EASSE-DE & EASSE-multi: Easier Automatic Sentence Simplification Evaluation for German & Multiple Languages

Regina Stodden
Department of Computational Linguistics
Faculty of Arts and Humanities
Heinrich Heine University Düsseldorf, Germany
regina.stodden@hhu.de

Abstract

In this work, we propose EASSE-multi, a framework for easier automatic sentence evaluation for languages other than English. Compared to the original EASSE framework, EASSE-multi does not focus only on English. It contains tokenizers and versions of text simplification evaluation metrics which are suitable for multiple languages. In this paper, we exemplify the usage of EASSE-multi for German TS resulting in EASSE-DE. Further, we compare text simplification results when evaluating with different language or tokenization settings of the metrics. Based on this, we formulate recommendations on how to make the evaluation of (German) TS models more transparent and better comparable. Additionally, we present a benchmark on German TS evaluated with EASSE-DE and make its resources (i.e., test sets, system outputs, and evaluation reports) available. The code of EASSE-multi and its German specialisation (EASSE-DE) can be found at https://github.com/rstodden/easse-multi and https://github.com/rstodden/easse-de.

1 Introduction

Automatic text simplification (TS) is a natural language processing (NLP) task that involves the development of algorithms and models to automatically transform complex textual content into more straightforward and accessible language. Manual or automatic evaluation is required to measure the quality of the generated simplifications. A good simplification should be grammatically correct, more simple and better readable than the original text and preserve the original meaning of it. For manual evaluation, people are asked to rate the extent of these three aspects for the generated simplification with respect to the original sentence. Because manual evaluation is very time-consuming, automatic metrics are used for a first quality check of sentence simplification models (Alva-Manchego et al., 2020). Compared to manual evaluation methods, automatic evaluation methods facilitate a quick assessment the output of various text simplification models, making it feasible to compare and iterate on different approaches efficiently. Further, with the increasing mass of evaluation data of different model approaches, it becomes challenging to evaluate this large number of generated texts manually. Automatic evaluation methods allow researchers to scale up their assessments to handle large datasets effectively (Alva-Manchego et al., 2020).

Alva-Manchego et al. (2019) proposed an evaluation framework for easier automatic sentence simplification evaluation, called EASSE, to facilitate a comparison of TS models on existing test sets and on the same evaluation metrics as well as to unify the implementation of the evaluation metrics. EASSE is nowadays the common standard for evaluating English TS models. Although it is specified for only English TS evaluation, it is often also used to evaluate TS models of other languages, e.g., German (see, e.g., Trienes et al. 2022), Spanish (see, e.g., Gonzalez-Dios et al. 2022), French (see, e.g., Cardon and Grabar 2020), Swedish (see, e.g., Holmer and Rennes 2023) or on a multi-lingual benchmark (Ryan et al., 2023). However, using EASSE on non-English texts raises some problems, e.g., the tokenizer is not adapted to the language of interest, the BERT-Score is evaluated on an English-only BERT model, and the readability scores are only designed for English.

In this paper, we present EASSE-multi, an adaptation of EASSE for languages other than English (i.e., more than 70 languages, these that are supported by SpaCy), to make the evaluation of non-English TS easier and more robust. We exemplify its usage for one language with several TS resources, i.e., German and the German EASSE variant, EASSE-DE. We further analyze the effects of different settings in EASSE-DE on TS metrics when evaluating German texts and presenting a German TS benchmark build with EASSE-DE.

2 Related Work

2.1 Automatic Evaluation

In order to automatically evaluate text simplification, SARI (Xu et al., 2016) is the primary metric to measure the overall simplicity quality. In more detail, SARI compares a generated simplification sentence with the source sentence and several references to estimate the quality of the lexical simplification. Further, most often BLEU (Papineni et al., 2002) and BERT-Score-Precision (Zhang* et al., 2020) are utilized to measure the similarity or meaning preservation between the original text and the system-generated simplification. Following Alva-Manchego et al. (2021), BERT-Score-Precision can also

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 107–116

measure overall simplicity even if not implemented for this use case. Recently, the LENS score (Maddela et al., 2023) has been proposed to measure the overall simplification quality of English simplifications; it is a trainable score trained on human assessments and English complex-simple pairs. However, human assessments are often missing for TS system outputs in other languages, hence, it is difficult to reproduce for other languages.

Readability formulas such as, FRE or FKGL (Flesch, 1948), are also often used to estimate the readability of the system output (Alva-Manchego et al., 2021). For a syntactical simplification evaluation, SAMSA (Sulem et al., 2018b) has been proposed: SAMSA is a reference-less metric based on the annotation of semantic structures.

The reliability of these metrics for English TS evaluation has been questioned in research, e.g., see Sulem et al. (2018a), Tanprasert and Kauchak (2021), or Alva-Manchego et al. (2021). Another issue with automatic metrics is that the reliability of the scores has only been evaluated against human annotations of English annotations and that the correlations are not yet reproduced or repeated in other languages. Therefore, the suitability of the scores is unclear for other languages than English. Stodden and Kallmeyer (2020) have indeed shown that the way how English sentences are simplified differs from the German or Spanish ways.

Hence, different simplification metrics might be required per language. An approach in this direction could be learnable metrics (per language) as LENS (Maddela et al., 2023), BETS (Zhao et al., 2023) or MeaningBERT (Beauchemin et al., 2023), which are currently only applied to English texts. But, as long as SARI, BLEU, and BERT-Score are still common practices in TS research, we will use them in our analysis but we are also open to replacing or extending the metrics in our evauluation framework, if available.

