ASSET: A Dataset for Tuning and Evaluation of Sentence Simplification Models with Multiple Rewriting Transformations

Fernando Alva-Manchego[1*] and **Louis Martin**[2,3*] and **Antoine Bordes**[3]
Carolina Scarton[1] and **Benoît Sagot**[2] and **Lucia Specia**[1,4]
[1]University of Sheffield, [2]Inria, [3]Facebook AI Research, [4]Imperial College London
f.alva@sheffield.ac.uk, louismartin@fb.com, abordes@fb.com
c.scarton@sheffield.ac.uk, benoit.sagot@inria.fr
l.specia@imperial.ac.uk

Abstract

In order to simplify a sentence, human editors perform multiple rewriting transformations: they split it into several shorter sentences, paraphrase words (i.e. replacing complex words or phrases by simpler synonyms), reorder components, and/or delete information deemed unnecessary. Despite these varied range of possible text alterations, current models for automatic sentence simplification are evaluated using datasets that are focused on a single transformation, such as lexical paraphrasing or splitting. This makes it impossible to understand the ability of simplification models in more realistic settings. To alleviate this limitation, this paper introduces ASSET, a new dataset for assessing sentence simplification in English. ASSET is a crowdsourced multi-reference corpus where each simplification was produced by executing several rewriting transformations. Through quantitative and qualitative experiments, we show that simplifications in ASSET are better at capturing characteristics of simplicity when compared to other standard evaluation datasets for the task. Furthermore, we motivate the need for developing better methods for automatic evaluation using ASSET, since we show that current popular metrics may not be suitable when multiple simplification transformations are performed.

1 Introduction

Sentence Simplification (SS) consists in modifying the content and structure of a sentence to make it easier to understand, while retaining its main idea and most of its original meaning (Alva-Manchego et al., 2020). Simplified texts can benefit non-native speakers (Paetzold, 2016), people suffering from aphasia (Carroll et al., 1998), dyslexia (Rello et al., 2013) or autism (Evans et al., 2014). They also help language processing tasks, such as parsing (Chandrasekar et al., 1996), summarisation (Silveira and Branco, 2012), and machine translation (Hasler et al., 2017).

In order simplify a sentence, several rewriting transformations can be performed: replacing complex words/phrases with simpler synonyms (i.e. lexical paraphrasing), changing the syntactic structure of the sentence (e.g. splitting), or removing superfluous information that make the sentence more complicated (Petersen, 2007; Aluísio et al., 2008; Bott and Saggion, 2011). However, models for automatic SS are evaluated on datasets whose simplifications are not representative of this variety of transformations. For instance, TurkCorpus (Xu et al., 2016), a standard dataset for assessment in SS, contains simplifications produced mostly by lexical paraphrasing, while reference simplifications in HSplit (Sulem et al., 2018a) focus on splitting sentences. The Newsela corpus (Xu et al., 2015) contains simplifications produced by professionals applying multiple rewriting transformations, but sentence alignments are automatically computed and thus imperfect, and its data can only be accessed after signing a restrictive public-sharing licence and cannot be redistributed, hampering reproducibility.

These limitations in evaluation data prevent studying models' capabilities to perform a broad range of simplification transformations. Even though most SS models are trained on simplification instances displaying several text transformations (e.g. WikiLarge (Zhang and Lapata, 2017)), we currently do not measure their performance in more *abstractive* scenarios, i.e. cases with substantial modifications to the original sentences.

In this paper we introduce **ASSET** (**A**bstractive **S**entence **S**implification **E**valuation and **T**uning), a new dataset for tuning and evaluation of automatic SS models. ASSET consists of 23,590 human simplifications associated with the 2,359 original sentences from TurkCorpus (10 simplifications per

[*]Equal Contribution

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4668–4679
July 5 - 10, 2020.

original sentence). Simplifications in ASSET were collected via crowdsourcing (§ 3), and encompass a variety of rewriting transformations (§ 4), which make them simpler than those in TurkCorpus and HSplit (§ 5), thus providing an additional suitable benchmark for comparing and evaluating automatic SS models. In addition, we study the applicability of standard metrics for evaluating SS using simplifications in ASSET as references (§ 6). We analyse whether BLEU (Papineni et al., 2002) or SARI (Xu et al., 2016) scores correlate with human judgements of fluency, adequacy and simplicity, and find that neither of the metrics shows a strong correlation with simplicity ratings. This motivates the need for developing better metrics for assessing SS when multiple rewriting transformations are performed.

We make the following contributions:

- A high quality large dataset for tuning and evaluation of SS models containing simplifications produced by applying multiple rewriting transformations.[1]
- An analysis of the characteristics of the dataset that turn it into a new suitable benchmark for evaluation.
- A study questioning the suitability of popular metrics for evaluating automatic simplifications in a multiple-transformation scenario.

2 Related Work

2.1 Studies on Human Simplification

A few corpus studies have been carried out to analyse how humans simplify sentences, and to attempt to determine the rewriting transformations that are performed.

