
Article:
Alva-Manchego, F., Scarton, C. and Specia, L. (2021) The (un)suitability of automatic evaluation metrics for text simplification. Computational Linguistics, 47 (4). pp. 861-889. ISSN 0891-2017

https://doi.org/10.1162/coli_a_00418

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Short Paper

The (Un)Suitability of Automatic Evaluation Metrics for Text Simplification

Fernando Alva-Manchego*
University of Sheffield

Carolina Scarton*
University of Sheffield

Lucia Specia**
Imperial College London

In order to simplify sentences, several rewriting operations can be performed such as replacing complex words per simpler synonyms, deleting unnecessary information, and splitting long sentences. Despite this multi-operation nature, evaluation of automatic simplification systems relies on metrics that moderately correlate with human judgements on the simplicity achieved by executing specific operations (e.g. simplicity gain based on lexical replacements). In this article, we investigate how well existing metrics can assess sentence-level simplifications where multiple operations may have been applied and which, therefore, require more general simplicity judgements. For that, we first collect a new and more reliable dataset for evaluating the correlation of metrics and human judgements of overall simplicity. Second, we conduct the first meta-evaluation of automatic metrics in Text Simplification, using our new dataset (and other existing data) to analyse the variation of the correlation between metrics' scores and human judgements across three dimensions: the perceived simplicity level, the system type and the set of references used for computation. We show that these three aspects affect the correlations and, in particular, highlight the limitations of commonly-used operation-specific metrics. Finally, based on our findings, we propose a set of recommendations for automatic evaluation of multi-operation simplifications, suggesting which metrics to compute and how to interpret their scores.

1. Introduction

Text Simplification consists of modifying the content and structure of a text in order to make it easier to read and understand, while preserving its main idea and as much as possible of its original meaning. Human editors simplify through several rewriting operations, such as lexical paraphrasing (i.e. replacing complex words/phrases with simpler synonyms and some rewording for fluency), changing the syntactic structure of sentences (e.g. splitting or reordering components), or removing information deemed non-essential to understand the main idea of the original text (Petersen 2007; Aluísio et al. 2008; Bott and Saggion 2011; Xu, Callison-Burch, and Napoles 2015). Modern sys-

* Department of Computer Science, University of Sheffield, UK. E-mail: {f.alva, c.scarton}@sheffield.ac.uk.
** Department of Computing, Imperial College London, UK. E-mail: l.specia@imperial.ac.uk

Submission received: 24 December 2020; revised version received: 14 July 2021; accepted for publication: 28 July 2021.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

tems for Automatic Text Simplification are sentence-level, and attempt to replicate this multi-operation rewriting process by leveraging corpora of parallel original-simplified sentence pairs (Alva-Manchego, Scarton, and Specia 2020). However, the simplicity of automatic sentence-level simplifications is measured with metrics that evaluate single specific operations. For instance, SARI (Xu et al. 2016) was designed to estimate simplicity gain when just lexical paraphrasing was being assessed, whilst SAMSA (Sulem, Abend, and Rappoport 2018b) attempts to quantify structural simplicity by verifying the correctness of sentence splitting. In a recent study, Alva-Manchego et al. (2020) showed that, for the same set of original sentences, human judges preferred manual simplifications where multiple edit operations had been applied over those where only one operation had been performed (i.e. only lexical paraphrasing or only splitting). However, the authors also provided preliminary evidence that both a general metric like BLEU (Papineni et al. 2002), and an operation-specific one like SARI had poor correlations with judgements of overall simplicity when computed using multi-operation manual references.

In this article, we study the extent to which evaluation metrics can estimate the simplicity of automatic **sentence-level simplifications** where **multiple rewriting operations** may have been applied. In order to do so, we: (1) create a new dataset with direct assessments of simplicity; (2) perform the first meta-evaluation of automatic metrics for sentence-level Text Simplification, focused on their correlation with human judgements on simplicity; and (3) propose a set of guidelines for automatic evaluation of sentence-level simplifications, seeking to improve the interpretation of automatic scores, especially for multi-operation simplifications.[1]

In the remainder of the paper, we first review manual and automatic evaluation methods in Sentence Simplification (Sec. 2). Then, we describe two existing datasets with human judgements on simplicity gain and structural simplicity of system outputs, whose limitations motivate the collection of a new dataset with overall simplicity scores crowdsourced through Direct Assessment (Sec. 3). After that, we study the variation in sentence-level correlations between automatic metrics and human judgements under three test conditions: the level of perceived simplicity, the approach implemented by the simplification systems, and the set of manual simplification references (Sec. 4). For direct assessments of simplicity, in particular, we show that: (a) metrics can more reliably score low-quality simplifications; (b) most metrics are better at scoring system outputs from neural sequence-to-sequence models; and (c) computing metrics using all available manual references for each original sentence does not significantly improve their correlations. We also propose explanations on the low-to-moderate correlations achieved by simplification-specific metrics. Based on our findings, we propose a set of recommendations for better evaluation of automatic sentence-level simplifications and suggest ways to improve current practices (Sec. 5). Among these, we suggest to first compute BERTScore (Zhang et al. 2020) to verify that the system output is of high quality, and then use SARI and/or SAMSA to measure the gains in simplicity. Finally, we summarise our results, highlighting our contributions and conclusions (Sec. 6).

1 Our new dataset and the code to reproduce our experiments are available in https://github.com/feralvam/metaeval-simplification.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

2. Background

The preferred method for evaluating the quality of automatic simplifications is eliciting human judgements on grammaticality, meaning preservation and simplicity. However, these can be costly to obtain while tuning simplification models, especially at large scale. This creates scenarios where automatic metrics act as proxies for human judgements, so it is important to understand how these metrics behave under different circumstances, to better interpret their scores. We first review common practices for collecting human judgements on the simplicity of system outputs against which metrics are evaluated, and motivate our choice of Direct Assessment as our data labelling methodology. Then, we briefly explain the benefits of conducting meta-evaluations of automatic metrics.

2.1 Human Evaluation of Simplicity

When obtaining human judgements on the simplicity of system outputs, there are three components to consider: the question to elicit the judgement, what the judges are shown, and how they submit their judgement. It is generally agreed to show both the original and simplified sentences so that raters can determine if the latter is simpler than the former. However, several variations have been tested for the other two components.

It is common to ask how much simpler the system output is compared to the original sentence, using Likert scales of 0-5, 1-4 or 1-5 (the higher the better) to submit discrete scores (Woodsend and Lapata 2011; Wubben, van den Bosch, and Krahmer 2012; Feblowitz and Kauchak 2013; Narayan and Gardent 2014, 2016; Zhang and Lapata 2017; Alva-Manchego et al. 2017; Vu et al. 2018; Guo, Pasunuru, and Bansal 2018; Dong et al. 2019; Kriz et al. 2019; Kumar et al. 2020; Jiang et al. 2020). A variation in the scale is presented in (Nisioi et al. 2017) with -2 to +2 scores instead, allowing to distinguish instances with no changes in their simplicity (0), and instances where the automatic system hurt the readability of the original sentence (-1 or -2).

Most work does not specify what "being simpler" entails, and trusts human judges to use their own understanding of the concept. In contrast, Xu et al. (2016) experimented with Simplicity Gain, asking judges to count *"how many successful lexical or syntactic paraphrases occurred in the simplification"*. The authors argue that this framing of the task allows for easier judgements and more informative interpretation of the scores, while reducing the bias towards models that perform minimal modifications. In a similar fashion, Nisioi et al. (2017) and Cooper and Shardlow (2020) asked judges to count the number of changes made by automatic systems, and then to identify how many of them were "correct" (i.e. preserved meaning and grammaticality, while making the sentence easier to understand). On a different line of work, Sulem, Abend, and Rappoport (2018b,c) focused on Structural Simplicity, requesting judges to use the -2 to +2 scale to answer *"is the output simpler than the input, ignoring the complexity of the words?"* This is intended to focus the evaluation in a specific operation: sentence splitting.

For the dataset collected as part of our study (Sec. 3.3), we follow common practice and present human judges with both original sentences and their automatic simplifications. Furthermore, since the focus of this article is on multi-operation simplifications, we rely on a general definition of simplicity instead of one for a specific (set of) operation(s). Finally, as in Alva-Manchego et al. (2020), we experiment with collecting continuous scores following the Direct Assessment methodology (Graham et al. 2017), since they can be standardised to remove individual rater's biases, resulting in higher inter-annotator agreement (Graham et al. 2013).

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

2.2 Automatic Evaluation of Simplicity

BLEU (Papineni et al. 2002) and SARI (Xu et al. 2016) are the most commonly used metrics in Sentence Simplification. Whilst BLEU scores can be misleading for several text generation tasks (Reiter 2018), in the case of Simplification they have been shown to correlate well with human assessments of grammaticality and meaning preservation (Wubben, van den Bosch, and Krahmer 2012; Štajner, Mitkov, and Saggion 2014; Xu et al. 2016; Sulem, Abend, and Rappoport 2018a; Alva-Manchego et al. 2020). SARI, on the other hand, is better suited for evaluating the simplicity of system outputs produced via lexical paraphrasing. It does so by comparing the automatic simplification to both the original sentence and multiple manual references, and measuring the correctness of the words added, kept and deleted. Although not widely adopted, SAMSA (Sulem, Abend, and Rappoport 2018b) is another simplicity-specific metric, but focused on sentence splitting. It validates that each simple sentence resulting from splitting a complex one is correctly formed (i.e. it corresponds to a single *Scene* with all its *Participants*).

Studies on the correlation of human judgements on simplicity and automatic scores have been performed when introducing new metrics or datasets.[2] Xu et al. (2016) argued that SARI correlates with crowdsourced judgements of Simplicity Gain when the simplification references had been produced by lexical paraphrasing, whilst SAMSA was shown to correlate with expert judgements of Structural Simplicity. When introducing HSplit (Sulem, Abend, and Rappoport 2018a), a dataset of manual references for sentence splitting, the authors argued that BLEU (Papineni et al. 2002) was not a good estimate for (Structural) Simplicity. However, these studies did not analyse if the absolute correlations varied in different subgroups of the data. In contrast, our study shows that correlations are affected by the perceived quality of the simplifications, the types of the simplification systems, and the set of manual references used.