2.2 Original EASSE Package

The original EASSE package (Alva-Manchego et al., 2019) is designed to ease the automatic evaluation of English sentence simplification. It contains the implementation of automatic evaluation metrics, including SARI, BLEU, SAMSA, FKGL, and BERT-Score, as well as a linguistic feature analysis on the simplification pairs utilizing the TS-eval package by Martin et al. (2018). EASSE also stores English TS test sets and outputs of English TS systems, as well as builds an evaluation report regarding all specified metrics of all specified TS models to facilitate the whole evaluation process. It is commonly used to evaluate TS system outputs in English and other languages. In this work, we will adapt EASSE in order to be better suitable for evaluation in other languages than English.

3 System Overview: EASSE-multi

In order to make EASSE language-independent and more robust for evaluating texts of languages other than English, we are proposing EASSE-multi (and its German variant EASSE-DE in the next section).

Therefore, we add a language constant to EASSE-multi to specify the currently evaluated language (e.g., "DE" for German in EASSE-DE). We also add SpaCy to the list of possible tokenizers to allow tokenization specified for languages other than English (see subsection 3.1).

The language constant also allows to choose language-specific evaluation metrics, e.g., readability metrics (see subsection 3.3), different models for BERT-Score (see subsection 3.2) and multi-lingual linguistic feature extraction (see subsection 3.4).

3.1 Tokenization

The original EASSE version currently supports 13a tokenization or white-space split tokenization (presuming pre-tokenized data). To include the language component into tokenization, we added the tokenizers of SpaCy (Montani et al., 2023) and the extension Spacy-Stanza (Qi et al., 2020)[1] as they currently support the tokenization of roughly 70 languages and also support linguistic annotations, e.g., part-of-speech tagging and dependency parsing, which will be relevant for the linguistic feature extraction.

3.2 Metrics

Evaluation metrics for TS are mostly language-independent, e.g., SARI, or BLEU, as they are n-gram-based methods. However, the n-grams depend on tokenization, which differs from language to language (see previous section). On the other hand, there are also language-specific evaluation metrics: Following Zhang* et al. (2020), BERT-Score can be used for a specific language (e.g., using the English-only model RoBERTa (Liu et al., 2019)) or in a multi-lingual setting (e.g., using a multi-lingual model such as BERT-multilingual (Devlin et al., 2019)).

In EASSE-multi, the usage of the metrics is optimized regarding the evaluated language, as based on the language constant, the tokenizer and the BERT-model are chosen to fit non-English languages better.

3.3 Readability

Readability scores and the LENS-Score (Maddela et al., 2023) are language-dependent, for the first due to included language-specific averages of word and sentence lengths and for the second due to training an evaluation score exclusively on English.

As an extension of EASSE, we also added readability formulas for languages other than English to EASSE-multi, which have already been implemented in the textstat package – a package for measuring readability and complexity in different languages. For example, common readability scores for German are the Amstad's adaption on the Flesch Reading Ease (FRE) or the Vienna non-fictional text formulas (Bamberger and

[1] https://github.com/explosion/spacy-stanza

Vanecek, 1984). LENS has not been reproduced for other languages due to missing required human assessment labels; hence, it makes no sense to include it in EASSE-multi.

Following the criticism of Tanprasert and Kauchak (2021) regarding readability metrics for TS evaluation, we follow their recommendation and include average sentence length, number of syllables and number of splits in our report. Hence, we add these features to the default report.

3.4 Multi-lingual Feature Extraction

As argued in Tanprasert and Kauchak (2021) and Alva-Manchego et al. (2019), we include a few linguistic features to get more insights into the system-generated simplification. For this, we are using the feature extraction toolkit of the reference-less quality estimation tool (further called TS-eval) by Martin et al. (2018) for the English analysis and its extended language-independent version TS-eval-multi by Stodden and Kallmeyer (2020). We decided to use TS-eval-multi for feature extraction and not the similar language-independent feature extraction toolkit called LFTK (Lee and Lee, 2023) as both versions of TS-eval focus more on features for text simplification, whereas LFTK focuses more on features for readability assessment. The TS-eval package has also already been integrated into the evaluation package EASSE, which facilitates its extension to the multilingual TS-eval. Further, most of LFTK's implemented features only apply to English. In future work, TS-eval-multi could be extended with features of LFTK. TS-eval-multi contains, for example, the parse tree height, cosine similarity between source and output based on pre-trained word embeddings, and length of phrases and clauses.

3.5 Additional Resources

The original EASSE framework also includes resources of English TS, i.e., English TS test sets, word lists, and system outputs of English TS models. With EASSE-multi, this component can be extended to the language of interest. We exemplify this with EASSE-DE and add only German resources (see section 4). However, the German resources can be easily replaced with resources of other languages.

3.6 Recommended Setting

At the moment, we cannot provide recommended settings per language except specifying the language constant, using SpaCy for tokenization, and using the multilingual BERT-Score. Further recommendations, for example, if case sensitivity is useful for the language of interest or determining which BERT version is more suitable for the language of interest, require more analysis which is out of the scope of this work. However, we recommend always naming which kind of settings have been used during evaluation as it can greatly influence the TS metrics. The settings should be reported in detail to ensure that the effect on the metric is due to the TS system and not the evaluation metrics' settings.

Furthermore, it could be helpful to report the results of the baselines, e.g., src2src (i.e., source-to-source or using the original complex sentence as input and output) or tgt2tgt (target-to-target or using the simple sentence as input and output). If the system outputs cannot be made available, it could help to verify on the gold data whether the applied evaluation method (e.g., in a replication experiment) is the same as the evaluation method used for an original experiment, as the results should be identical. Additionally, it could be helpful to re-evaluate the data comparing to. Therefore, we recommend making the system outputs publicly available (if the data is not restricted by license or copyright), e.g., as part of the EASSE-DE resources.