Petersen and Ostendorf (2007) analysed a corpus of 104 original and professionally simplified news articles in English. Sentences were manually aligned and each simplification instance was categorised as dropped (1-to-0 alignment), split (1-to-N), total (1-to-1) or merged (2-to-1). Some splits were further sub-categorised as edited (i.e. the sentence was split and some part was dropped) or different (i.e. same information but very different wording). This provides evidence that sentence splitting and deletion of information can be performed simultaneously.

Aluísio et al. (2008) studied six corpora of simple texts (different genres) and a corpus of complex news texts in Brazilian Portuguese, to produce a manual for Portuguese text simplification (Specia et al., 2008). It contains several rules to perform the task focused on syntactic alterations: to split adverbial/coordinated/subordinated sentences, to reorder clauses to a subject-verb-object structure, to transform passive to active voice, among others.

Bott and Saggion (2011) worked with a dataset of 200 news articles in Spanish with their corresponding manual simplifications. After automatically aligning the sentences, the authors determined the simplification transformations performed: change (e.g. difficult words, pronouns, voice of verb), delete (words, phrases or clauses), insert (word or phrases), split (relative clauses, coordination, etc.), proximisation (add locative phrases, change from third to second person), reorder, select, and join (sentences).

From all these studies, it can be argued that the scope of rewriting transformations involved in the simplification process goes beyond only replacing words with simpler synonyms. In fact, human perception of complexity is most affected by syntactic features related to sentence structure (Brunato et al., 2018). Therefore, since human editors make several changes to both the lexical content and syntactic structure of sentences when simplifying them, we should expect that models for automatic sentence simplification can also make such changes.

2.2 Evaluation Data for SS

Most datasets for SS (Zhu et al., 2010; Coster and Kauchak, 2011; Hwang et al., 2015) consist of automatic sentence alignments between related articles in English Wikipedia (EW) and Simple English Wikipedia (SEW). In SEW, contributors are asked to write texts using simpler language, such as by shortening sentences or by using words from Basic English (Ogden, 1930). However, Yasseri et al. (2012) found that the syntactic complexity of sentences in SEW is almost the same as in EW. In addition, Xu et al. (2015) determined that automatically-aligned simple sentences are sometimes just as complex as their original counterparts, with only a few words replaced or dropped and the rest of the sentences left unchanged.

More diverse simplifications are available in the Newsela corpus (Xu et al., 2015), a dataset of 1,130 news articles that were each manually simplified

[1] ASSET is released with a CC-BY-NC license at https://github.com/facebookresearch/asset.

to up to 5 levels of simplicity. The parallel articles can be automatically aligned at the sentence level to train and test simplification models (Alva-Manchego et al., 2017; Štajner et al., 2018). However, the Newsela corpus can only be accessed after signing a restrictive license that prevents publicly sharing train/test splits of the dataset, which impedes reproducibility.

Evaluating models on automatically-aligned sentences is problematic. Even more so if only one (potentially noisy) reference simplification for each original sentence is available. With this concern in mind, Xu et al. (2016) collected the TurkCorpus, a dataset with 2,359 original sentences from EW, each with 8 manual reference simplifications. The dataset is divided into two subsets: 2,000 sentences for validation and 359 for testing of sentence simplification models. TurkCorpus is suitable for automatic evaluation that involves metrics requiring multiple references, such as BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016). However, Xu et al. (2016) focused on simplifications through lexical paraphrasing, instructing annotators to rewrite sentences by reducing the number of difficult words or idioms, but without deleting content or splitting the sentences. This prevents evaluating a model's ability to perform a more diverse set of rewriting transformations when simplifying sentences. HSplit (Sulem et al., 2018a), on the other hand, provides simplifications involving only splitting for sentences in the test set of TurkCorpus. We build on TurkCorpus and HSplit by collecting a dataset that provides several manually-produced simplifications involving multiple types of rewriting transformations.

2.3 Crowdsourcing Manual Simplifications

A few projects have been carried out to collect manual simplifications through crowdsourcing. Pellow and Eskenazi (2014a) built a corpus of everyday documents (e.g. driving test preparation materials), and analysed the feasibly of crowdsourcing their sentence-level simplifications. Of all the quality control measures taken, the most successful was providing a training session to workers, since it allowed to block spammers and those without the skills to perform the task. Additionally, they proposed to use workers' self-reported confidence scores to flag submissions that could be discarded or reviewed. Later on, Pellow and Eskenazi (2014b) presented a preliminary study on producing simplifications through a collaborative process. Groups of four workers were assigned one sentence to simplify, and they had to discuss and agree on the process to perform it. Unfortunately, the data collected in these studies is no longer publicly available.

Simplifications in TurkCorpus were also collected through crowdsourcing. Regarding the methodology followed, Xu et al. (2016) only report removing bad workers after manual check of their first several submissions. More recently, Scarton et al. (2018) used volunteers to collect simplifications for SimPA, a dataset with sentences from the Public Administration domain. One particular characteristic of the methodology followed is that lexical and syntactic simplifications were performed independently.