3. Datasets with Human Judgements on Simplicity

In this Section, we describe the datasets that will be used in our meta-evaluation study. Each dataset is composed of a set of original sentences, their automatic simplifications produced by various simplification systems, and human evaluations on (some form of) simplicity for all system outputs. These datasets were chosen (or created) since:

1. Each provides a different simplicity judgement: Simplicity Gain (Xu et al. 2016), Structural Simplicity (Sulem, Abend, and Rappoport 2018c), and Direct Assessments of Simplicity (new). This allows studying the behaviour of metrics along varied ways of measuring simplicity (Sec. 4.2).
2. Each includes system outputs from different types of simplification approaches. This allows analysing the impact of the system type in the correlation of metrics (Sec. 4.3). Table 1 presents brief descriptions of the most representative models in these datasets.
3. All original sentences come from TurkCorpus (Xu et al. 2016). This allows exploiting the alignment between TurkCorpus, HSplit (Sulem, Abend, and Rappoport 2018a) and ASSET (Alva-Manchego et al. 2020) to investigate

2 Štajner, Mitkov, and Saggion (2014) analysed several MT metrics without introducing a new resource, but focused on human judgements of grammaticality and meaning preservation.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 1
Descriptions of simplification systems included in the studied datasets. Similarly to Alva-Manchego, Scarton, and Specia (2020), we classified them into phrase-based MT (PBMT), syntax-based MT (SBMT), neural sequence-to-sequence (S2S), and semantics-informed rules (Sem) by themselves or coupled with one of the previous types (i.e. Sem+PBMT, Sem+S2S).

Type	Name	Description
PBMT	PBMT-R (Wubben, van den Bosch, and Krahmer 2012)	Phrase-based MT model that chooses the candidate simplification that is most dissimilar to the original sentence.
SBMT	SBMT-SARI (Xu et al. 2016) SBMT-BLEU (Xu et al. 2016) SBMT-FKBLEU (Xu et al. 2016)	Syntax-based MT models trained on paraphrases from the Paraphrase Database (Ganitkevitch, Van Durme, and Callison-Burch 2013) and tuned using SARI, BLEU or FKBLEU.
S2S	NTS (Nisioi et al. 2017)	Neural models with standard encoder-decoder architectures with attention.
	Dress-Ls (Zhang and Lapata 2017)	RNN-based encoder-decoder with attention combined with reinforcement learning.
	DMASS-DCSS (Zhao et al. 2018)	Transformer-based encoder-decoder (Vaswani et al. 2017) and memory-augmentation with paraphrasing rules from the Simple Paraphrase Database (Pavlick and Callison-Burch 2016).
	ACCESS (Martin et al. 2020)	Transformer-based encoder-decoder that conditions the generation of simplifications on explicit desired text attributes (e.g. length and/or dissimilarity with original input).
Sem	DSS (Sulem, Abend, and Rappoport 2018c)	Hand-crafted rules for sentence splitting based on either automatic or manual UCCA (Abend and Rappoport 2013) semantic annotations.
Sem+PBMT	Hybrid (Narayan and Gardent 2014)	Phrase-based statistical MT model coupled with semantic analysis to learn to split sentences.
Sem+S2S	SENTS (Sulem, Abend, and Rappoport 2018c)	Uses DSS for sentence splitting and then the resulting output goes through a MT-based model for further paraphrasing.

the effect in the correlations of using different sets of manual references when computing the metrics (Sec. 4.4).

Furthermore, we compare the datasets in terms of their human evaluation reliability using both Inter-Annotator Agreement (IAA) and correlation coefficients, as suggested in Amidei, Piwek, and Willis (2019). For IAA, we compute Intraclass Correlation (ICC, Shrout and Fleiss 1979) with the implementation available in `pingouin` (Vallat 2018).[3] For computing ratings correlations, we account for multiple annotators per instance by simulating two raters as follows: (1) we randomly choose one score as rater A and the average of the others as rater B; (2) we compute the Spearman's rank correlation coefficient between raters A and B using `SciPy`;[4] and (3) we repeat this process 1,000 times to report the mean and variance of all iterations. For interpreting the values of

3 https://pingouin-stats.org/generated/pingouin.intraclass_corr.html
4 https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.spearmanr.html

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

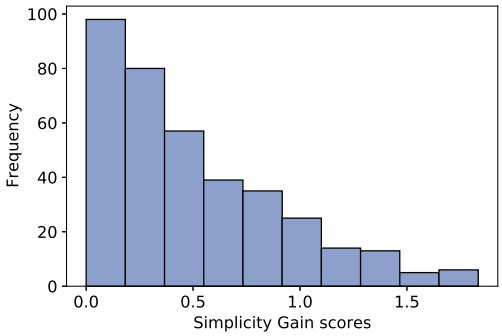

Figure 1
Distribution of Simplicity Gain scores in the dataset of (Xu et al. 2016).

both calculations, we use the scale of Landis and Koch (1977) for IAA and the scale of Rosenthal (1996) for nonparametric correlation coefficients.

3.1 Simplicity Gain Dataset

Xu et al. (2016) created this dataset to study the suitability of metrics for measuring the Simplicity Gain of automatic simplifications. The authors simplified 93 original sentences using four Sentence Simplification systems: PBMT-R, SBMT-BLEU, SBMT-FKBLEU and SBMT-SARI. For the Simplicity Gain judgements, workers on Amazon Mechanical Turk (AMT) were asked to count the number of *"successful lexical or syntactic paraphrases occurred in the simplification"* (Xu et al. 2016). The judgements from five different workers were averaged to get the final score for each instance. In order to measure human evaluation reliability, we computed an ICC of 0.176 and a Spearman's ρ of 0.299 ± 0.036. The ICC only points to *slight* agreement between the annotators, and the Spearman's ρ implies a *small* correlation between the human ratings.

This dataset has limitations that could prevent generalising findings based on its data. For instance, the number of evaluated instances (372) is small, and they were produced by only four automatic systems, three of which have very similar characteristics. In addition, as shown in Figure 1, the evaluated systems did not perform significant simplification changes (as judged by humans), since most instances were rated with Simplicity Gain scores below 1, with a high frequency of values between 0 and 0.25.

3.2 Structural Simplicity Dataset

Sulem, Abend, and Rappoport (2018c) created this dataset to evaluate the performance of Sentence Simplification models that mix hand-crafted rules (based on a semantic parsing) for sentence splitting, with standard MT-based architectures for lexical paraphrasing. Sulem, Abend, and Rappoport (2018a) further exploited this data to examine the suitability of BLEU for assessing Structural Simplicity.[5] The authors simplified

5 Sulem, Abend, and Rappoport (2018b) also collected a Structural Simplicity dataset, but used simplification instances from PWKP (Zhu, Bernhard, and Gurevych 2010), making it unsuitable for our analysis since: (1) it does not contain manual references, but automatic alignments to sentences from Simple Wikipedia; and (2) it is single-reference, which is unfair to reference-based metrics.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

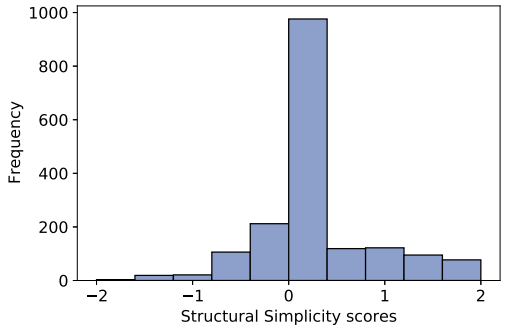

Figure 2
Distribution of Structural Simplicity scores in the dataset of (Sulem, Abend, and Rappoport 2018c).

70 sentences using 25 automatic systems: Hybrid; SBMT-SARI; four versions of NTS mixing initialisation with default or word2vec embeddings, and selecting the highest or fourth-best hypothesis according to SARI; two versions of DSS, with either automatic or manual semantic annotations; eight versions of SENTS that first use a version of DSS for sentence splitting and then the resulting output goes through a version of NTS; and many variations of SENTS where NTS is replaced by Moses (Koehn et al. 2007).

Native English speakers were asked to use a 5-point Likert scale (-2 to +2 scores) to measure Structural Simplicity: *"is the output simpler than the input, ignoring the complexity of the words?"* (Sulem, Abend, and Rappoport 2018c). The judgements from three different annotators are averaged to get the final score for each instance. Our computation of human evaluation reliability found an ICC of 0.465 and a Spearman's ρ of 0.508 ± 0.013. The ICC points to a *moderate* agreement between the annotators, and the Spearman's ρ implies a *medium* correlation between the human ratings.

Compared to the Simplicity Gain dataset, this one is bigger (1,750 instances) and with more variability in the system outputs collected. In addition, Figure 2 shows that the distribution of scores spans across all possible values, indicating that some systems even hurt the Structural Simplicity of the original sentence. Despite the over-representation of simplifications with scores between 0 and 0.5, around 32% of instances improve Structural Simplicity, indicating that an analysis based on perceived quality across different levels is possible.