3.7 Usage

In order to customize EASSE-multi for a specific language (e.g., EASSE-DE for German or EASSE-ES for Spain), a few steps are necessary. First, the framework needs to be updated with language-specific data, i.e., TS test sets, (optionally) system outputs, and a SpaCy model[2] in the language of interest. Next, the settings[3] should be edited to fit the language, i.e., a) set the language constant, b) decide on considering or ignoring casing, c) edit metric scores (e.g., add language-specific readability scores), and d) (optionally) specify test set names and paths. Then, you can either run EASSE-multi to evaluate one single model or generate a report of scores for several models. More instructions on how to use EASSE-multi can be found in the GitHub repository[4].

4 EASSE-DE: Using EASSE-multi for German TS Evaluation

We will exemplify the usage of EASSE-multi for one language, i.e., German, resulting in EASSE-DE[5]. We have decided on German, as it is well-researched language in the research field of TS and enough resources (i.e., TS models, test sets, and system outputs) are available for a reasonable showcase project.

Therefore, we add German resources to EASSE-DE (see subsection 4.1), i.e., German sentence simplification test sets (see subsubsection 4.1.1), and available outputs of German TS systems regarding these test sets (see subsubsection 4.1.2). Further, we analyze whether and to what extent differences exist when evaluating German TS with the original evaluation framework EASSE or its adaptation EASSE-DE (see subsection 4.2).

[2] https://spacy.io/usage/models

[3] You can find the settings file here: https://github.com/rstodden/easse-multi/blob/master/easse/utils/constants.py

[4] https://github.com/rstodden/easse-multi

[5] https://github.com/rstodden/easse-de

name	target group	domain	size	# ref.	n:m	complex			simple		
						FRE↓	sent. len.↑	word len.↑	FRE↑	sent. len.↓	word len.↓
ABGB	non-experts	law	448	2	40%	42.75	24.85	1.83	44.6	22.39	1.89
APA_LHA-or-a2	Non-native speaker	news	500	1	6 %	44.7	20.2	1.92	69.55	11.27	1.78
APA_LHA-or-b1	Non-native speaker	news	500	1	8 %	43.7	20.48	1.93	62.6	12.82	1.83
BiSECT	people w. reading problems	politics	753	1	100 %	**8.55**	**30.24**	2.01	35.85	15.72	1.98
DEplain-APA	Non-native speaker	news	1,231	1	27 %	58.75	11.92	1.86	65.8	10.55	1.79
DEplain-web	mixed	web/mixed	1,846	1	57 %	62.95	19.13	1.64	**77.9**	10.76	**1.57**
GEOlino	children	encyclopedia	663	1	40 %	61.5	13.31	1.7	66.0	9.94	1.66
simple-german-corpus	mixed	web/mixed	391	1	73 %	41.15	13.96	2.0	65.4	**9.31**	1.83
TextComplexityDE	Non-native speaker	encyclopedia	250	1	83 %	28.1	27.75	**2.08**	51.2	14.17	1.9

Table 1: Overview Test Sets for German Sentence Simplification which are included in EASSE-DE. Including the target group, domain, size in sentence pairs, number of references, percentage of $n : m$ alignments, word length measured in syllables, and sentence length measured in words.

System Name	Reference	Type	Training Data	# Simp. Pairs	URL
hda-etr	Siegel et al. (2019)	rule-based	-	-	https://github.com/hdaSprachtechnologie/easy-to-understand_language
sockeye-APA-LHA	Spring et al. (2021) & Ebling et al. (2022)	seq2seq	APA-LHA OR-A2 & APA-LHA OR-B1	8,455 & 9,268	https://github.com/ZurichNLP/RANLP2021-German-ATS
sockeye-DEplain-APA	Stodden (2024)	seq2seq	DEplain-APA	10,660	https://huggingface.co/DEplain
mBART-DEplain-APA	Stodden et al. (2023)	fine-tuned seq2seq	DEplain-APA	10,660	https://huggingface.co/DEplain/trimmed_mbart_sents_apa
mBART-DEplain-APA+web	Stodden et al. (2023)	fine-tuned seq2seq	DEplain-APA+web	10,660 + 1,594	https://huggingface.co/DEplain/trimmed_mbart_sents_apa_web
mT5-DEplain-APA	Stodden (2024)	fine-tuned seq2seq	DEplain-APA	10,660	https://huggingface.co/DEplain
mT5-SGC	Stodden (2024)	fine-tuned seq2seq	SGC	4,430	https://huggingface.co/DEplain
BLOOM-zero	Ryan et al. (2023)	zero-shot AR model	-	-	https://github.com/XenonMolecule/MultiSim
BLOOM-sim-10	Ryan et al. (2023)	few-shot AR model	TCDE19 & GEOlino	200 & 959	https://github.com/XenonMolecule/MultiSim
BLOOM-random 10	Ryan et al. (2023)	few-shot AR model	TCDE19 & GEOlino	200 & 959	https://github.com/XenonMolecule/MultiSim
custom-decoder-ats	Anschütz et al. (2023)	AR model + fine-tuned seq2seq	Simplified, monolingual German data & 20Minuten	544,467 & 17,905	https://huggingface.co/josh-oo/custom-decoder-ats

Table 2: Overview of German TS models including training details (i.e., training data and size of training samples). Each line separates different model types. Adaptation from Stodden (2024).