3 Creating ASSET

We extended TurkCorpus (Xu et al., 2016) by using the same original sentences, but crowdsourced manual simplifications that encompass a richer set of rewriting transformations. Since TurkCorpus was adopted as the standard dataset for evaluating SS models, several system outputs on this data are already publicly available (Zhang and Lapata, 2017; Zhao et al., 2018; Martin et al., 2020). Therefore, we can now assess the capabilities of these and other systems in scenarios with varying simplification expectations: lexical paraphrasing with TurkCorpus, sentence splitting with HSplit, and multiple transformations with ASSET.

3.1 Data Collection Protocol

Manual simplifications were collected using Amazon Mechanical Turk (AMT). AMT allows us to publish HITs (Human Intelligence Tasks), which workers can choose to work on, submit an answer, and collect a reward if the work is approved. This was also the platform used for TurkCorpus.

Worker Requirements. Participants were workers who: (1) have a HIT approval rate $>= 95\%$; (2) have a number of HITs approved > 1000; (3) are residents of the United States of America, the United Kingdom or Canada; and (4) passed the corresponding Qualification Test designed for our task (more details below). The first two requirements are measured by the AMT platform and ensure that the workers have experience on different tasks and have had most of their work approved by previous requesters. The last two requirements are intended

Original	Their eyes are quite small, and their visual acuity is poor.
TurkCorpus	Their eyes are very little, and their sight is inferior.
HSplit	Their eyes are quite small. Their visual acuity is poor as well.
ASSET	They have small eyes and poor eyesight.
Original	His next work, Saturday, follows an especially eventful day in the life of a successful neurosurgeon.
TurkCorpus	His next work at Saturday will be a successful Neurosurgeon.
HSplit	His next work was Saturday. It follows an especially eventful day in the life of a successful Neurosurgeon.
ASSET	"Saturday" records a very eventful day in the life of a successful neurosurgeon.
Original	He settled in London, devoting himself chiefly to practical teaching.
TurkCorpus	He rooted in London, devoting himself mainly to practical teaching.
HSplit	He settled in London. He devoted himself chiefly to practical teaching.
ASSET	He lived in London. He was a teacher.

Table 1: Examples of simplifications collected for ASSET together with their corresponding version from TurkCorpus and HSplit for the same original sentences.

to ensure that the workers have a proficient level of English, and are capable of performing the simplification task.

Qualification Test. We provided a training session to workers in the form of a Qualification Test (QT). Following Pellow and Eskenazi (2014a), we showed them explanations and examples of multiple simplification transformations (see details below). Each HIT consisted of three sentences to simplify, and all submissions were manually checked to filter out spammers and workers who could not perform the task correctly. The sentences used in this stage were extracted from the QATS dataset (Štajner et al., 2016). We had 100 workers take the QT, out of which 42 passed the test (42%) and worked on the task.

Annotation Round. Workers who passed the QT had access to this round. Similar to Pellow and Eskenazi (2014a), each HIT now consisted of four original sentences that needed to be simplified. In addition to the simplification of each sentence, workers were asked to submit confidence scores on their simplifications using a 5-point likert scale (1:Very Low, 5:Very High). We collected 10 simplifications (similar to Pellow and Eskenazi (2014a)) for each of the 2,359 original sentences in TurkCorpus.

Simplification Instructions. For both the QT and the Annotation Round, workers received the same set of instructions about how to simplify a sentence. We provided examples of lexical paraphrasing (lexical simplification and reordering), sentence splitting, and compression (deleting unimportant information). We also included an example where all transformations were performed. However, we clarified that it was at their discretion to decide which types of rewriting to execute in any given original sentence.[2]

Table 1 presents a few examples of simplifications in ASSET, together with references from TurkCorpus and HSplit, randomly sampled for the same original sentences. It can be noticed that annotators in ASSET had more freedom to change the structure of the original sentences.

3.2 Dataset Statistics

ASSET contains 23,590 human simplifications associated with the 2,359 original sentences from TurkCorpus (2,000 from the validation set and 359 from the test set). Table 2 presents some general statistics from simplifications in ASSET. We show the same statistics for TurkCorpus and HSplit for comparison.[3]

In addition to having more references per original sentence, ASSET's simplifications offer more variability, for example containing many more instances of natural sentence splitting than TurkCorpus. In addition, reference simplifications are shorter on average in ASSET, given that we allowed annotators to delete information that they considered unnecessary. In the next section, we further compare these datasets with more detailed text features.

4 Rewriting Transformations in ASSET

We study the simplifications collected for ASSET through a series of text features to measure the

[2]Full instructions are available in the dataset's repository.

[3]HSplit is composed of two sets of simplifications: one where annotators were asked to split sentences as much as they could, and one where they were asked to split the original sentence only if it made the simplification easier to read and understand. However, we consider HSplit as a whole because differences between datasets far outweigh differences between these two sets.

