162/coli_a_00418
3. Amplayo, R.K., Liu, P.J., Zhao, Y., Narayan, S.: SMART: sentences as basic units for text evaluation. In: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023. OpenReview.net (2023)
4. Beauchemin, D., Saggion, H., Khoury, R.: Meaningbert: assessing meaning preservation between sentences. Frontiers in Artificial Intelligence **6** (2023). https://doi.org/10.3389/frai.2023.1223924

5. Brown, T.B., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems. pp. 1877–1901 (2020)
6. Cripwell, L., Legrand, J., Gardent, C.: Document-level planning for text simplification. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 993–1006. ACL (2023). https://doi.org/10.18653/v1/2023.eacl-main.70
7. Cripwell, L., Legrand, J., Gardent, C.: Simplicity level estimate (SLE): A learned reference-less metric for sentence simplification. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023. pp. 12053–12059. ACL (2023). https://doi.org/10.18653/V1/2023.EMNLP-MAIN.739
8. Devaraj, A., Marshall, I., Wallace, B., Li, J.J.: Paragraph-level simplification of medical texts. In: Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4972–4984. ACL (2021). https://doi.org/10.18653/v1/2021.naacl-main.395
9. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019)
10. Di Nunzio, G.M., Vezzani, F., Bonato, V., Azarbonyad, H., , Kamps, J., Ermakova, L.: Overview of the CLEF 2024 SimpleText task 2: Identify and explain difficult concepts. In: Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2024)
11. Ermakova, L., Bertin, S., McCombie, H., Kamps, J.: Overview of the CLEF 2023 SimpleText Task 3: Simplification of Scientific Texts. In: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023. CEUR Workshop Proceedings, vol. 3497, pp. 2855–2875. CEUR-WS.org (2023)
12. Ermakova, L., Laimé, V., McCombie, H., Kamps, J.: Overview of the CLEF 2024 SimpleText task 3: Simplify scientific text. In: Working Notes of CLEF 2024: Conference and Labs of the Evaluation Forum. CEUR Workshop Proceedings, CEUR-WS.org (2024)
13. Ermakova, L., SanJuan, E., Huet, S., Azarbonyad, H., Augereau, O., Kamps, J.: Overview of the CLEF 2023 SimpleText Lab: Automatic Simplification of Scientific Texts. In: Experimental IR Meets Multilinguality, Multimodality, and Interaction. pp. 482–506. Springer Nature Switzerland, Cham (2023)
14. Ermakova, L., et al.: CLEF 2024 SimpleText Track: Improving Access to Scientific Texts for Everyone. In: Advances in Information Retrieval: 46th European Conference on Information Retrieval. pp. 28–35 (2024). https://doi.org/10.1007/978-3-031-56072-9_4
15. Garimella, A., Sancheti, A., Aggarwal, V., Ganesh, A., Chhaya, N., Kambhatla, N.: Text simplification for legal domain: Insights and challenges. In: Proceedings of the Natural Legal Language Processing Workshop 2022. pp. 296–304. ACL (2022). https://doi.org/10.18653/v1/2022.nllp-1.28
16. Grabar, N., Saggion, H.: Evaluation of automatic text simplification: Where are we now, where should we go from here. In: Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale. pp. 453–463. ATALA (2022)
17. Kincaid, J., Fishburne Jr, R., Rogers, R., Chissom, B.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)