4.1 German TS Resources

4.1.1 German TS Test Sets

For a better overview of available test sets for German sentence simplification, we have added gold data, i.e., manually simplified complex-simple sentence pairs, to EASSE-DE. In more detail, EASSE-DE refers to nine test sets, i.e., ABGB (Meister, 2023), APA-LHA-OR-A2 (Spring et al., 2021), APA-LHA-OR-B1 (Spring et al., 2021), BiSECT (Kim et al., 2021), DEplain-APA (Stodden et al., 2023), DEplain-web (Stodden et al., 2023), TextComplexityDE (Naderi et al., 2019), GEOlino (Mallinson et al., 2020), and Simple-German-Corpus (Toborek et al., 2023). We refer to Table 1 for more meta data of the test sets.

4.1.2 German TS Models

For German TS, a few models are available or reproducible, e.g., ZEST, by Mallinson et al. (2020), sockeye by Spring et al. (2021), custom-decoder-ats by Anschütz et al. (2023), the few-shot approaches on BLOOM by Ryan et al. (2023), or the mBART models by Stodden et al. (2023). A more detailed description and analysis of German TS models, including their reproduction, has been recently proposed by Stodden (2024). The system outputs of all reproduced German TS models (see Table 2) have been added to EASSE-DE to facilitate a better comparison between existing models and models which will be newly proposed in future.

4.2 Comparison of EASSE and EASSE-DE

In the following section, we present and analyse the metric scores when using either the original EASSE or the adapted version EASSE-DE, including different settings on three German test sets of one German TS model.

4.2.1 Method

Evaluation Settings. In the comparative analysis, we focus on the settings in EASSE regarding i) language specification (i.e., English vs. German), ii) tokenization method (i.e., none vs 13a vs SpaCy), iii) BERT model version (i.e., RoBERTa-large vs BERT-base-multilingual-cased), iv) FRE version (English vs German). Due to their n-gram-based approach, we expect the tokenization method to have an effect on SARI and BLEU but not on BERT-Score-Precision.

German TS Test Sets. In the analysis, we evaluate on three available German TS test sets: DEplain-APA (Stodden et al., 2023), DEplain-web (Stodden et al., 2023), and TextComplexityDE (Naderi et al., 2019).

These test sets are all manually simplified and manually aligned, and, therefore, we expect a higher simplification quality for them as for other test sets, e.g., BiSECT (Kim et al., 2021)[6] or APA-LHA (Spring et al., 2021)[7]. Further, these three test sets include texts of different domains (news, web, and Wikipedia), and their simplification addresses different target groups (non-native speakers and people with cognitive disabilities). Hence, they represent different kinds of simplifications and therefore seem to be a good choice for our analysis.

German TS Model. Further, we have selected the generated simplifications of one model, i.e., mBART-DEplain-APA+web. Reasons for the choice of this model are that it is ready-to-use without additional examples, and, following Stodden (2024), this model achieves the best BERT-Scores across several test sets. In comparison, the BLOOM models by Ryan et al. (2023) are few-shot models that require additional complex-simple pairs to generate simplifications.

4.2.2 Results

The results of the mBART-APA+web model with different settings are presented in Table 3.[8] In the following, we analyse the differences regarding tokenization, readability scores, multi-lingual BERT-Score, and system rankings.

	Tok.	Lang.	BLEU↑	SARI↑	BS-P↑	FRE↑
TCDE19 (n = 250)	spacy	EN	**18.56**	**37.69**	0.39	**57.37**
	spacy	DE	17.75	37.37	**0.55**	43.65
	13a	DE	18.04	37.41	**0.55**	43.55
	none	DE	16.04	37.47	**0.55**	43.65
DEplain-APA (n = 1231)	spacy	EN	**30.59**	**34.79**	0.48	**78.25**
	spacy	DE	28.03	33.81	**0.64**	65.2
	13a	DE	28.37	33.92	**0.64**	65.2
	none	DE	24.69	32.88	**0.64**	65.2
DEplain-web (n = 1846)	spacy	EN	18.37	**34.21**	0.27	**76.52**
	spacy	DE	17.99	34.07	**0.44**	69.05
	13a	DE	18.17	34.10	**0.44**	69.05
	none	DE	15.97	33.67	**0.44**	69.05

Table 3: Scores of trimmed-mbart-DEplain-APA+web when using different language settings and tokenizers.

Tokenization. As expected, different tokenization methods (including language specification) affect the calculation of metrics used for TS evaluation. The last three rows in each block of Table 3 show the differences in the scores when using different tokenization strategies. We can see that the BERT-score is always the same for all settings due to the sub-word tokenization in BERT. The FRE scores are also robust across all test sets when looking at the trimmed-mBART results, but in Appendix A Table 5, we see slightly more differences. The SARI scores also change slightly, i.e., to less than 1 point in all settings, whereas the differences in the BLEU scores range between 2 to 3 points in all test sets. In conclusion, when comparing one model against another with a slightly different evaluation setting (here, the tokenizer), even these small changes can be wrongly interpreted as an improvement of the model idiosyncrasy. However, it is only due to the different settings. Therefore, we recommend stating all settings chosen for evaluation for a more reliable comparison.

Readability Metrics. As can be seen in Table 3, the scores are quite different wrt. to FRE for the English and German settings (see first two rows in each block). The results are different due to the different constants of the formulas and their dependency on different tokenization and syllable splitting. When interpreting the readability scores, they also result in different categories: Following (Amstad, 1978), the simplifications with the English setting on DEplain-APA and DEplain-web can be described as "ease" whereas they are categorized as "simple" using the German setting. In summary, the language adaptation of readability scores can make a noticeable difference when interpreting the simplification results.