	ASSET	TurkCorpus	HSplit
Original Sentences	2,359	2,359	359
Num. of References	10	8	4
Type of Simp. Instances			
1-to-1	17,245	18,499	408
1-to-N	6,345	373	1,028
Tokens per Reference	19.04	21.29	25.49

Table 2: General surface statistics for ASSET compared with TurkCorpus and HSplit. A simplification instance is an original-simplified sentence pair.

abstractiveness of the rewriting transformations performed by the annotators. From here on, the analysis and statistics reported refer to the test set only (i.e. 359 original sentences), so that we can fairly compare ASSET, TurkCorpus and HSplit.

4.1 Text Features

In order to quantify the rewriting transformations, we computed several low-level features for all simplification instances using the `tseval` package (Martin et al., 2018):

- **Number of sentence splits:** Corresponds to the difference between the number of sentences in the simplification and the number of sentences in the original sentence. In `tseval`, the number of sentences is calculated using NLTK (Loper and Bird, 2002).
- **Compression level:** Number of characters in the simplification divided by the number of characters in the original sentence.
- **Replace-only Levenshtein distance:** Computed as the normalised character-level Levenshtein distance (Levenshtein, 1966) for replace operations only, between the original sentence and the simplification. Replace-only Levenshtein distance is computed as follows (with o the original sentence and s the simplification):

$$\frac{replace\_ops(o,s)}{min(len(o),len(s))}$$

  We do not consider insertions and deletions in the Levenshtein distance computation so that this feature is independent from the compression level. It therefore serves as a proxy for measuring the lexical paraphrases of the simplification.
- **Proportion of words deleted, added and reordered:** Number of words deleted/reordered from the original sentence divided by the number of words in the original sentence; and the number of words that were added to the original sentence divided by the number of words in the simplification.
- **Exact match:** Boolean feature that equals to true when the original sentence and the simplification are exactly the same, to account for unchanged sentences.
- **Word deletion only:** Boolean feature that equals to true when the simplification is obtained only by deleting words from the original sentence. This feature captures extractive compression.
- **Lexical complexity score ratio:** We compute the score as the mean squared log-ranks of content words in a sentence (i.e. without stopwords). We use the 50k most frequent words of the FastText word embeddings vocabulary (Bojanowski et al., 2016). This vocabulary was originally sorted with frequencies of words in the Common Crawl. This score is a proxy to the lexical complexity of the sentence given that word ranks (in a frequency table) have been shown to be best indicators of word complexity (Paetzold and Specia, 2016). The ratio is then the value of this score on the simplification divided by that of the original sentence.
- **Dependency tree depth ratio:** We compute the ratio of the depth of the dependency parse tree of the simplification relative to that of the original sentence. When a simplification is composed by more than one sentence, we choose the maximum depth of all dependency trees. Parsing is performed using spaCy.[4] This feature serves as a proxy to measure improvements in structural simplicity.

Each feature was computed for all simplification instances in the dataset and then aggregated as a histogram (Figure 1) and as a percentage (Table 3).

4.2 Results and Analysis

Figure 1 shows the density of all features in ASSET, and compares them with those in TurkCorpus and

[4] `github.com/explosion/spaCy`

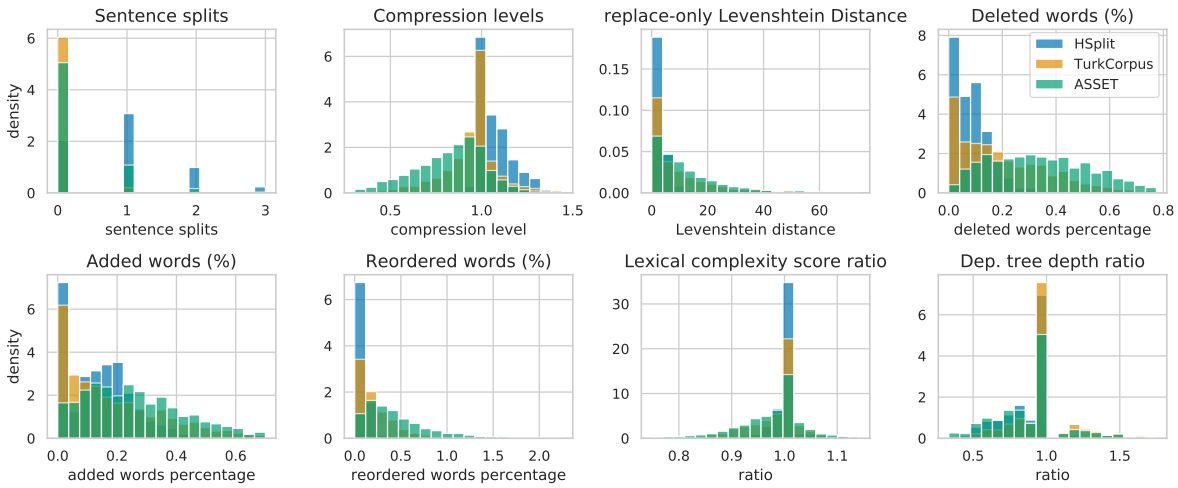

Figure 1: Density of text features in simplifications from HSplit, TurkCorpus, and ASSET.