3.3 The New Simplicity-DA Dataset

We introduce a new dataset with human judgements of simplification quality elicited via Direct Assessment (DA, Graham et al. 2017), a commonly-used methodology in Machine Translation Shared Tasks (Bojar et al. 2018; Barrault et al. 2019). Leveraging publicly-available system outputs on the test set of TurkCorpus (Xu et al. 2016), we collected simplifications from six systems: PBMT-R, Hybrid, SBMT-SARI, Dress-Ls, DMASS-DCSS, and ACCESS. For each system, we randomly sampled 100 automatic simplifications, not necessarily all from the same set of original sentences, but ensuring that the system output was not identical to the original sentence. Then, we crowd-sourced human ratings using AMT. Workers were asked to assess the quality of the automatic simplifications in three aspects: fluency, meaning preservation and simplicity. For each aspect, raters needed to submit a score between 0 and 100, depending on how

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

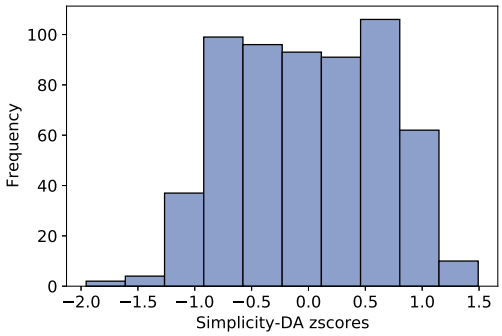

Figure 3
Distribution of Simplicity zscores in the newly-collected Simplicity-DA dataset.

much they agreed with a specific question. For simplicity, in particular, they were asked: *Rate your level of agreement to the statement: "The Simplified sentence is easier to understand than the Original sentence"*. This is inspired by the DA methodology and, thus, we refer to this kind of simplicity judgements as Simplicity-DA. Each HIT in AMT consisted of five sentences, with a maximum time of one hour for completion, and a payment of $0.5 per HIT. For quality control, workers had to pass a qualification test before participating in the rating task. All submissions to this test were manually reviewed to ensure undestandability of the instructions.[6] This crowdsourcing methodology is similar to the preliminary metrics' correlation study in (Alva-Manchego et al. 2020). However, our new Simplicity-DA dataset includes more automatic simplifications than those collected before (600 vs 100), allowing better generalisation of our findings.

For each simplification instance, we collected 15 ratings per quality aspect (fluency, meaning preservation and simplicity), which are then standardised by the mean and standard deviation of each worker to reduce individual biases. The average of all 15 standardised ratings (also called zscore) is the final score for the instance per quality aspect. Our computation of human evaluation reliability found an ICC of 0.386 and a Spearman's ρ of 0.607 $\pm$ 0.026. The ICC points to a *fair* agreement between the annotators, and the Spearman's ρ implies a *large* correlation between the human ratings.

The annotation reliability for the collected ratings in our dataset is higher than that for the Simplicity Gain dataset, and comparable to that of the Structural Simplicity dataset. In addition, our dataset is bigger in size and offers more variability of system outputs than the Simplicity Gain dataset. In particular, we included state-of-the-art neural sequence-to-sequence models, the current trend in automatic simplification systems. See Table 2 for a summary comparing the characteristics of the three datasets. Furthermore, Figure 3 shows that the Simplicity-DA ratings are more diversely-distributed across all scores values than the other datasets. This benefits our meta-evaluation since one of our intended dimensions of study is the perceived low or high quality (in terms of simplicity) of the system outputs. Overall, we argue that the newly-collected Simplicity-DA dataset provides a valid alternative view at human judgements of simplicity. In particular, it is more reliable for analysing automatic metrics in a multi-operation sim-

6 The Qualification Test and Rating Task, with the instructions given to the workers, can be found in: https://github.com/feralvam/metaeval-simplification/tree/main/HIT_designs.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 2
Summary of characteristics of the datasets with human ratings of simplicity used for the meta-evaluation study.

	Simplicity Gain	Structural Simplicity	Simplicity-DA
General Statistics			
Number of Instances	372	1,750	600
Ratings per Instance	5	3	15
Type of Rating	Discrete (count)	Discrete (Likert scale)	Continuous
System Types	PBMT, SBMT	PBMT, SBMT, S2S, Sem, Sem+PBMT, Sem+S2S	PBMT, SBMT, S2S, Sem+PBMT
Human Evaluation Reliability			
ICC	0.176	0.465	0.386
Spearman's ρ	0.299 ± 0.036	0.508 ± 0.013	0.607 ± 0.026

plification scenario since the judgements are not tied to the correctness of a specific rewriting operation.

4. Meta-Evaluation of Automatic Evaluation Metrics

In this Section, we study how the correlations between automatic scores and human judgements vary across different dimensions. Our investigation is inspired by research in Machine Translation evaluation. In particular, by the WMT Metrics Shared Tasks that compare standard and new metrics in a common setting using human judgements collected through Direct Assessment (Graham et al. 2017), primarily in the latest years (Bojar et al. 2016; Bojar, Graham, and Kamran 2017; Ma, Bojar, and Graham 2018; Ma et al. 2019; Mathur et al. 2020). Data from these WMT Shared Tasks has allowed to further study the behaviour of metrics at sentence level across different dimensions (Fomicheva and Specia 2019), to analyse the protocols for evaluating metrics at system level (Mathur, Baldwin, and Cohn 2020), to study the effect of the quality of references used to compute metrics (Freitag, Grangier, and Caswell 2020), among others.

In our study, we analyse the behaviour of automatic metrics at sentence level since the datasets described previously contain human judgements for each individual simplification instance. Also, metrics explicitly-developed to measure some form of simplicity, such as SARI and SAMSA, operate by-definition at the sentence-level.[7] Our meta-evaluation analyses the variation of correlations between automatic metrics with human judgements across three dimensions: the level of simplicity of the system outputs, the approaches used by the simplification systems, and the set of manual references used to compute the metrics.

4.1 Experimental Setting

Our study focuses on metrics developed to estimate the simplicity of system outputs, or that have been traditionally-used for this task:[8] BLEU, SARI, SAMSA, FKGL (Kin-

7 SARI has a corpus-level version that is commonly-reported to compare the performances of automatic systems. However, its authors only validated the correlation of SARI with human judgements at sentence-level, not system-level (https://github.com/cocoxu/simplification/issues/9).
8 Details of these and other metrics can be found in (Alva-Manchego, Scarton, and Specia 2020).

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

caid et al. 1975), FKBLEU (Xu et al. 2016) and iBLEU (Sun and Zhou 2012).[9] We also experiment with the arithmetic mean (AM) and geometric mean (GM) of BLEU-SARI and SARI-SAMSA. Finally, we include BERTScore (Zhang et al. 2020), a reference-based metric that computes the cosine similarity between tokens in a system output and in a manual reference using contextual embeddings, namely BERT (Devlin et al. 2019). This metric provides three types of scores: $\text{BERTScore}_{\text{Recall}}$ matches each token in the reference to its most similar in the system output, $\text{BERTScore}_{\text{Precision}}$ matches each token in the system output to its most similar in the reference, and $\text{BERTScore}_{\text{F1}}$ combines the two. When multiple references are available, BERTScore compares the system output against all references and returns the highest value. In the context of Sentence Simplification, a modified version of BERTScore has been used to create artificial data for training a model that ranks candidate simplifications, obtaining promising results (Maddela, Alva-Manchego, and Xu 2021).

We used the implementations of these metrics provided by EASSE (Alva-Manchego et al. 2019).[10] Most of the metrics are sentence-level by definition, with the exception of BLEU and derivations. In this case, we used a smoothed version with method `floor` and default value 0.0 in SacreBLEU (Post 2018).[11] For a fair comparison, we detokenised and recased all original sentences and system outputs in the three datasets. Then, we set EASSE to compute all metrics with the same configuration: tokenisation using SacreMoses[12] and case-sensitive calculation.

In order to compare the automatic evaluation metrics, we followed the methodology of recent editions of the WMT Metrics Shared Task (Ma, Bojar, and Graham 2018; Ma et al. 2019). First, we computed the correlations between automatic scores and human judgements via Pearson's r for each metric. Since the simplicity ratings in our human evaluation datasets are absolute instead of relative rankings between instances, this method is better suited and easier to apply than Kendall's Tau. Furthermore, we performed Williams significance tests (Williams 1959) to determine if the increase in correlation between two metrics is statistically significant or not.

4.2 Metrics across Simplicity Quality Levels

Our first dimension of analysis is the perceived quality of the automatic simplifications. We investigate whether it is easier or harder for metrics to evaluate low-quality or high-quality simplifications, as determined by their human judgements on simplicity. In order to do this, we split the instances in each dataset into two groups according to their simplicity score, and compute the Pearson's r between metrics and human judgements for the top 50% ("High"), the bottom 50% ("Low") and "All" available instances.

4.2.1 Simplicity-DA. Table 3 presents the correlations in each quality split of this dataset. Reference-based metrics were computed using manual simplifications from ASSET, since the Simplicity-DA judgement is not limited to a particular operation being performed, and simplifications in ASSET were created applying several of them.

When "All" instances are considered, $\text{BERTScore}_{\text{Precision}}$ shows a strong correlation with direct assessments of Simplicity, and no metric is better than it. Flesch-based

9 Even though FKGL is a document-level metric, we include it in our study following Xu et al. (2016).
10 https://github.com/feralvam/easse
11 https://github.com/mjpost/sacrebleu
12 https://github.com/alvations/sacremoses

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 3
Absolute Pearson correlations between **Simplicity-DA** and metrics scores computed using references from **ASSET**, for **low/high/all quality splits** (N is the number of instances in the split). Correlations of metrics not significantly outperformed by any other in the quality split are boldfaced.