BERT-Score. As shown in the first two rows of each row-block in Table 3, changing the transformer model of the BERT-Score significantly affects the BERT-Score. The scores using the multi-lingual model are much higher than those using the only-English model. Hence, the choice of the BERT model seems to have a high effect on the TS evaluation.

System Rankings. When evaluating TS systems, often their ranks are compared to each other instead of the exact scores. Therefore, we have analysed whether the ranks changes when evaluating 11 German TS systems (and 2 baselines) either with the original EASSE or with EASSE-DE.[9] As can be seen in Table 4, the ranks of the models wrt. BLEU, and BERT-Score-Precision are slightly changing depending on the EASSE version whereas the ranks for SARI are constant. Contrary to the ranks, changes are visible wrt. the scores. When evaluating more similar systems (e.g., during hyperparameter tuning) the differences might get more meaningful and relevant also with respect to the ranks. Therefore, it is important to specify the settings used for evaluation to have a reliable comparison.

5 Benchmark for German TS

EASSE-DE facilitates modeling German text simplification by providing a unified evaluation framework as well as storing data of several German test sets (see Table 1). Additionally with the provided system outputs of reproduced German TS systems (Stodden, 2024), a

[6] BiSECT is generated using machine translation of English texts. Due to this augmentation strategy, the German version includes encoding errors.

[7] The training and validation sets of APA-LHA are automatically aligned, and, hence, more faulty compared to manually aligned corpora.

[8] To ensure that the effects are not due to the system but to the evaluation changes, we also add the results of the identity baseline (see Table 5).

[9] The system outputs, which have been used for this analysis, are available upon request at https://doi.org/10.5281/zenodo.13891495.

	BLEU↑		SARI↑		BS-P↑	
	S	R	S	R	S	R
hda_LS	22.3	5	26.06	12	0.55	7
sockeye-APA-LHA	11.84	11	40.16	3	0.37	12
sockeye-DEplain-APA	19.58	7	44.14	1	0.53	9
mbart_DEplain_apa	28.49	1	38.72	5	0.64	1
mbart_DEplain_apa_web	28.03	2	33.81	10	0.64	1
mT5-DEplain-APA	22.32	4	39.41	4	0.61	4
mt5-simple-german-corpus	8.12	12	37.92	6	0.48	11
BLOOM-zero	16.14	9	35.43	9	0.53	9
BLOOM-10-random	17.97	8	35.93	8	0.57	5
BLOOM-10-similarity	20.97	6	41.27	2	0.57	5
custom-decoder-ats	1.24	13	36.42	7	0.16	13
Identity baseline	26.89	3	15.25	13	0.63	3
Truncate baseline	16.11	10	27.2	11	0.55	7

(a) Evaluated with default settings of EASSE-DE, i.e., no lower-casing and SpaCy tokenizer.

	BLEU↑		SARI↑		BS-P↑	
	S	R	S	R	S	R
hda_LS	23.77	4	26.82	12	0.38	7
sockeye-APA-LHA	12.42	11	40.27	3	0.13	12
sockeye-DEplain-APA	20.97	7	44.89	1	0.36	9
mbart_DEplain_APA	30.01	1	39.12	5	0.47	1
mbart_DEplain_APA_web	29.62	2	34.44	10	0.47	1
mT5-DEplain-APA	23.7	5	39.8	4	0.46	3
mt5-simple-german-corpus	8.92	12	38.2	6	0.29	11
BLOOM-zero	17.23	10	35.19	9	0.36	9
BLOOM-10-random	19.23	8	35.52	8	0.38	7
BLOOM-10-similarity	22.21	6	41.21	2	0.39	6
custom-decoder-ats	1.29	13	36.65	7	-0.13	13
Identity baseline	28.5	3	15.88	13	0.45	4
Truncate baseline	18.94	9	28.31	11	0.41	5

(b) Evaluated with default settings of original EASSE, i.e., lower-casing and 13a tokenizer.

Table 4: Scores (S) and ranks (R) of German TS models on the DEplain-APA test set.

benchmark for German TS can be easily build and updated using EASSE-DE. In Appendix B, we provide a German TS benchmark including results of 7 German TS models (see Table 2) on 7 German test sets of the domains of news, web, and Wikipedia texts.

As discussed in Stodden (2024), there is no clear picture regarding best performing models across all domains or test sets. As expected, models achieve the best scores if they are evaluated and trained on the same corpus. However, corresponding to the ranks following the metrics' scores the models are ranked differently, e.g., a model gets the highest SARI score but lower BS_P scores and vice versa. For a reliable interpretation of the metrics, there is more research to be done regarding finding new evaluation metrics and checking the suitability of existing metrics on languages other than English.

6 Discussion & Conclusion

We have proposed EASSE-multi, which facilitates easy evaluation of sentence simplification in multiple languages. Therefore, we have extended the original EASSE package with a language-constant tokenizer, language-dependent version of BERT-Score, and language-wise readability scores.

Further, we have exemplified using EASSE-multi for German TS evaluation in the form of EASSE-DE. In comparing the results generated by EASSE and EASSE-DE, we have shown that it is important to consider the text's language when evaluating. Following that, we recommend using EASSE-DE over EASSE when evaluating German sentence simplification models as it includes language-sensitive evaluation metrics. Even if the scores per metric might be lower when using EASSE-DE than EASSE, we argue that these are more reliable due to the language-sensitive metrics.