	ASSET	TurkCorpus	HSplit
Sentence Splitting	20.2%	4.6%	68.2%
Compression (<75%)	31.2%	9.9%	0.1%
Word Reordering	28.3%	19.4%	10.1%
Exact Match	0.4%	16.3%	26.5%
Word Deletion Only	4.5%	3.9%	0.0%

Table 3: Percentage of simplifications featuring one of different rewriting transformations operated in ASSET, TurkCorpus and HSplit. A simplification is considered as compressed when its character length is less than 75% of that of the original sentence.

HSplit. Table 3 highlights some of these statistics. In particular, we report the percentage of sentences that: have at least one sentence split, have a compression level of 75% or lower, have at least one reordered word, are exact copies of the original sentences, and operated word deletion only (e.g. by removing only an adverb).

Sentence splits are practically non-existent in TurkCorpus (only 4.6% have one split or more), and are more present and distributed in HSplit. In ASSET, annotators tended to not split sentences, and those who did mostly divided the original sentence into just two sentences (1 split).

Compression is a differentiating feature of ASSET. Both TurkCorpus and HSplit have high density of a compression ratio of 1.0, which means that no compression was performed. In fact, HSplit has several instances with compression levels greater than 1.0, which could be explained by splitting requiring adding words to preserve fluency. In contrast, ASSET offers more variability, perhaps signalling that annotators consider deleting information as an important simplification operation.

By analysing replace-only Levenshtein distance, we can see that simplifications in ASSET paraphrase the input more. For TurkCorpus and HSplit, most simplifications are similar to their original counterparts (higher densities closer to 0). On the other hand, ASSET's simplifications are distributed in all levels, indicating more diversity in the rewordings performed. This observation is complemented by the distributions of deleted, added and reordered words. Both TurkCorpus and HSplit have high densities of ratios close to 0.0 in all these features, while ASSET's are more distributed. Moreover, these ratios are rarely equal to 0 (low density), meaning that for most simplifications, at least some effort was put into rewriting the original sentence. This is comfirmed by the low percentage of exact matches in ASSET (0.4%) with respect to TurkCorpus (16.3%) and HSplit (26.5%). Once again, it suggests that more rewriting transformations are being performed in ASSET.

In terms of lexical complexity, HSplit has a high density of ratios close to 1.0 due to its simplifications being structural and not lexical. TurkCorpus offers more variability, as expected, but still their simplifications contain a high number of words that are equally complex, perhaps due to most simplifications just changing a few words. On the other hand, ASSET's simplifications are more distributed across different levels of reductions in lexical complexity.

Finally, all datasets show high densities of a 1.0 ratio in dependency tree depth. This could mean that significant structural changes were not made, which is indicated by most instances corresponding

to operations other than splitting. However, ASSET still contains more simplifications that reduce syntactic complexity than TurkCorpus and HSplit.

5 Rating Simplifications in ASSET

Here we measure the quality of the collected simplifications using human judges. In particular, we study if the *abstractive* simplifications in ASSET (test set) are preferred over lexical-paraphrase-only or splitting-only simplifications in TurkCorpus (test set) and HSplit, respectively.

5.1 Collecting Human Preferences

Preference judgments were crowdsourced with a protocol similar to that of the simplifications (§ 3.1).

Selecting Human Judges. Workers needed to comply with the same basic requirements as described in § 3.1. For this task, the Qualification Test (QT) consisted in rating the quality of simplifications based on three criteria: fluency (or grammaticality), adequacy (or meaning preservation), and simplicity. Each HIT consisted of six original-simplified sentence pairs, and workers were asked to use a continuous scale (0-100) to submit their level of agreement (0: Strongly disagree, 100: Strongly agree) with the following statements:

1. The Simplified sentence adequately expresses the meaning of the Original, perhaps omitting the least important information.
2. The Simplified sentence is fluent, there are no grammatical errors.
3. The Simplified sentence is easier to understand than the Original sentence.

Using continuous scales when crowdsourcing human evaluations is common practice in Machine Translation (Bojar et al., 2018; Barrault et al., 2019), since it results in higher levels of inter-annotator consistency (Graham et al., 2013). The six sentence pairs for the Rating QT consisted of:

- Three submissions to the Annotation QT, manually selected so that one contains splitting, one has a medium level of compression, and one contains grammatical and spelling mistakes. These allowed to check that the particular characteristics of each sentence pair affect the corresponding evaluation criteria.
- One sentence pair extracted from WikiLarge (Zhang and Lapata, 2017) that contains several sentence splits. This instance appeared twice in the HIT and allowed checking for intra-annotator consistency.
- One sentence pair from WikiLarge where the Original and the Simplification had no relation to each other. This served to check the attention level of the worker.