	Metric	Low ($N = 300$)	High ($N = 300$)	All ($N = 600$)
Reference-based	$\text{BERTScore}_{\text{Precision}}$	0.512	0.287	**0.617**
	$\text{BERTScore}_{\text{Recall}}$	0.471	0.172	0.500
	$\text{BERTScore}_{\text{F1}}$	0.518	0.224	0.573
	BLEU	0.405	0.235	0.496
	iBLEU	0.398	0.253	0.504
	SARI	0.336	0.139	0.359
	BLEU-SARI (AM)	0.417	0.239	0.503
	BLEU-SARI (GM)	0.408	0.215	0.476
	SARI-SAMSA (AM)	0.203	0.050	0.166
	SARI-SAMSA (GM)	0.222	0.024	0.156
	FKBLEU	0.131	0.006	0.098
Non-Reference-based	FKGL	0.272	0.093	0.117
	SAMSA	0.103	0.010	0.058

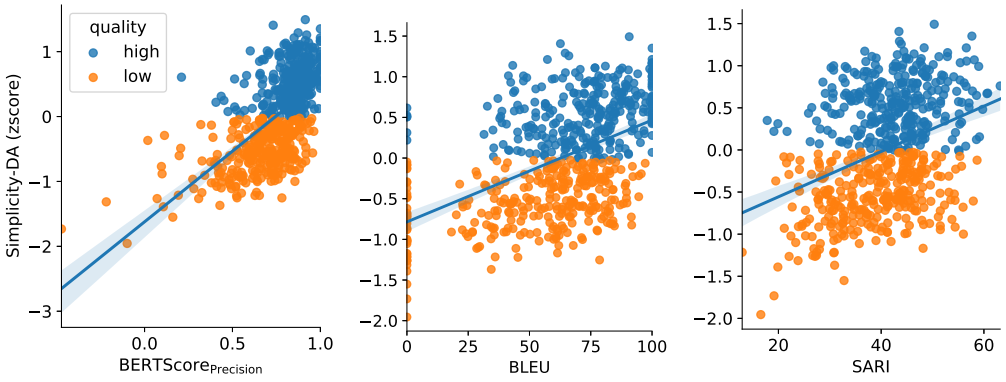

Figure 4
Scatter plots showing the correlation (r) between $\text{BERTScore}_{\text{Precision}}$, BLEU and SARI, with human rating of Simplicity-DA, for different quality levels.

metrics (FKGL and FKBLEU) have the lowest correlations, providing further evidence that these type of metrics are unsuitable for sentence-level evaluation. Simplification-specific metrics, SARI and SAMSA, also fair poorly. One possible explanation is that they were developed to assess the execution of particular simplification operations (lexical paraphrasing and sentence splitting, respectively), whilst the Simplicity-DA judgements are not operation-specific, but rather perceptions of general simplicity. Computing their arithmetic or geometric means does not yield good correlations in this dataset either. BLEU shows a moderate correlation, and combining it with SARI through arithmetic or geometric mean does not significantly improve the correlation with Simplicity-DA judgements in this dataset.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 4
Examples of original sentences with some of their simplification references in ASSET, and system outputs with corresponding human and automatic scores from the Simplicity-DA dataset. The reference selected by the automatic metric as most-similar to the system output is emphasised.

Original Sentence	In 1998, Culver ran for Iowa Secretary of State and was victorious.			
System Output	In 1998, Culver ran for Iowa Secretary of State.			
Sample References	Culver ran and won Iowa's secretary of State in 1998.			
	In 1998, Culver successfully ran for Iowa Secretary of State.			
	In 1998, Culver ran for Iowa Secretary of State. He won the election.			
	Simplicity-DA	0.551	**BERTScore$_{\text{Precision}}$**	0.984
Original Sentence	Below are some useful links to facilitate your involvement.			
System Output	Below is some useful links to help with your involvement			
Sample References	Here are good links to get you to do it.			
	Below are some useful links to help with your involvement.			
	Here are some useful links to help you.			
	Simplicity-DA	0.327	**BERTScore$_{\text{Precision}}$**	0.934
Original Sentence	He was appointed Companion of Honour (CH) in 1988.			
System Output	He was appointed Companion of Honour in 1988.			
Sample References	He was made the Companion of Honour (CH) in 1988.			
	He was appointed Companion of Honour in 1988.			
	In 1988 he was chosen as a Companion of Honour.			
	Simplicity-DA	0.436	**BERTScore$_{\text{Precision}}$**	1.000

When comparing the correlations between the "Low" and "High" splits, we can notice that the ones in the latter are much lower. This could be interpreted as: **low scores of some metrics indicate "bad" quality of a simplification (in terms of Simplicity-DA), but high scores do not necessarily imply "good" quality.** Figure 4 further illustrates this behaviour for three representative metrics. This could be explained by how (most of) the metrics assess the system outputs (i.e. by computing their similarity to the manual references), and by the question used to elicit Simplicity-DA judgements.

One possible reason is that simplifying a sentence may be limited to a few important changes that improve its readability (e.g. replacing some words or splitting a long sentence into two), whilst keeping the rest of the original sentence as-is. Not performing these key modifications or executing unnecessary ones would be penalised by the human judges, resulting in low Simplicity-DA scores. However, similarity-based metrics could still provide high scores that, in fact, are indicative of the overlap between the system output and the references due to some degree of meaning preservation, but not of the changes that improve simplicity. The first example in Table 4 illustrates this scenario. The reference selected by BERTScore$_{\text{Precision}}$ as the most similar to the system output is a clever simplification that uses the adverb "succesfully" to replace the clause "and was victorious" from the original sentence. Since the rest of the sentence is unchanged, it has a high overlap with the system output that merely deleted the "and was victorious" clause.

Finally, there could be a disagreement between the changes the human judges deemed necessary for a good Simplicity-DA score, and what the editors that created ASSET considered as valid simplifications. The second and third examples in Table 4 illustrate this scenario. The selected references are almost identical to the corresponding system outputs, and thus BERTScore$_{\text{Precision}}$ scored them very high. However, the

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 5
Absolute Pearson correlations between **Simplicity Gain** and metrics scores computed using references from **TurkCorpus**, for **low/high/all quality splits** (N is the number of instances in the split). Correlations of metrics not significantly outperformed by any other in the quality split are boldfaced.

	Metric	Low ($N = 186$)	High ($N = 186$)	All ($N = 372$)
Reference-based	$\text{BERTScore}_{\text{Precision}}$	0.209	0.231	0.241
	$\text{BERTScore}_{\text{Recall}}$	0.221	0.217	0.241
	$\text{BERTScore}_{\text{F1}}$	0.215	0.236	0.247
	BLEU	0.178	0.132	0.123
	iBLEU	0.181	0.136	0.128
	SARI	0.292	0.240	**0.331**
	BLEU-SARI (AM)	0.223	0.172	0.187
	BLEU-SARI (GM)	0.246	0.177	0.214
	FKBLEU	0.041	0.007	0.092
Non-Reference-based	FKGL	0.045	0.101	0.147
	SAMSA	0.120	0.042	0.013

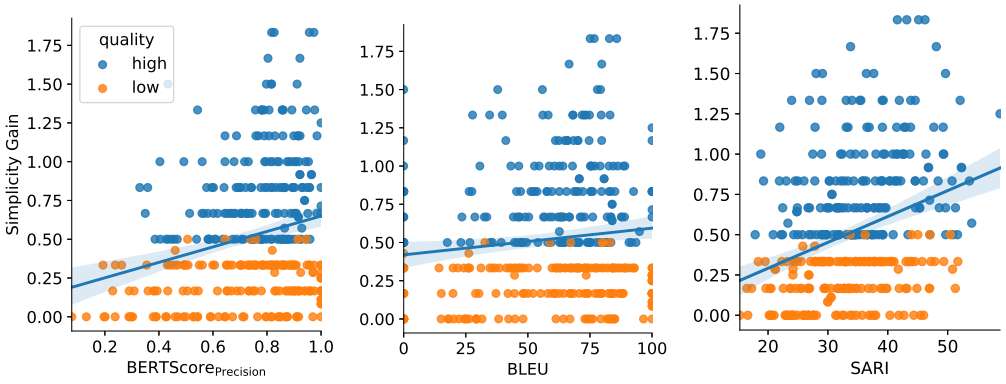

Figure 5
Scatter plots showing the correlation (r) between $\text{BERTScore}_{\text{Precision}}$, BLEU and SARI, with human rating of Simplicity Gain, for different quality levels.

human judges considered the changes insufficient to grant a high value of Simplicity-DA for improved simplicity. This may not be indicative that references in ASSET are incorrect, but rather that not all of them have the same degree of simplicity.

4.2.2 Simplicity Gain. Table 5 presents the correlations in each quality split of this dataset. Reference-based metrics were computed using manual simplifications from TurkCorpus, since the Simplicity Gain judgement is limited to counting lexical paraphrases, and references in TurkCorpus were created by only applying that operation.

In this dataset, SARI has a moderate correlation, and the highest among all metrics when "All" evaluation instances are considered, similar to the results in (Xu et al. 2016). Just like in the Simplicity-DA dataset, Flesch-based metrics and SAMSA show low correlations, whilst BLEU and its variants have correlations in the middle of the group. The different versions of BERTScore are second-best, and have similar performances, i.e.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 6
Examples of original sentences and system outputs with corresponding human and automatic scores from the Simplicity Gain dataset. Changes related to lexical paraphrasing are boldfaced.

Original Sentence	Jeddah is the **principal** gateway to Mecca, Islam's holiest city, which able-bodied Muslims **are required to** visit at least once in their **lifetime**.			
System Output	Jeddah is the **main** gateway to Mecca, Islam's holiest city, which sound Muslims **must** visit at least once in **life**.			
	Simplicity Gain	1.83	**SARI**	0.462
Original Sentence	The Great Dark Spot is thought to **represent** a hole in the methane cloud deck of Neptune.			
System Output	The Great Dark Spot is thought to **be** a hole in the methane cloud deck of Neptune.			
	Simplicity Gain	1.25	**SARI**	0.587

there is no statistically-significant difference between them. Also, combining SARI with BLEU does not improve its individual correlation. When comparing the correlations between the "Low" and "High" quality splits (also see Figure 5), most metrics have lower Person's r in "High". However, this is not a consistent behaviour, and the differences are not as considerable as observed in the Simplicity-DA dataset.