Further, we argue that it is unreliable to compare scores (maybe originating from different papers) as they might be generated by using different evaluation settings. Before making a comparison, we recommend verifying whether the same settings of the metric have been used in both experiments (the referenced and the new one). Otherwise, the differences in the scores might not be dependent on the model changes (which is the question of interest) but on, for example, different kinds of tokenization. Therefore, we strongly recommend always specifying the settings or, even better, the implementation of the metrics used for the evaluation, as it can have a huge impact on the reported scores. We identified the following aspects which should be reported accompanied with automatic evaluation: 1. language setting (e.g., EN, or DE) for features (e.g., BERT-Score, FRE, or word length), 2. tokenizer (e.g., none, 13a, or SpaCy), 3. lower casing (True or False), 4. BERT-Score model (e.g., RoBERTa-large, mT5, or BERT-base-multilingual-cased)

7 Future Work

Even if most of the scores are language-independent or can be easily adapted to work for other languages, as shown previously, there still might be problems in using the same scores for different languages due to language idiosyncrasies and different simplification operations per language. Approaches in the direction of language-wise evaluation of non-English TS could be learnable metrics (per language) as already proposed for English, e.g., LENS, BETS, or MeaningBERT. In future work, we want to investigate learnable metrics for non-English languages to fit the language idiosyncrasies better and add them to EASSE-DE.

Further, we would like to extend EASSE-DE to include more German TS resources. We hope that EASSE-DE will be useful for German TS researchers and invite them to contribute their test sets or system outputs to EASSE-DE.

Acknowledgements

First, we want to thank the authors of the original EASSE paper, without their highly valuable contribution (openly licensed code, and clearly structured and easy adaptable code) this extension would have not been possible. We also want to thank the reviewers for their valuable feedback, which has helped to strengthen the paper.

Lay Summary

The process of automatically rewriting texts is also called "automatic text simplification". Automatic text simplification can be defined as: the change of word choice in a text and/or the restructuring of a sentence to be better understandable for a given target group. Often, research in text simplification focuses on the simplification of English texts. In this work, we facilitate the research on text simplification in multiple languages. In more detail, we have focused on the evaluation of automatic text simplification systems for multiple languages. Therefore, we have provided an evaluation toolkit which can be used to evaluate the output of text simplification systems.

Additionally, we have showcased the usage of this toolkit for German. We are providing an easy-to-use framework for German text simplification including a selection of test sets, system outputs of several German TS models and a report regarding their quality.

Limitations

In this work, we have just showcased the usage of EASSE-multi for German, although it is also applicable to other languages. Furthermore, we have focused on openly licensed TS models and, hence, we have not included proprietary language models, e.g., ChatGPT.

References

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

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.

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service. PhD Thesis.

Miriam Anschütz, Joshua Oehms, Thomas Wimmer, Bartłomiej Jezierski, and Georg Groh. 2023. Language models for German text simplification: Overcoming parallel data scarcity through style-specific pre-training. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1147–1158, Toronto, Canada. Association for Computational Linguistics.

Richard Bamberger and Erich Vanecek. 1984. *Lesen - Verstehen - Lernen - Schreiben. Die Schwierigkeitsstufen von Texten in deutscher Sprache.* Jugend u. Volk Sauerlaender, Wien.

David Beauchemin, Horacio Saggion, and Richard Khoury. 2023. Meaningbert: Assessing meaning preservation between sentences. *Frontiers in Artificial Intelligence*, 6.

Rémi Cardon and Natalia Grabar. 2020. French biomedical text simplification: When small and precise helps. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 710–716, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sarah Ebling, Alessia Battisti, Marek Kostrzewa, Dominik Pfütze, Annette Rios, Andreas Säuberli, and Nicolas Spring. 2022. Automatic text simplification for german. *Frontiers in Communication*, 7.

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

Itziar Gonzalez-Dios, Iker Gutiérrez-Fandiño, Oscar m. Cumbicus-Pineda, and Aitor Soroa. 2022. IrekiaLFes: a new open benchmark and baseline systems for Spanish automatic text simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 86–97, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Daniel Holmer and Evelina Rennes. 2023. Constructing pseudo-parallel Swedish sentence corpora for automatic text simplification. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 113–123, Tórshavn, Faroe Islands. University of Tartu Library.

Joongwon Kim, Mounica Maddela, Reno Kriz, Wei Xu, and Chris Callison-Burch. 2021. BiSECT: Learning to split and rephrase sentences with bitexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6193–6209, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bruce W. Lee and Jason Lee. 2023. LFTK: Handcrafted features in computational linguistics. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 1–19, Toronto, Canada. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

Mounica Maddela, Yao Dou, David Heineman, and Wei Xu. 2023. 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)*, pages 16383–16408, Toronto, Canada. Association for Computational Linguistics.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

Fabian Meister. 2023. ABGB-TextSimplification-Datasets. GitHub repository: `https://github.com/MeisterFa/ABGB-TextSimplification-Datasets`. Visited on 2023-12-01.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

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

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.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics.

Melanie Siegel, Dorothee Beermann, and Lars Hellan. 2019. Aspects of linguistic complexity: A german - norwegian approach to the creation of resources for easy-to-understand language. In *11th International Conference on Quality of Multimedia Experience QoMEX 2019, Berlin, Germany, June 5-7, 2019*, pages 1–3. IEEE.

Nicolas Spring, Annette Rios, and Sarah Ebling. 2021. Exploring German multi-level text simplification. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1339–1349, Held Online. INCOMA Ltd.

Regina Stodden. 2024. Reproduction & Benchmarking of German Text Simplification Systems. In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 1–15, Torino, Italia. ELRA and ICCL.