All submitted ratings were manually reviewed to validate the quality control established and to select the qualified workers for the task.

Preference Task. For each of the 359 original sentences in the test set, we randomly sampled one reference simplification from ASSET and one from TurkCorpus, and then asked qualified workers to choose which simplification answers best each of the following questions:

- **Fluency**: Which sentence is more fluent?
- **Meaning**: Which sentence expresses the original meaning the best?
- **Simplicity**: Which sentence is easier to read and understand?

Workers were also allowed to judge simplifications as "similar" when they could not determine which one was better. The same process was followed to compare simplifications in ASSET against those in HSplit. Each HIT consisted of 10 sentence pairs.

5.2 Results and Analysis

Table 4 (top section) presents, for each evaluation dimension, the percentage of times a simplification from ASSET or TurkCorpus was preferred over the other, and the percentage of times they were judged as "similar". In general, judges preferred ASSET's simplifications in terms of fluency and simplicity. However, they found TurkCorpus' simplifications more meaning preserving. This is expected since they were produced mainly by replacing words/phrases with virtually no deletion of content.

A similar behaviour was observed when comparing ASSET to HSplit (bottom section of Table 4). In this case, however, the differences in preferences are greater than with TurkCorpus. This could indicate that changes in syntactic structure are not enough for a sentence to be consider simpler.

	Fluency	Meaning	Simplicity
ASSET	**38.4%***	23.7%	**41.2%***
TurkCorpus	22.8%	**37.9%***	20.1%
Similar	38.7%	38.4%	38.7%
ASSET	**53.5%***	17.0%	**59.0%***
HSplit	19.5%	**51.5%***	14.8%
Similar	27.0%	31.5%	26.2%

Table 4: Percentages of human judges who preferred simplifications in ASSET or TurkCorpus, and ASSET or HSplit, out of 359 comparisons. * indicates a statistically significant difference between the two datasets (binomial test with p-value < 0.001).

6 Evaluating Evaluation Metrics

In this section we study the behaviour of evaluation metrics for SS when using ASSET's simplifications (test set) as references. In particular, we measure the correlation of standard metrics with human judgements of fluency, adequacy and simplicity, on simplifications produced by automatic systems.

6.1 Experimental Setup

Evaluation Metrics. We analysed the behaviour of two standard metrics in automatic evaluation of SS outputs: BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016). BLEU is a precision-oriented metric that relies on the number of n-grams in the output that match n-grams in the references, independently of position. SARI measures improvement in the simplicity of a sentence based on the n-grams added, deleted and kept by the simplification system. It does so by comparing the output of the simplification model to multiple references and the original sentence, using both precision and recall. BLEU has shown positive correlation with human judgements of grammaticality and meaning preservation (Štajner et al., 2014; Wubben et al., 2012; Xu et al., 2016), while SARI has high correlation with judgements of simplicity gain (Xu et al., 2016). In our experiments, we used the implementations of these metrics available in the EASSE package for automatic sentence simplification evaluation (Alva-Manchego et al., 2019).[5] We computed all the scores at sentence-level as in the experiment by Xu et al. (2016), where they compared sentence-level correlations of FKGL, BLEU and SARI with human ratings. We used a smoothed sentence-level version of BLEU so that comparison is possible, even though BLEU was designed as a corpus-level metric.

System Outputs. We used publicly-available simplifications produced by automatic SS systems: PBSMT-R (Wubben et al., 2012), which is a phrase-based MT model; Hybrid (Narayan and Gardent, 2014), which uses phrase-based MT coupled with semantic analysis; SBSMT-SARI (Xu et al., 2016), which relies on syntax-based MT; NTS-SARI (Nisioi et al., 2017), a neural sequence-to-sequence model with a standard encoder-decoder architecture; and ACCESS (Martin et al., 2020), an encoder-decoder architecture conditioned on explicit attributes of sentence simplification.

Collection of Human Ratings. We randomly chose 100 original sentences from ASSET and, for each of them, we sampled one system simplification. The automatic simplifications were selected so that the distribution of simplification transformations (e.g. sentence splitting, compression, paraphrases) would match that from human simplifications in ASSET. That was done so that we could obtain a sample that has variability in the types of rewritings performed. For each sentence pair (original and automatic simplification), we crowdsourced 15 human ratings on fluency (i.e. grammaticality), adequacy (i.e. meaning preservation) and simplicity, using the same worker selection criteria and HIT design of the Qualification Test as in § 5.1.