We hypothesise that the overall moderate-to-low correlations is due to most of the metrics not directly measuring Simplicity Gain. Almost all metrics compute the similarity between the system output and the references. However, measuring Simplicity Gain implies identifying the changes made by the system, and then verifying that they are correct. In order to do this, it is necessary to take the original sentence into consideration, and not just the system output and the references. SARI is the only metric that attempts to follow this logic, by computing the correctness of the n-grams kept, deleted and added. Lexical paraphrasing is however strongly related to performing replacements, an operation that SARI does not directly identify and measure. The examples in Table 6 show how this limitation hurts the metric: whilst in the second instance there are fewer correct replacements than in the first one ($1 < 3$), the SARI score is higher ($0.587 > 0.462$). By not directly counting correct replacements, the metric is affected by the conservative nature of the outputs and references that copy most of the original sentences. It is the correctness of kept and deleted n-grams that contributes to getting a high score. Consequently, SARI is not measuring Simplicity Gain, which explains why the correlation with human judgements is barely moderate.[13]

The concept of Simplicity Gain is easy to understand: it is the number of correct changes. If metrics were able to measure it accurately, automatic scores would be more straightforward to interpret, facilitating the comparison of simplifications generated by different systems. However, collecting this type of human judgements is difficult, especially in instances where multiple rewriting operations may have been applied, and identifying where the changes happened (and counting them) is not trivial. In addition, the Simplicity Gain dataset from Xu et al. (2016) that we use in this study is quite small (only 372 evaluated instances), and contains automatic simplifications from only four

13 While the examples in Table 6 were cherry-picked, the low overall correlation shows that SARI and Simplicity Gain are not measuring the same thing. The examples are only illustrative of where this is the case. A more in-depth analysis is left for future work.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 7
Absolute Pearson correlations between **Structural Simplicity** and metrics scores computed using references from **HSplit**, for **low/high/all quality splits** (N is the number of instances in the split). Correlations of metrics not significantly outperformed by any other in the quality split are boldfaced.

	Metric	Low ($N = 875$)	High ($N = 875$)	All ($N = 1,750$)
Reference-based	$\text{BERTScore}_{\text{Precision}}$	**0.552**	0.310	0.090
	$\text{BERTScore}_{\text{Recall}}$	0.411	0.601	0.430
	$\text{BERTScore}_{\text{F1}}$	0.483	0.529	0.325
	BLEU	0.421	**0.643**	0.443
	iBLEU	0.408	0.635	0.436
	SARI	0.137	0.418	0.313
	BLEU-SARI (AM)	0.346	0.599	0.431
	BLEU-SARI (GM)	0.329	0.589	0.438
	BLEU-SAMSA (AM)	0.289	0.608	0.420
	BLEU-SAMSA (GM)	0.293	0.569	0.370
	FKBLEU	0.395	0.608	0.364
Non-Reference-based	FKGL	0.070	0.165	0.228
	SAMSA	0.103	0.431	0.284

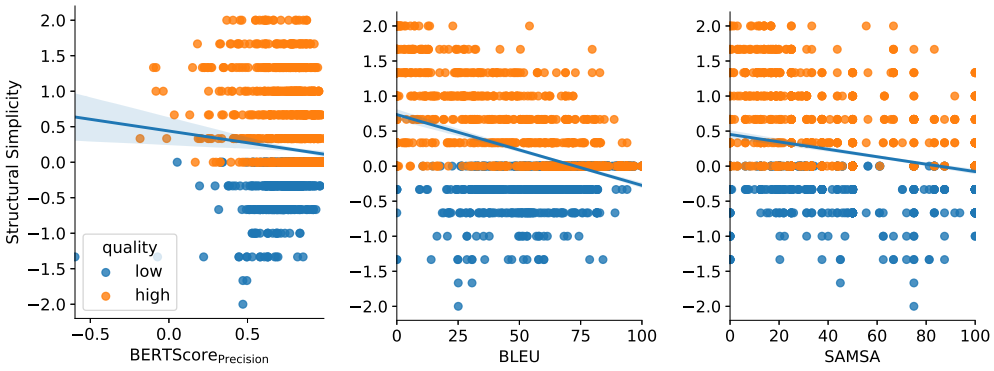

Figure 6
Scatter plots showing the correlation (r) between $\text{BERTScore}_{\text{Precision}}$, BLEU and SAMSA, with human rating of Structural Simplicity, for different quality levels.

systems, three of which are of similar characteristics (SBMT-based), without any current state-of-the-art neural models. All of this impedes generalisations that could be relevant in Sentence Simplification research.

4.2.3 Structural Simplicity. Table 7 presents the correlations in each quality split of this dataset. Reference-based metrics were computed using manual simplification from HSplit, since the Structural Simplicity judgement is limited to qualifying sentence splitting, and references in HSplit were created by only applying that operation.

In this dataset, most metrics have moderate correlations with human judgements when "All" evaluated instances are used. BLEU obtains the highest correlation, but its not the best overall since its differences with BLEU-SARI (GM) and $\text{BERTScore}_{\text{Recall}}$ are

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 8
Examples of original sentences and system outputs with corresponding human and automatic SAMSA scores from the Structural Simplicity dataset.

Original Sentence	Orton and his wife welcomed Alanna Marie Orton on July 12 2008.			
System Output	Orton and his wife welcomed Alanna Marie Orton on July 12 2008.			
	Structural Simplicity	0.00	**SAMSA**	1.00
Original Sentence	Graham attended Wheaton College from 1939 to 1943, when he graduated with a BA in anthropology.			
System Output	Graham attended Wheaton College from 1939 to 1943. He graduated with a BA in anthropology.			
	Structural Simplicity	0.33	**SAMSA**	1.00
Original Sentence	Jeddah is the principal gateway to Mecca, Islam's holiest city, which able-bodied Muslims are required to visit at least once in their lifetime.			
System Output	Jeddah is the principal gateway to Mecca.			
	Structural Simplicity	2.00	**SAMSA**	0.14

not statistically significant. This would seem to contradict the findings of Sulem, Abend, and Rappoport (2018a), who argued that BLEU does not correlate well with Structural Simplicity. However, as will be shown in the next Section, the magnitude of the correlation depends on the approach of the systems included in the study. Whilst Sulem, Abend, and Rappoport (2018a) only used models tailored for sentence splitting to reach that conclusion, in this first analysis we are using all available system outputs in the dataset. The low correlation of SAMSA is surprising, since this metric was specifically-designed to evaluate sentence splitting, and it showed better performance in the dataset of Sulem, Abend, and Rappoport (2018b). However, they measured the correlation at the system-level, whilst we are analysing it at the sentence-level. Finally, $\text{BERTScore}_{\text{Precision}}$, the best metric in the Simplicity-DA dataset, has the poorest correlation in the "All" data split. From previous results, we know that $\text{BERTScore}_{\text{Precision}}$ is good at measuring the similarity between a system output and a reference. As such, its low correlation would indicate that simple similarity matching is not enough to measure Structural Simplicity.

When comparing the correlations between the "Low" and "High" splits (also see Figure 6), we can notice that the ones in the former are much lower for all metrics but $\text{BERTScore}_{\text{Precision}}$. In fact, this metric has the highest correlation in the "Low" split, with a substantial increase over its own correlation in the "All" data split. This could also be explained by our previous argument. A low score in Structural Simplicity implies that the system output does not contain any sentence splitting, or that the changes made are not structural. In these situations, $\text{BERTScore}_{\text{Precision}}$ would not be able to match a reference in HSplit, since they most likely contain only sentence splitting. In turn, the metric returns a low score that correlates well with a low human judgement.

We further analyse the behaviour of SAMSA, a metric specifically-designed to evaluate Structural Simplicity. By design, SAMSA first uses a semantic parser to identify the Scenes in the original sentence, and a syntactic parser to identify the sentence splits in the system output. Then, it counts how many of the words corresponding to the Participants of each Scene align with words in each sentence split. Ideally, all Participants of a single Scene should appear in a single sentence split. The first example in Table 8 illustrates a case where this logic may be problematic. SAMSA identifies that there is only one Scene in the original sentence and only one sentence split in

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

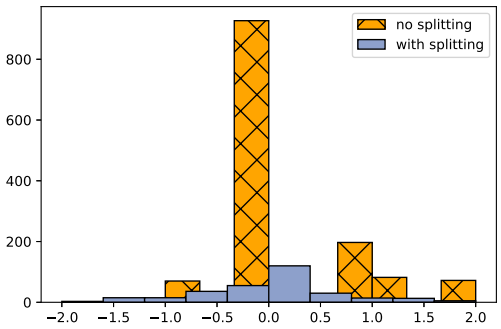

Figure 7
Distribution of Structural Simplicity scores in the dataset of (Sulem, Abend, and Rappoport 2018c) for instances with and without sentence splitting in the system output.

the system output. Since both sentences are identical, the word alignment is perfect and SAMSA gives the simplification the highest possible score. However, the human judges gave the instance a score of 0 since no changes were performed. On the one hand, this could suggest that SAMSA should only be used when sentence splitting was actually performed in the simplification instance. On the other hand, it could be argued that the original sentence was already structurally simple, and that no splitting was necessary, making the human score of 0 unfair. This points out to possible issues in the data collection, and that perhaps using a -2 to +2 scale is unsuitable for these scenarios.

We further explore our last argument of potential incompatibilities between what Structural Simplicity should measure, and what the human judges qualified as such. The second and third examples in Table 8 suggest that there are indeed problems. The second example shows that a perfectly-reasonable and correct splitting (with a SAMSA score of 1.0) received a low score from the judges. More worryingly, the third example presents a sentence where no splitting was performed (and with substantial compression) that received the highest score for Structural Simplicity. This could indicate that the human judges did not consider sentence splitting as the only mechanism for improving the simplicity of the structure of a sentence. In an attempt to quantify this phenomenon, Figure 7 presents the distribution of Structural Simplicity scores for instances where sentence splitting was performed and where it was not. Instances with splitting only amount to 17% (306/1,750) of the total of instances in the dataset. Whilst this is a low quantity, their human scores span along all possible values for Structural Simplicity. It is encouraging that most instances where no splitting was performed received a human score close to 0. However, there are many that were judged with high values of Structural Simplicity. We hypothesise that this is caused by misunderstanding of the rating instructions, since many of these instances also contain substantial levels of compression (as in the third example of Table 8), which could not an type of rewriting that improves the structural simplicity of a sentence.