Regina Stodden and Laura Kallmeyer. 2020. A multilingual and cross-domain analysis of features for text simplification. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 77–84, Marseille, France. European Language Resources Association.

Regina Stodden, Omar Momen, and Laura Kallmeyer. 2023. DEplain: A German parallel corpus with intralingual translations into plain language for sentence and document simplification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16441–16463, Toronto, Canada. Association for Computational Linguistics.

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.

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.

Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. 2023. A new aligned simple German corpus. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

11393–11412, Toronto, Canada. Association for Computational Linguistics.

Jan Trienes, Jörg Schlötterer, Hans-Ulrich Schildhaus, and Christin Seifert. 2022. Patient-friendly clinical notes: Towards a new text simplification dataset. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 19–27, Abu Dhabi, United Arab Emirates (Virtual). 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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Xinran Zhao, Esin Durmus, and Dit-Yan Yeung. 2023. Towards reference-free text simplification evaluation with a BERT Siamese network architecture. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13250–13264, Toronto, Canada. Association for Computational Linguistics.

A Results of Identity Baseline

	Tok.	Lang.	BLEU↑	SARI↑	BS-P↑	FRE↑
TCDE19 (n = 250)	spacy	EN	**28.22**	15.31	0.37	**39.16**
	spacy	DE	27.31	14.99	**0.55**	28.1
	13a	DE	27.49	15.05	**0.55**	28.0
	none	DE	24.43	13.78	**0.55**	28.1
DEplain-APA (n = 1231)	spacy	EN	**29.28**	**16.17**	0.45	**77.64**
	spacy	DE	26.89	15.25	**0.63**	58.75
	13a	DE	27.25	15.35	**0.63**	64.6
	none	DE	23.33	13.75	**0.63**	58.75
DEplain-web (n = 1846)	spacy	EN	21.24	12.09	0.25	**70.33**
	spacy	DE	20.85	11.93	0.42	62.95
	13a	DE	20.89	11.94	0.42	62.95
	none	DE	18.82	10.9	0.42	62.95

Table 5: Scores of identity baseline on three test sets when using different language settings and tokenizers.

B German TS Benchmark

B.1 Evaluation on News Corpora

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
hda_LS	3.02	14.02	0.12	37.55	1.14	1.04
sockeye-APA-LHA	**13.59**	**51.77**	**0.35**	68.65	0.64	0.99
sockeye-DEplain-APA	4.79	40.32	0.25	**70.25**	0.71	1.25
mBART-DEplain-APA	4.73	30.28	0.23	57.55	0.85	1.33
mBART-DEplain-APA+web	4.56	25.89	0.23	56.35	0.84	1.16
mT5-DEplain-APA	4.65	34.47	0.24	58.10	0.58	1.09
mT5-SGC	2.78	39.79	0.28	**70.25**	**0.48**	1.00
BLOOM-zero	2.44	26.83	0.19	51.85	0.82	1.29
BLOOM-10-random	2.64	33.05	0.24	57.95	0.64	0.98
BLOOM-10-similarity	5.10	38.05	0.29	64.60	0.59	0.98
custom-decoder-ats	0.28	37.05	0.08	52.60	3.16	**2.91**
Identity baseline	3.50	3.90	0.18	44.70	1.00	1.00
Reference baseline	100	100	1.00	69.55	0.60	0.97
Truncate baseline	2.60	17.49	0.19	54.25	0.79	1.00

Table 6: Evaluation on APA-LHA-OR-A2 (copied from Stodden (2024)).

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
hda_LS	4.54	15.49	0.15	36.15	1.15	1.10
sockeye-APA-LHA	**11.00**	**44.93**	**0.32**	61.90	0.70	0.97
sockeye-DEplain-APA	3.57	39.4	0.25	**70.65**	0.68	1.26
mBART-DEplain-APA	5.32	30.94	0.26	57.65	0.86	1.37
mBART-DEplain-APA+web	5.81	26.61	0.25	56.05	0.85	1.19
mT5-DEplain-APA	4.92	35.70	0.26	57.70	0.57	1.10
mT5-SGC	2.54	39.36	0.29	70.45	**0.48**	1.00
BLOOM-zero	3.41	27.56	0.21	56.80	0.84	1.34
BLOOM-10-random	5.18	32.43	0.26	56.25	0.71	0.98
BLOOM-10-similarity	6.21	37.22	0.27	62.00	0.72	0.98
custom-decoder-ats	0.52	37.59	0.07	49.70	3.78	**3.51**
Identity baseline	5.47	4.89	0.22	43.70	1.00	1.00
Reference baseline	100	100	1.00	62.60	0.68	0.98
Truncate baseline	4.59	18.36	0.22	53.85	0.79	1.00

Table 7: Evaluation on APA-LHA-OR-B1 (copied from Stodden (2024)).

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
hda_LS	22.3	26.06	0.55	64.60	1.00	1.00
sockeye-APA-LHA	11.84	40.16	0.37	63.70	0.94	0.97
sockeye-DEplain-APA	19.58	**44.14**	0.53	71.45	0.94	1.09
mBART-DEplain-APA	**28.49**	38.72	**0.64**	65.30	0.99	1.07
mBART-DEplain-APA+web	28.03	33.81	**0.64**	65.20	0.98	1.05
mT5-DEplain-APA	22.32	39.41	0.61	63.20	0.87	1.04
mt5-SGC	8.12	37.92	0.48	**71.65**	**0.74**	1.00
BLOOM-zero	16.14	35.43	0.53	65.10	0.87	1.14
BLOOM-10-random	17.97	35.93	0.57	65.50	0.91	1.00
BLOOM-10-similarity	20.97	41.27	0.57	65.70	0.93	1.07
custom-decoder-ats	1.24	36.42	0.16	53.00	7.41	**5.07**
Identity baseline	26.89	15.25	0.63	58.75	1.00	1.00
Reference baseline	100.00	100.00	1.00	65.80	1.03	1.20
Truncate baseline	16.11	27.20	0.55	66.10	0.80	1.01

Table 8: Evaluation on DEplain-APA (copied from Stodden (2024)).