6.2 Inter-Annotator Agreement

We followed the process suggested in (Graham et al., 2013). First, we normalised the scores of each rater by their individual mean and standard deviation, which helps eliminate individual judge preferences. Then, the normalised continuous scores were converted to five interval categories using equally spaced bins. After that, we followed Pavlick and Tetreault (2016) and computed quadratic weighted Cohen's κ (Cohen, 1968) simulating two raters: for each sentence, we chose one worker's rating as the category for annotator A, and selected the rounded average scores for the remaining workers as the category for annotator B. We then computed κ for this pair over the whole dataset. We repeated the process 1,000 times to compute the mean and variance of κ. The resulting values are: 0.687 ± 0.028 for Fluency, 0.686 ± 0.030 for Meaning and 0.628 ± 0.032 for Simplicity. All values point to a moderate level

[5] https://github.com/feralvam/easse

Metric	References	Fluency	Meaning	Simplicity
BLEU	ASSET	0.42*	0.61*	0.31*
	TurkCorpus	0.35*	0.59*	0.18
SARI	ASSET	0.16	0.13	0.28*
	TurkCorpus	0.14	0.10	0.17

Table 5: Pearson correlation of human ratings with **automatic metrics** on system simplifications. * indicates a significance level of p-value < 0.05.

of agreement, which is in line with the subjective nature of the simplification task.

6.3 Correlation with Evaluation Metrics

We computed the Pearson correlation between the normalised ratings and the evaluation metrics of our interest (BLEU and SARI) using ASSET or TurkCorpus as the set of references. We refrained from experimenting with HSplit since neither BLEU nor SARI correlate with human judgements when calculated using that dataset as references (Sulem et al., 2018a). Results are reported in Table 5.

BLEU shows a strong positive correlation with Meaning Preservation using either simplifications from ASSET or TurkCorpus as references. There is also some positive correlation with Fluency judgements, but that is not always the case for Simplicity: no correlation when using TurkCorpus and moderate when using ASSET. This is in line with previous studies that have shown that BLEU is not a good estimate for simplicity (Wubben et al., 2012; Xu et al., 2016; Sulem et al., 2018b).

In the case of SARI, correlations are positive but low with all criteria and significant only for simplicity with ASSET's references. Xu et al. (2016) showed that SARI correlated with human judgements of simplicity gain, when instructing judges to *"grade the quality of the variations by identifying the words/phrases that are altered, and counting how many of them are good simplifications"*.[6] The judgements they requested differ from the ones we collected, since theirs were tailored to rate simplifications produced by lexical paraphrasing only. These results show that SARI might not be suitable for the evaluation of automatic simplifications with multiple rewrite operations.

In Table 6, we further analyse the human ratings collected, and compute their correlations with similar text features as in § 4. The results shown reinforce our previous observations that judgements on Meaning correlate with making few changes to the sentence: strong negative correlation with Levenshtein distance, and strong negative correlation with proportion of words added, deleted, and reordered. No conclusions could be drawn with respect to Simplicity.

[6] https://github.com/cocoxu/simplification/tree/master/HIT_MTurk_crowdsourcing

Feature	Fluency	Meaning	Simplicity
Length	0.12	0.31*	0.03
Sentence Splits	-0.13	-0.06	-0.08
Compression Level	0.26*	0.46*	0.04
Levenshtein Distance	-0.40*	-0.67*	-0.18
Replace-only Lev. Dist.	-0.04	-0.17	-0.06
Prop. Deleted Words	-0.43*	-0.67*	-0.19
Prop. Added Words	-0.19	-0.38*	-0.12
Prop. Reordered Words	-0.37*	-0.57*	-0.18
Dep. Tree Depth Ratio	0.20	0.24	0.06
Word Rank Ratio	0.04	0.08	-0.05

Table 6: Pearson correlation of human ratings with **text features** on system simplifications. * indicates a significance level of p-value < 0.01.

7 Conclusion

We have introduced ASSET, a new dataset for tuning and evaluation of SS models. Simplifications in ASSET were crowdsourced, and annotators were instructed to apply multiple rewriting transformations. This improves current publicly-available evaluation datasets, which are focused on only one type of transformation. Through several experiments, we have shown that ASSET contains simplifications that are more *abstractive*, and that are consider simpler than those in other evaluation corpora. Furthermore, we have motivated the need to develop new metrics for automatic evaluation of SS models, especially when evaluating simplifications with multiple rewriting operations. Finally, we hope that ASSET's multi-transformation features will motivate the development of SS models that benefit a variety of target audiences according to their specific needs such as people with low literacy or cognitive disabilities.

Acknowledgements

This work was partly supported by Benoît Sagot's chair in the PRAIRIE institute, funded by the French national agency ANR as part of the "Investissements d'avenir" programme under the reference ANR-19-P3IA-0001.

References

Sandra M. Aluísio, 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, Lisbon, Portugal. ACM.

Fernando Alva-Manchego, 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, Taipei, Taiwan. Asian Federation of Natural Language Processing.

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

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

Loïc Barrault, 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, Florence, Italy. Association for Computational Linguistics.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Ondřej Bojar, 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, Belgium, Brussels. Association for Computational Linguistics.

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

Dominique Brunato, Lorenzo De Mattei, Felice Dell'Orletta, Benedetta Iavarone, and Giulia Venturi. 2018. Is this sentence difficult? do you agree? In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2690–2699, Brussels, Belgium. Association for Computational Linguistics.