18. Kriz, R., Apidianaki, M., Callison-Burch, C.: Simple-qe: Better automatic quality estimation for text simplification. CoRR **abs/2012.12382** (2020)
19. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (2019)
20. Liu, Y., Lapata, M.: Text summarization with pretrained encoders (2019)
21. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized BERT pretraining approach. CoRR **abs/1907.11692** (2019)
22. Lu, J., Li, J., Wallace, B., He, Y., Pergola, G.: NapSS: Paragraph-level medical text simplification via narrative prompting and sentence-matching summarization. In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 1079–1091. ACL (2023). https://doi.org/10.18653/v1/2023.findings-eacl.80
23. Maddela, M., Dou, Y., Heineman, D., Xu, W.: LENS: A learnable evaluation metric for text simplification. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023. pp. 16383–16408. ACL (2023). https://doi.org/10.18653/V1/2023.ACL-LONG.905
24. Mucida, L., Oliveira, A., Possi, M.: Language-independent metric for measuring text simplification that does not require a parallel corpus. In: The International FLAIRS Conference Proceedings. pp. 1–4 (2022). https://doi.org/10.32473/flairs.v35i.130608
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. ACL (2002). https://doi.org/10.3115/1073083.1073135
26. Popović, M.: chrF: character n-gram F-score for automatic MT evaluation. In: Proceedings of the Tenth Workshop on Statistical Machine Translation. pp. 392–395. ACL (2015). https://doi.org/10.18653/v1/W15-3049
27. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**, 140:1–140:67 (2020)
28. Scialom, T., Martin, L., Staiano, J., de la Clergerie, É.V., Sagot, B.: Rethinking automatic evaluation in sentence simplification. CoRR **abs/2104.07560** (2021)
29. Sellam, T., Das, D., Parikh, A.: Bleurt: Learning robust metrics for text generation. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7881–7892. ACL (2020)
30. Siddharthan, A., Mandya, A.: Hybrid text simplification using synchronous dependency grammars with hand-written and automatically harvested rules. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics. pp. 722–731. ACL (2014). https://doi.org/10.3115/v1/E14-1076
31. Sulem, E., Abend, O., Rappoport, A.: BLEU is not suitable for the evaluation of text simplification. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. pp. 738–744. ACL (2018). https://doi.org/10.18653/v1/D18-1081
32. Sulem, E., Abend, O., Rappoport, A.: 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). pp. 685–696. ACL (2018). https://doi.org/10.18653/v1/N18-1063

33. Sulem, E., Abend, O., Rappoport, A.: Simple and effective text simplification using semantic and neural methods. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 162–173. ACL (2018). https://doi.org/10.18653/v1/P18-1016
34. Sun, H., Zhou, M.: Joint learning of a dual SMT system for paraphrase generation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). pp. 38–42. ACL (2012)
35. Sun, R., Jin, H., Wan, X.: Document-level text simplification: Dataset, criteria and baseline. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 7997–8013. ACL (2021). https://doi.org/10.18653/v1/2021.emnlp-main.630
36. Wan, X.: Automatic Text Simplification. Computational Linguistics **44**(4), 659–661 (12 2018). https://doi.org/10.1162/coli_r_00332
37. Wubben, S., van den Bosch, A., Krahmer, E.: Sentence simplification by monolingual machine translation. In: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 1015–1024. ACL (2012)
38. Xu, W., Napoles, C., Pavlick, E., Chen, Q., Callison-Burch, C.: Optimizing statistical machine translation for text simplification. Transactions of the Association for Computational Linguistics **4**, 401–415 (2016). https://doi.org/10.1162/tacl_a_00107
39. Yuan, W., Neubig, G., Liu, P.: Bartscore: Evaluating generated text as text generation. In: Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual. pp. 27263–27277 (2021)
40. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: Bertscore: Evaluating text generation with BERT. In: 8th International Conference on Learning Representations. OpenReview.net (2020)
41. Zhao, X., Durmus, E., Yeung, D.Y.: Towards reference-free text simplification evaluation with a BERT Siamese network architecture. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 13250–13264. ACL (2023). https://doi.org/10.18653/v1/2023.findings-acl.838
42. Zhu, Z., Bernhard, D., Gurevych, I.: A monolingual tree-based translation model for sentence simplification. In: Proceedings of the 23rd International Conference on Computational Linguistics. pp. 1353–1361 (2010)
43. Zuo, T., Yosinaga, N.: Leveraging word representation for text simplification evaluation. In: Proceedings of Forum on Data Engineering and Information Management (2021)