B.2 Evaluation on Web Corpora

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
sockeye-APA-LHA	0.24	32.41	0.13	69.55	0.74	0.90
sockeye-DEplain-APA	3.44	36.24	0.24	76.7	0.76	1.32
mBART-DEplain-APA	13.50	33.11	0.40	69.65	0.90	1.30
mBART-DEplain-APA+web	**17.99**	34.07	**0.44**	69.05	0.85	1.16
mT5-DEplain-APA	6.80	**37.15**	0.36	70.90	0.63	1.10
mt5-SGC	2.50	36.56	0.37	**78.10**	**0.47**	0.93
BLOOM-zero	10.88	30.58	0.35	70.30	0.85	1.28
BLOOM-10-random	11.06	30.90	0.39	68.55	0.69	0.98
BLOOM-10-similarity	11.62	37.03	0.42	70.05	0.63	0.98
custom-decoder-ats	0.72	34.92	0.10	57.15	5.41	**3.79**
Identity baseline	20.85	11.93	0.42	62.95	1.00	1.00
Reference baseline	100.00	100.00	1.00	77.90	0.94	1.84
Truncate baseline	17.28	24.58	0.40	67.05	0.82	1.02

Table 9: Evaluation on DEplain-web (copied from Stodden (2024)).

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio↓	Sent. splits↑
hda_LS	6.34	20.22	0.25	41.15	1.00	1.03
sockeye-APA-LHA	0.33	35.50	0.13	63.70	0.80	0.82
sockeye-DEplain-APA	1.35	37.86	0.18	**71.05**	0.79	1.01
mBART-DEplain-APA	5.70	32.77	0.31	58.15	0.97	1.00
mBART-DEplain-APA+web	6.56	29.80	0.33	44.95	1.61	1.09
mT5-DEplain-APA	2.81	35.92	0.30	51.45	0.76	0.88
mt5-SGC	3.30	43.62	0.37	58.55	**0.61**	0.85
BLOOM-zero	3.76	31.95	0.25	53.55	0.81	1.07
BLOOM-10-random	4.64	33.16	0.30	51.50	0.75	0.92
BLOOM-10-similarity	**13.32**	**44.66**	**0.38**	58.65	0.92	1.13
custom-decoder-ats	0.44	36.53	0.06	32.05	8.83	**3.68**
Identity baseline	7.46	6.51	0.29	41.15	1.00	1.00
Reference baseline	100.00	100.00	1.00	65.40	1.25	1.81
Truncate baseline	4.66	20.12	0.28	50.50	0.81	0.87

Table 10: Evaluation on SGC (copied from Stodden (2024)).

B.3 Evaluation on Knowledge Acquiring Corpora

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio ↓	Sent. splits↑
hda_LS	55.22	34.20	0.76	61.50	1.00	1.00
sockeye-APA-LHA	0.69	18.94	0.15	69.45	1.05	0.92
sockeye-DEplain-APA	7.27	24.71	0.33	77.3	0.96	1.15
mBART-DEplain-APA	50.56	**44.29**	0.74	70.75	1.04	1.15
mBART-DEplain-APA+web	**55.35**	44.28	**0.79**	64.60	0.97	1.08
mT5-DEplain-APA	28.43	36.93	0.65	67.95	0.80	1.04
mt5-SGC	11.92	28.75	0.55	**78.30**	**0.70**	0.94
BLOOM-zero	28.18	32.15	0.59	67.85	0.87	1.26
custom-decoder-ats	0.77	22.05	0.08	46.55	14.61	**4.76**
Identity baseline	67.12	26.81	0.86	61.50	1.00	1.00
Reference baseline	100.00	100.00	1.00	66.00	0.95	1.32
Truncate baseline	45.39	29.78	0.75	63.80	0.83	1.00

Table 11: Evaluation on GEOlino (n=663) (copied from Stodden (2024)).

	BLEU↑	SARI↑	BS_P↑	FRE↑	Compr. ratio ↓	Sent. splits↑
hda_LS	20.66	26.92	0.45	33.65	1.00	1.01
sockeye-APA-LHA	0.13	29.87	0.14	**69.05**	0.43	0.97
sockeye-DEplain-APA	0.68	31.79	0.19	65.0	0.51	1.42
mBART-DEplain-APA	13.69	**39.14**	0.50	51.10	0.76	1.57
mBART-DEplain-APA+web	**17.75**	37.37	**0.55**	43.65	0.74	1.29
mT5-DEplain-APA	2.84	35.09	0.40	46.60	0.40	1.14
mt5-SGC	1.05	32.98	0.38	64.40	**0.31**	0.97
BLOOM-zero	9.46	34.96	0.42	45.55	0.78	1.75
custom-decoder-ats	1.73	32.87	0.22	27.70	1.54	**4.22**
Identity baseline	27.31	14.99	0.55	28.10	1.00	1.00
Reference baseline	100.00	100.00	1.00	51.20	0.95	2.04
Truncate baseline	20.17	26.45	0.52	37.65	0.81	1.00

Table 12: Evaluation on TCDE19 (n=250) (copied from Stodden (2024)).