John Carroll, Guido Minnen, Yvonne Canning, Siobhan Devlin, and John Tait. 1998. Practical simplification of english newspaper text to assist aphasic readers. In *Proceedings of AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics*, volume 2 of *COLING '96*, pages 1041–1044, Copenhagen, Denmark. Association for Computational Linguistics.

Jacob Cohen. 1968. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological Bulletin*, 70(4):213–220.

William Coster and David Kauchak. 2011. Simple english wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Short Papers - Volume 2*, HLT '11, pages 665–669, Stroudsburg, PA, USA. Association for Computational Linguistics.

Richard Evans, Constantin Orasan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations*, PIT 2014, pages 131–140, Gothenburg, Sweden. Association for Computational Linguistics.

Yvette Graham, 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, Sofia, Bulgaria. Association for Computational Linguistics.

Eva Hasler, Adri de Gispert, Felix Stahlberg, Aurelien Waite, and Bill Byrne. 2017. Source sentence simplification for statistical machine translation. *Computer Speech & Language*, 45(C):221–235.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 211–217, Denver, Colorado. Association for Computational Linguistics.

VI Levenshtein. 1966. Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10:707.

Edward Loper and Steven Bird. 2002. NLTK: the natural language toolkit. *CoRR*, cs.CL/0205028.

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

Louis Martin, Benoît Sagot, Éric de la Clergerie, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*.

Shashi Narayan 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, Baltimore, Maryland. Association for Computational Linguistics.

Sergiu Nisioi, 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, Vancouver, Canada. Association for Computational Linguistics.

Charles Kay Ogden. 1930. *Basic English: A General Introduction with Rules and Grammar*. Kegan Paul, Trench, Trubner & Co.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Gustavo Henrique Paetzold. 2016. *Lexical Simplification for Non-Native English Speakers*. Ph.D. thesis, University of Sheffield, Sheffield, UK.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Philadelphia, Pennsylvania. ACL.

Ellie Pavlick and Joel Tetreault. 2016. An empirical analysis of formality in online communication. *Transactions of the Association for Computational Linguistics*, 4:61–74.

David Pellow and Maxine Eskenazi. 2014a. An open corpus of everyday documents for simplification tasks. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 84–93, Gothenburg, Sweden. Association for Computational Linguistics.

David Pellow and Maxine Eskenazi. 2014b. Tracking human process using crowd collaboration to enrich data. In *Human Computation and Crowdsourcing: Works in Progress and Demonstration Abstracts. An Adjunct to the Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing*, pages 52–53.

Sarah E. Petersen. 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.

Sarah E. Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Proceedings of the Speech and Language Technology for Education Workshop*, SLaTE 2007, pages 69–72.

Luz Rello, Clara Bayarri, Azuki Gòrriz, Ricardo Baeza-Yates, Saurabh Gupta, Gaurang Kanvinde, Horacio Saggion, Stefan Bott, Roberto Carlini, and Vasile Topac. 2013. "dyswebxia 2.0!: More accessible text for people with dyslexia". In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, W4A '13, pages 25:1–25:2, Rio de Janeiro, Brazil. ACM.

Carolina Scarton, Gustavo H. Paetzold, and Lucia Specia. 2018. Simpa: A sentence-level simplification corpus for the public administration domain. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sara Botelho Silveira and António Branco. 2012. Enhancing multi-document summaries with sentence simplificatio. In *Proceedings of the 14th International Conference on Artificial Intelligence,*, ICAI 2012, pages 742–748, Las Vegas, USA.

Lúcia Specia, Sandra Maria Aluísio, and Thiago A. Salgueiro Pardo. 2008. Manual de simplificação sintática para o português. Technical Report NILC-TR-08-06, NILC–ICMC–USP, São Carlos, SP, Brasil. Available in `http://www.nilc.icmc.usp.br/nilc/download/NILC_TR_08_06.pdf`.

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

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the*

North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 685–696, New Orleans, Louisiana. Association for Computational Linguistics.

Sanja Štajner, Marc Franco-Salvador, Paolo Rosso, and Simone Paolo Ponzetto. 2018. Cats: A tool for customized alignment of text simplification corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Sanja Štajner, Maja Popović, Horacio Saggion, Lucia Specia, and Mark Fishel. 2016. Shared task on quality assessment for text simplification. In *Proceeding of the Workshop on Quality Assessment for Text Simplification - LREC 2016*, QATS 2016, pages 22–31, Portorož, Slovenia. European Language Resources Association (ELRA).

Sander Wubben, 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, Stroudsburg, PA, USA. Association for Computational Linguistics.

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

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

Taha Yasseri, András Kornai, and János Kertész. 2012. A practical approach to language complexity: A wikipedia case study. *PLOS ONE*, 7(11):1–8.

Xingxing Zhang 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, Copenhagen, Denmark. Association for Computational Linguistics.

Sanqiang Zhao, 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, Brussels, Belgium. Association for Computational Linguistics.

Zhemin Zhu, 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, Stroudsburg, PA, USA. Association for Computational Linguistics.