Improvement in Structural Simplicity is a relevant feature to evaluate in automatic simplifications. Isolating its assessment both manually and through metrics can contribute to a more fine-grained analysis of the performance of automatic systems. However, it is important to establish adequate quality control mechanisms that ensure the trustworthiness of the collected data, so that we can develop metrics that accurately resemble the intended human judgements.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 9
Pearson correlations between **Simplicity-DA** human judgements and automatic metrics scores computed using references from **ASSET**, for **splits based on system type** (N is the number of instances in the split). Correlations of metrics not significantly outperformed by any other in the system type split are boldfaced. Metrics are grouped in Reference-based (top) and Non-Reference-based (bottom).

Metric	SBMT ($N = 100$)	PBMT ($N = 100$)	S2S ($N = 300$)	Sem+PBMT ($N = 100$)
$\text{BERTScore}_{\text{Precision}}$	0.537	0.459	**0.650**	**0.624**
$\text{BERTScore}_{\text{Recall}}$	0.527	0.375	0.484	0.470
$\text{BERTScore}_{\text{F1}}$	0.528	0.400	0.588	0.568
BLEU	0.295	0.347	0.546	0.333
iBLEU	0.315	0.336	0.536	0.335
SARI	0.228	0.173	0.310	0.240
BLEU-SARI (AM)	0.315	0.336	0.536	0.335
BLEU-SARI (GM)	0.298	0.320	0.508	0.308
SARI-SAMSA (AM)	0.243	0.121	0.209	0.291
SARI-SAMSA (GM)	0.250	0.080	0.190	0.333
FKBLEU	0.006	0.058	0.092	0.138
FKGL	0.055	0.063	0.104	0.062
SAMSA	0.184	0.067	0.126	0.248

4.3 Metrics across Types of Systems

We now investigate if metrics' correlations are affected by the type of system that generated the simplifications. For this study, we do not use the Simplicity Gain dataset since it only provides simplifications produced by PBMT and SBMT systems.

4.3.1 Simplicity-DA. Table 9 presents the correlations of each metric for the different system types in this dataset, with reference-based metrics computed using simplifications from ASSET. $\text{BERTScore}_{\text{Precision}}$ achieves the highest correlations in all groups, and for S2S and Sem+PBMT models, in particular, no other metric is statistically-equal. Most metrics show higher correlations in the S2S group than in others. However, because the number of data points is smaller in the latter, stronger conclusions cannot be formulated. Overall, since the current trend is to develop S2S models, it is encouraging that modern metrics are capable of evaluating them, but keeping in mind the nuances we signalled in the previous Section regarding quality levels.

4.3.2 Structural Simplicity. Table 10 presents the correlations of each metric in the different system type groups in this dataset. Reference-based metrics were computed using manual simplifications from HSplit. All metrics achieve their highest correlations in the S2S group, except for $\text{BERTScore}_{\text{Precision}}$. As presented in the previous Section, this metric is particularly good at judging instances with low Structural Simplicity, which seem to be those from the PBMT and SBMT groups, mainly.

Previously, we observed that BLEU had high correlation with high-scoring quality judgements (in terms of Structural Simplicity). Here, we notice that this behaviour is limited to simplifications produced by S2S and Sem+S2S systems. This appears to contradict the observations of Sulem, Abend, and Rappoport (2018a), who used this same dataset to conclude that BLEU is a bad estimator of Structural Simplicity. The reason

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 10
Pearson correlations between **Structural Simplicity** judgements and automatic metrics scores computed using references from **HSplit**, for **splits based on system type** (N is the number of instances in the split). Correlations of metrics not significantly outperformed by any other in the system type split are boldfaced. Metrics are grouped in Reference-based (top) and Non-Reference-based (bottom).

Metric	PBMT ($N = 70$)	SBMT ($N = 70$)	S2S ($N = 280$)	Sem ($N = 140$)	Sem+PBMT ($N = 630$)	Sem+S2S ($N = 560$)
$\text{BERTScore}_{\text{Precision}}$	**0.501**	**0.571**	0.292	**0.330**	0.096	0.111
$\text{BERTScore}_{\text{Recall}}$	0.339	0.418	0.635	0.066	0.134	0.480
$\text{BERTScore}_{\text{F1}}$	0.405	0.497	0.553	0.180	0.049	0.362
BLEU	0.284	0.380	0.661	0.130	0.147	0.540
iBLEU	0.252	0.380	0.642	0.130	0.145	0.536
SARI	0.015	0.286	0.330	0.028	0.166	0.355
BLEU-SARI (AM)	0.184	0.364	0.603	0.100	0.175	0.507
BLEU-SARI (GM)	0.157	0.341	0.589	0.097	0.185	0.515
BLEU-SAMSA (AM)	0.240	0.334	0.603	0.095	0.072	0.573
BLEU-SAMSA (GM)	0.216	0.279	0.563	0.109	0.075	0.561
FKBLEU	0.215	0.344	0.617	0.009	0.119	0.539
FKGL	0.205	0.016	0.251	0.083	0.155	0.242
SAMSA	0.141	0.177	0.368	0.052	0.009	0.497

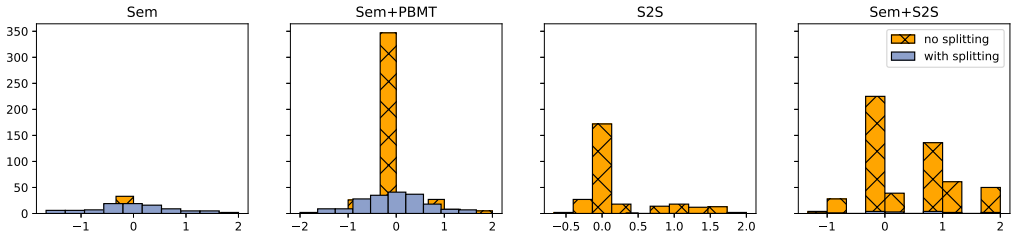

Figure 8
Distribution of Structural Simplicity scores in the dataset of (Sulem, Abend, and Rappoport 2018c) for instances with and without sentence splitting in the system output and for each system type.

behind this disagreement is that for their sentence-level study "HSplit as Reference Setting", the systems they chose were those within the Sem and Sem+PBMT groups, for which BLEU, indeed, shows poor correlations. A possible reason for choosing this setup is explained by Figure 8. Whilst S2S and Sem+S2S have more instances that were scored with good Structural Simplicity, these groups contain very few system outputs where sentence splitting was performed. Therefore, we believe that Sulem, Abend, and Rappoport (2018a)'s conclusion should be more nuanced: *BLEU is a bad metric to estimate Structural Simplicity in system outputs where sentence splitting was performed*.

Nevertheless, not considering system outputs in the S2S and Sem+S2S groups reduces the future impact of the previous statement, since the current trend in Sentence Simplification research is developing that type of models. For their system-level study "Standard Reference Setting", Sulem, Abend, and Rappoport (2018a) included systems from the S2S group, but computed BLEU using references from Simple Wikipedia and TurkCorpus, which are not focused on sentence splitting. We believe that this experi-

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Table 11
Pearson correlations between **Simplicity-DA** judgements and reference-based metrics scores **grouped by the set of manual references used**. Within each group, we divide the data into Low/High/All quality splits. Correlations of metrics not significantly outperformed by any other in their group and quality split are boldfaced. The scores in the left-hand side (under ASSET) are the same ones as in Table 3.

Metric	ASSET			All References			Selected References		
	Low	High	All	Low	High	All	Low	High	All
$\text{BERTScore}_{\text{Precision}}$	0.512	0.287	**0.617**	0.541	0.280	**0.629**	0.543	0.276	**0.635**
$\text{BERTScore}_{\text{Recall}}$	0.471	0.172	0.500	0.476	0.165	0.506	0.479	0.165	0.511
$\text{BERTScore}_{\text{F1}}$	0.518	0.224	0.573	0.530	0.202	0.576	0.534	0.202	0.584
BLEU	0.405	0.235	0.496	0.404	0.230	0.526	0.402	0.223	0.525
iBLEU	0.398	0.253	0.504	0.398	0.250	0.537	0.396	0.244	0.536
SARI	0.336	0.139	0.359	0.366	0.097	0.353	0.352	0.115	0.350
BLEU-SARI (AM)	0.417	0.239	0.503	0.418	0.218	0.519	0.418	0.221	0.523
BLEU-SARI (GM)	0.408	0.215	0.476	0.410	0.195	0.490	0.410	0.205	0.496

mental setting is unfair to BLEU, and that more cautious analysis should be performed to determine if a metric should be used to assess Structural Simplicity in S2S models.

4.4 Effect of Simplification References

The third dimension of analysis for our meta-evaluation is the set of simplification references used to compute automatic evaluation scores. Since there can be multiple correct simplifications for the same original sentence, it is possible that a reference-based metric becomes more reliable if it has access to more manual references for comparison. It is worth remembering that whilst BLEU and SARI take all references for each original sentences into account when computing their scores, BERTScore takes one at a time and returns the maximum score. In this Section, we investigate whether the correlations of reference-based metrics vary depending on using all available simplification references or particular subsets of them. We only experiment with the Simplicity-DA dataset, because its simplicity judgements are not tied to performing a specific type of simplification operation, as is the case for the other datasets. Thus, having a more varied set of references could be beneficial for reference-based metrics in this scenario. In addition, we take advantage of the fact that the original sentences in the Simplicitiy-DA dataset have corresponding manual simplifications in three multi-reference datasets: ASSET (10 references), TurkCorpus (8 references) and HSplit (4 references). Recall that the manual simplifications in each dataset were produced via different operations: lexical paraphrasing in TurkCorpus; sentence splitting in HSplit; and lexical paraphrasing, compression, and sentence splitting in ASSET.

4.4.1 ASSET vs All References. Table 11 presents the correlations of each reference-based metric computed using the 10 manual references from ASSET or their union with those from TurkCorpus (8 references) and HSplit (4 references), i.e. what we refer to "All References" (22). We further divide this data into "Low", "High" and "All" quality splits as in a previous Section. As such, the left-hand side of Table 11 is the same as Table 3. We do not add the system type dimension since the number of instances in each subgroup would be too small to allow drawing strong conclusions.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

When using "All" instances, most metrics have a slight increase in their Pearson's r when All References are used, with BERTScore$_{\text{Precision}}$ achieving the highest correlations, and being statistically superior to every other metric. This improvement seems to be caused by better detection of "Low" quality simplifications. In fact, using All References slightly affects BERTScore$_{\text{Precision}}$ and most metrics when detecting system outputs of "High" Simplicity-DA. As in a previous Section, we hypothesise that this is caused by the different degrees of simplicity that each manual reference has in each dataset. By having more references available, BERTScore$_{\text{Precision}}$ is more-likely to match one with a system output, and then return a high score. However, high similarity with a reference does not necessarily mean high improvements in simplicity, since the manual reference could correspond to a valid simplification but with a relatively-low degree of simplicity.

4.4.2 ASSET vs Selected References. In the previous analysis, we changed the set of references for all sentences that are being assessed at the same time. We now analyse the effect of **changing the set of references for each sentence individually**. More concretely, we devise an experiment where, for each automatically-simplified sentence, reference-based metrics compare it to a subset of all available references based on the simplification operations that were performed. Therefore, for each sentence:

1. **Identify the operations that were performed.** We use the annotation algorithms in EASSE to label deletions, replacements and splits at the sentence-level. For deletions and replacements, these algorithms leverage automatic word alignments between the original sentence and the automatic simplification, extracted using SimAlign (Jalili Sabet et al. 2020). If a word in the original sentence is aligned but not to an exact match in the simplification, then it is considered a replacement. If a word in the original sentence is not aligned, then it is considered as deleted. For identifying splits, we compute the number of sentences in the original and simplified sides using NLTK,[14] and register a split if the number in the simplified side is higher than the one in the original side. In preliminary experiments with a sample of 250 sentences, these algorithms achieved F1 scores of 0.76 for deletions, 0.78 for replacements, and 0.87 for splits. More details can be found in (Alva-Manchego 2020, chap 3).
2. **Determine the references to use.** Based on the operations identified, we treat three possible cases: (1) the system performed only sentence splitting; (2) the system performed only lexical paraphrasing and/or deletion; (3) the system performed another possible combination of operations.[15] Depending on the case, a different set of references would be used: HSplit for (1), TurkCorpus and ASSET for (2), and ASSET for (3). ASSET was added for case (2) since it also contains manual references where only lexical paraphrasing was applied.
3. **Compute the metrics.** Calculate the metrics' scores using the selected set of manual references.

14 https://www.nltk.org/api/nltk.tokenize.html
15 There is one more possible case: (4) the system did not perform any operation (i.e. the original sentence and the system output are the same). However, there are no such instances in the Simplicity-DA dataset.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Column "Selected References" in Table 11 presents the correlations of reference-based metrics computed following the previous process. All metrics but SARI improve their correlations when instances of "All" qualities are used. As before, this is caused by better detection of "Low" quality simplifications.

5. Recommendations for Automatic Evaluation

Our meta-evaluation has allowed us to better understand the behaviour of traditional and more modern metrics for assessing automatic simplifications. Based on those findings, in this Section we set a list of recommendations related to the present and future of automatic evaluation of Sentence Simplification systems.

5.1 Evaluation of Current Simplification Systems

Automatic Metrics. It is difficult to determine an overall "best" metric across all types of simplicity judgements. For Simplicity-DA, $\text{BERTScore}_{\text{Precision}}$ achieved the highest correlations in all dimensions of analysis. For Simplicity Gain, SARI is better than all BERTScore variants, but that difference is not statistically significant when assessing low and high quality simplifications separately. In addition, there is not enough data to determine if that behaviour translates to modern sequence-to-sequence models. The comparison is even less clear for Structural Simplicity, since the correlations are heavily dependent on the system type or, rather, on evaluating simplifications where sentence splitting was actually performed, instances of which are insufficient in the dataset used. SAMSA was specifically-developed for this type of simplicity evaluation, and manual inspection suggests that it is doing what it was designed for. As such, even though our analysis does not seem to support its use, we argue that this is caused by the lack of adequate data with judgements on Structural Simplicity. Overall, we suggest to use multiple metrics, and mainly $\text{BERTScore}_{\text{Precision}}$ for reference-based evaluation. SARI could be used when the simplification system only executed lexical paraphrasing, whilst SAMSA may be useful when it is guaranteed that splitting was performed.

Simplification References. Simplifications in ASSET are well suited for reference-based evaluation. Incorporating references from TurkCorpus and HSplit seems to only slightly improve the correlations. In addition, it appears that selecting which references to use for each sentence individually benefits the computation of metrics. However, for both cases, the improvements are limited to evaluation of low-quality simplifications.

Interpretation of Automatic Scores. For Simplicity-DA, low scores of most metrics appear to be good estimators of low quality, whilst high scores do not necessarily suggest high quality. This indicates that metrics could be more useful for development stages of simplification models. Following the recommendation of using multiple metrics, we suggest to use $\text{BERTScore}_{\text{Precision}}$ to get a first evaluation. If the score is low, then it signals that the quality of the output is also low. However, when the score is high, it is important to look at other metrics, such as SARI or SAMSA, to verify the correctness of the simplification operations. Nevertheless, for final arguments on the superiority of one system over another, human evaluation should be preferred. For Simplicity Gain, metrics' correlations are low to moderate in general, so it is unclear if they are actually measuring this type of human judgement. In the case of Structural Simplicity, inconsistencies in the human judgements (i.e. high scores for instances where no splitting was performed) hinders the interpretation of results.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

5.2 Development of New Metrics

Considering the advantages and disadvantages of current metrics, as well as the problems identified in the data used for evaluating them, we provide some suggestions for the development of new resources for automatic evaluation.

Collection of New Human Judgements. We experimented with crowdsourcing simplicity judgements following a methodology inspired by Direct Assessment, which has been successful in Machine Translation research. We believe that submitting continuous scores on how much simpler a system output is over the original sentence gives raters more flexibility on their judgements, and facilitates subsequent analyses. However, whilst the type of score collected (continuous or discrete) influences the ratings, it is even more important to ensure that raters submit judgements that follow the kind of simplicity that is intended to be measured. As such, it is paramount to train raters before they perform the actual task, and establish quality control mechanisms through out the data collection process. In relation to the kind of simplicity judgement to elicit, both Simplicity Gain and Structural Simplicity have advantages over requesting absolute simplicity scores. Therefore, we recommend to collect more human judgements based on them, using modern simplification models and simplification instances with adequate characteristics for what we are trying to evaluate.

Characteristics of New Metrics. For Simplicity-DA, Simplicity Gain and Structural Simplicity, raters had to compare the automatic simplification to the original sentence, and then submit a particular kind of judgement. Therefore, if humans submit evaluations taking both the original sentence and the simplification into consideration, then we should expect that automatic metrics do so too. Both SARI and SAMSA follow this logic, and we would expect that new metrics take that idea even further. For example, by replacing n-gram matching in SARI and syntax-based word alignments in SAMSA by similarity of contextual word embeddings, as is done in BERTScore. Furthermore, we have explained that not every manual simplification in multi-reference datasets (i.e. ASSET, TurkCorpus and HSplit) has the same simplicity level. Therefore, it could be useful to enrich references with human judgements on their simplicity. In this way, an automatic score would not be only based on the similarity to a reference, but also on the potential level of simplicity that the system output could achieve if it were an exact match with that particular reference. Perhaps, metrics could even combine how similar the system output is to a reference with the simplicity level that could be achieved.

Analysis Beyond Absolute Correlations. Our meta-evaluation has shown that different factors influence the correlation of human judgements with automatic scores, namely: perceived quality level, system type, and set of references used for computation. As such, new automatic metrics should not only be evaluated on their absolute overall correlation. It is important to also analyse the reasons behind that value considering the different factors that could be affecting it. In this way, we can determine in which situations the new metrics prove more advantageous than others.

6. Conclusions

In this article, we studied the degree in which current evaluation metrics measure the ability of automatic systems to perform sentence-level simplifications, especially when multiple operations were applied.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

We collected a **new dataset for evaluation of automatic metrics** following the Direct Assessment methodology to crowdsource human ratings on fluency, meaning preservation and simplicity. The dataset consists of **600 automatic simplifications generated by six different systems**, three of which are based on modern neural sequence-to-sequence architectures. This makes it bigger and more varied than the Simplicity Gain dataset. In addition, we collected **15 ratings per simplification instance** to increase annotation reliability, contrasting with the Simplicity Gain dataset that has five raters, and the Structural Simplicity dataset that only has three. Our data collection process can be fine-tuned, and more system outputs should be included. However, our dataset's features are sufficient to offer an alternative view at simplicity judgements over system outputs.

We used our newly-collected dataset (Simplicity-DA), together with the Simplicity Gain and Structural Simplicity datasets to conduct, to the best of our knowledge, **the first meta-evaluation study of automatic metrics in Sentence Simplification**. We analysed the variations of the correlations of sentence-level metrics with human judgements along three dimensions: the perceived simplicity level, the system type, and the set of references used to compute the automatic scores. For the first dimension, we found that **metrics can more reliably score low-quality simplifications** in terms of Simplicity-DA, whilst this effect is not apparent in Simplicity Gain and no strong conclusions could be drawn for Structural Simplicity due to inconsistencies in the ratings. For the second dimension, **correlations change based on the system type**. In the Simplicity-DA dataset, most metrics are better at scoring system outputs from neural sequence-to-sequence models. Whilst this difference in correlation is more significant in the Structural Simplicity dataset, it seems to be caused by low representation of sentence splitting in the data, rather than differences in system type. This highlights the importance of analysing outputs of several types of systems (e.g. neural and non-neural) with all the characteristics under study (e.g. split sentences), to prevent obtaining conclusions that are limited to a certain subgroup of models. For the third dimension, **combining all multi-reference datasets does not significantly improve metrics' correlations over using only ASSET** in the Simplicity-DA dataset. Further analyses on the diversity of the manual references across ASSET, TurkCorpus and HSplit should be performed in order to explain this result. In addition, preliminary experiments on per-sentence reference selection based on the performed operations showed promising results.

Based on the findings of our meta-evaluation, we designed a set of **guidelines for automatic evaluation of current simplification models**. In particular for multi-operation simplifications, we suggest to use BERTScore with references from ASSET during the development stage of simplification models, and manual evaluation for final comparisons. The main reason is that BERTScore is very good at identifying references that are similar to a system output. However, since not all references have the same simplicity level, a high similarity with a reference does not necessarily indicate high (improvements in) simplicity. Finally, we proposed a **desiderata for the characteristics of new resources for automatic evaluation**. Namely: (1) to further collect Simplicity Gain and Structural Simplicity ratings with better quality controls and diversity of system outputs; (2) to develop metrics that take both the original sentence and the automatic simplification into consideration, (3) to enrich manual references with their simplicity level; and (4) to evaluate new metrics along several dimensions and not just overall absolute correlation with human ratings of (some form) of simplicity.

References

Abend, Omri and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

228–238, Association for Computational Linguistics, Sofia, Bulgaria.

Aluísio, Sandra M., Lucia Specia, Thiago A. S. Pardo, Erick G. Maziero, Helena M. Caseli, and Renata P. M. Fortes. 2008. A corpus analysis of simple account texts and the proposal of simplification strategies: First steps towards text simplification systems. In *Proceedings of the 26th Annual ACM International Conference on Design of Communication*, SIGDOC '08, pages 15–22, ACM, Lisbon, Portugal.

Alva-Manchego, Fernando. 2020. *Automatic Sentence Simplification with Multiple Rewriting Transformations*. Phd thesis, University of Sheffield, Sheffield, UK.

Alva-Manchego, Fernando, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305, Asian Federation of Natural Language Processing, Taipei, Taiwan.

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

Alva-Manchego, Fernando, 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, Association for Computational Linguistics, Hong Kong, China.

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

Amidei, Jacopo, Paul Piwek, and Alistair Willis. 2019. Agreement is overrated: A plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation*, pages 344–354, Association for Computational Linguistics, Tokyo, Japan.

Barrault, Loïc, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Philipp Koehn, Shervin Malmasi, Christof Monz, Mathias Müller, Santanu Pal, Matt Post, and Marcos Zampieri. 2019. Findings of the 2019 conference on machine translation (WMT19). In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 1–61, Association for Computational Linguistics, Florence, Italy.

Bojar, Ondřej, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Philipp Koehn, and Christof Monz. 2018. Findings of the 2018 conference on machine translation (WMT18). In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 272–303, Association for Computational Linguistics, Belgium, Brussels.

Bojar, Ondřej, Yvette Graham, and Amir Kamran. 2017. Results of the WMT17 metrics shared task. In *Proceedings of the Second Conference on Machine Translation*, pages 489–513, Association for Computational Linguistics, Copenhagen, Denmark.

Bojar, Ondřej, Yvette Graham, Amir Kamran, and Miloš Stanojević. 2016. Results of the WMT16 metrics shared task. In *Proceedings of the First Conference on Machine Translation: Volume 2, Shared Task Papers*, pages 199–231, Association for Computational Linguistics, Berlin, Germany.

Bott, Stefan and Horacio Saggion. 2011. Spanish text simplification: An exploratory study. *Procesamiento del Lenguaje Natural*, 47:87–95.

Cooper, Michael and Matthew Shardlow. 2020. CombiNMT: An exploration into neural text simplification models. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 5588–5594, European Language Resources Association, Marseille, France.

Devlin, Jacob, 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, Association for

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Computational Linguistics, Minneapolis, Minnesota.

Dong, Yue, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Association for Computational Linguistics, Florence, Italy.

Feblowitz, Dan and David Kauchak. 2013. Sentence simplification as tree transduction. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 1–10, Association for Computational Linguistics, Sofia, Bulgaria.

Fomicheva, Marina and Lucia Specia. 2019. Taking mt evaluation metrics to extremes: Beyond correlation with human judgments. *Computational Linguistics*, 45(3):515–558.

Freitag, Markus, David Grangier, and Isaac Caswell. 2020. BLEU might be guilty but references are not innocent. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 61–71, Association for Computational Linguistics, Online.

Ganitkevitch, Juri, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Association for Computational Linguistics, Atlanta, Georgia.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Association for Computational Linguistics, Sofia, Bulgaria.

Graham, Yvette, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2017. Can machine translation systems be evaluated by the crowd alone. *Natural Language Engineering*, 23(1):3–30.

Guo, Han, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 462–476, Association for Computational Linguistics, Santa Fe, New Mexico, USA.

Jalili Sabet, Masoud, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. SimAlign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643, Association for Computational Linguistics, Online.

Jiang, Chao, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF model for sentence alignment in text simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Association for Computational Linguistics, Online.

Kincaid, J.P., R.P. Fishburne, R.L. Rogers, and B.S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical Report 8-75, Chief of Naval Technical Training: Naval Air Station Memphis. 49 p.

Koehn, Philipp, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, ACL '07, pages 177–180, Association for Computational Linguistics, Stroudsburg, PA, USA.

Kriz, Reno, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. 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 3137–3147, Association for Computational Linguistics, Minneapolis, Minnesota.

Kumar, Dhruv, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928, Association for Computational Linguistics, Online.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Landis, J. Richard and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Ma, Qingsong, Ondřej Bojar, and Yvette Graham. 2018. Results of the WMT18 metrics shared task: Both characters and embeddings achieve good performance. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 671–688, Association for Computational Linguistics, Belgium, Brussels.

Ma, Qingsong, Johnny Wei, Ondřej Bojar, and Yvette Graham. 2019. Results of the WMT19 metrics shared task: Segment-level and strong MT systems pose big challenges. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 62–90, Association for Computational Linguistics, Florence, Italy.

Maddela, Mounica, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Association for Computational Linguistics, Online.

Martin, Louis, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4689–4698, European Language Resources Association, Marseille, France.

Mathur, Nitika, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Association for Computational Linguistics, Online.

Mathur, Nitika, Johnny Wei, Markus Freitag, Qingsong Ma, and Ondřej Bojar. 2020. Results of the WMT20 metrics shared task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 688–725, Association for Computational Linguistics, Online.

Narayan, Shashi and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Association for Computational Linguistics, Baltimore, Maryland.

Narayan, Shashi and Claire Gardent. 2016. Unsupervised sentence simplification using deep semantics. In *Proceedings of the 9th International Natural Language Generation conference*, pages 111–120, Association for Computational Linguistics, Edinburgh, UK.

Nisioi, Sergiu, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Association for Computational Linguistics, Vancouver, Canada.

Papineni, Kishore, 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 on Association for Computational Linguistics*, ACL '02, pages 311–318, ACL, Philadelphia, Pennsylvania.

Pavlick, Ellie and Chris Callison-Burch. 2016. Simple ppdb: A paraphrase database for simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Association for Computational Linguistics, Berlin, Germany.

Petersen, Sarah E. 2007. *Natural Language Processing Tools for Reading Level Assessment and Text Simplification for Bilingual Education*. Ph.D. thesis, University of Washington, Seattle, WA, USA. AAI3275902.

Post, Matt. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Association for Computational Linguistics, Brussels, Belgium.

Reiter, Ehud. 2018. A structured review of the validity of BLEU. *Computational Linguistics*, 44(3):393–401.

Rosenthal, James A. 1996. Qualitative descriptors of strength of association and effect size. *Journal of social service research*, 21(4):37–59.

Shrout, Patrick E. and Joseph L. Fleiss. 1979. Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2):420–428.

Sulem, Elior, Omri Abend, and Ari Rappoport. 2018a. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on*

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

Empirical Methods in Natural Language Processing, pages 738–744, Association for Computational Linguistics, Brussels, Belgium.

Sulem, Elior, 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, Association for Computational Linguistics, New Orleans, Louisiana.

Sulem, Elior, Omri Abend, and Ari Rappoport. 2018c. 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)*, pages 162–173, Association for Computational Linguistics, Melbourne, Australia.

Sun, Hong and Ming Zhou. 2012. Joint learning of a dual smt system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2*, ACL '12, pages 38–42, Association for Computational Linguistics, Stroudsburg, PA, USA.

Vallat, Raphael. 2018. Pingouin: statistics in python. *Journal of Open Source Software*, 3(31):1026.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*. Curran Associates, Inc., pages 5998–6008.

Vu, Tu, Baotian Hu, Tsendsuren Munkhdalai, and Hong Yu. 2018. Sentence simplification with memory-augmented neural networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 79–85, Association for Computational Linguistics, New Orleans, Louisiana.

Williams, Evan James. 1959. *Regression Analysis*, volume 14. Wiley, New York.

Woodsend, Kristian and Mirella Lapata. 2011. Learning to Simplify Sentences with Quasi-Synchronous Grammar and Integer Programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Association for Computational Linguistics, Edinburgh, Scotland, UK.

Wubben, Sander, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers - Volume 1*, ACL '12, pages 1015–1024, Association for Computational Linguistics, Stroudsburg, PA, USA.

Xu, Wei, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xu, Wei, 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.

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

Zhang, Xingxing and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 595–605, Association for Computational Linguistics, Copenhagen, Denmark.

Zhao, Sanqiang, Rui Meng, Daqing He, Andi Saptono, and Bambang Parmanto. 2018. Integrating transformer and paraphrase rules for sentence simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3164–3173, Association for Computational Linguistics, Brussels, Belgium.

Zhu, Zhemin, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1353–1361, Association for Computational Linguistics, Stroudsburg, PA, USA.

Štajner, Sanja, Ruslan Mitkov, and Horacio Saggion. 2014. One step closer to automatic evaluation of text simplification systems. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/coli_a_00418

1–10, Association for Computational Linguistics, Gothenburg, Sweden.