1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI2020), pages 41–48
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020

Benchmarking Data-driven Automatic Text Simplification for German

Andreas Säuberli, Sarah Ebling, Martin Volk
Department of Computational Linguistics, University of Zurich
Andreasstrasse 15, 8050 Zurich, Switzerland
andreas.saeuberli@uzh.ch, {ebling,volk}@cl.uzh.ch

Abstract

Automatic text simplification is an active research area, and there are first systems for English, Spanish, Portuguese, and Italian. For German, no data-driven approach exists to this date, due to a lack of training data. In this paper, we present a parallel corpus of news items in German with corresponding simplifications on two complexity levels. The simplifications have been produced according to a well-documented set of guidelines. We then report on experiments in automatically simplifying the German news items using state-of-the-art neural machine translation techniques. We demonstrate that despite our small parallel corpus, our neural models were able to learn essential features of simplified language, such as lexical substitutions, deletion of less relevant words and phrases, and sentence shortening.

Keywords: Simplified German, automatic text simplification, neural machine translation

1 Introduction

Simplified language is a variety of standard language characterized by reduced lexical and syntactic complexity, the addition of explanations for difficult concepts, and clearly structured layout.[1] Among the target groups of simplified language are persons with cognitive impairment and learning disabilities, prelingually deaf persons, functionally illiterate persons, and foreign language learners (Bredel and Maaß, 2016).

Automatic text simplification, the process of automatically producing a simplified version of a standard-language text, was initiated in the late 1990s (Carroll et al., 1998; Chandrasekar et al., 1996) and since then has been approached by means of rule-based and statistical methods. As part of a rule-based approach, the operations carried out typically include replacing complex lexical and syntactic units by simpler ones. A statistical approach generally conceptualizes the simplification task as one of converting a standard-language into a simplified-language text using machine translation techniques.

Research on automatic text simplification has been documented for English (Zhu et al., 2010), Spanish (Saggion et al., 2015), Portuguese (Aluisio and Gasperin, 2010), French (Brouwers et al., 2014), and Italian (Barlacchi and Tonelli, 2013). To the authors' knowledge, the work of Suter (2015) and Suter et al. (2016), who presented a prototype of a rule-based text simplification system, is the only proposal for German.

The paper at hand presents the first experiments in data-driven simplification for German, relying on neural machine translation. The data consists of news items manually simplified according to a well-known set of guidelines. Hence, the contribution of the paper is twofold:

1. Introducing a parallel corpus as data for automatic text simplification for German
2. Establishing a benchmark for automatic text simplification for German

[1]The term *plain language* is avoided, as it refers to a specific level of simplification. *Simplified language* subsumes all efforts of reducing the complexity of a text.

Section 2 presents the research background with respect to parallel corpora (Section 2.1) and monolingual sentence alignment tools (Section 2.2) for automatic text simplification. Section 3 introduces previous approaches to data-driven text simplification. Section 4 presents our work on automatic text simplification for German, introducing the data (Section 4.1), the models (Section 4.2), the results (Section 4.3), and a discussion (Section 4.4).

2 Parallel Corpora and Alignment Tools for Automatic Text Simplification

2.1 Parallel Corpora

Automatic text simplification via machine translation requires pairs of standard-language/simplified-language texts aligned at the sentence level, i.e., parallel corpora. A number of parallel corpora have been created to this end. Gasperin et al. (2010) compiled the PorSimples Corpus consisting of Brazilian Portuguese texts (2,116 sentences), each with two different levels of simplifications ("natural" and "strong"), resulting in around 4,500 aligned sentences. Bott and Saggion (2012) produced the Simplext Corpus consisting of 200 Spanish/simplified Spanish document pairs, amounting to a total of 1,149 (Spanish) and 1,808 (simplified Spanish) sentences (approximately 1,000 aligned sentences).

A large parallel corpus for automatic text simplification is the Parallel Wikipedia Simplification Corpus (PWKP) compiled from parallel articles of the English Wikipedia and the Simple English Wikipedia (Zhu et al., 2010), consisting of around 108,000 sentence pairs. Application of the corpus has been criticized for various reasons (Štajner et al., 2018); the most important among these is the fact that Simple English Wikipedia articles are often not translations of articles from the English Wikipedia. Hwang et al. (2015) provided an updated version of the corpus that includes a total of 280,000 full and partial matches between the two Wikipedia versions.

Another frequently used data collection, available for English and Spanish, is the Newsela Corpus (Xu et al., 2015) consisting of 1,130 news articles, each simplified into four school grade levels by professional editors.

Klaper et al. (2013) created the first parallel corpus for German/simplified German, consisting of 256 texts each (approximately 70,000 tokens) downloaded from the Web. More recently, Battisti et al. (2020) extended the corpus to 6,200 documents (nearly 211,000 sentences).

The above-mentioned PorSimples and Newsela corpora present standard-language texts simplified into multiple levels, thus accounting for a recent consensus in the area of simplified-language research, according to which a single level of simplified language is not sufficient; instead, multiple levels are required to account for the heterogeneous target usership. For simplified German, *capito*,[2] the largest provider of simplification services (translations and translators' training) in Austria, Germany, and Switzerland, distinguishes between three levels along the Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2009): A1, A2, and B1.[3] Each level is linguistically operationalized, i.e., specified with respect to linguistic constructions permitted or not permitted at the respective level.

2.2 Sentence Alignment Tools for Simplified Texts

A freely available tool exists for generating sentence alignments of standard-language/simplified-language document pairs: *Customized Alignment for Text Simplification (CATS)* (Štajner et al., 2018). *CATS* requires a number of parameters to be specified:

- **Similarity strategy:** *CATS* offers a lexical (character-n-gram-based, CNG) and two semantic similarity strategies. The two semantic similarity strategies, WAVG (Word Average) and CWASA (Continuous Word Alignment-based Similarity Analysis), both require pretrained word embeddings. WAVG averages the word vectors of a paragraph or sentence to obtain the final vector for the respective text unit. CWASA is based on the alignment of continuous words using directed edges.
- **Alignment strategy:** *CATS* allows for adhering to a monotonicity restriction, i.e., requiring the order of information to be identical on the standard-language and simplified-language side, or abandoning it.

3 Data-Driven Automatic Text Simplification

Specia (2010) introduced statistical machine translation to the automatic text simplification task, using data from a small parallel corpus (roughly 4,500 parallel sentences) for Portuguese. Coster and Kauchak (2011) used the original PWKP Corpus (cf. Section 2.1) to train a machine translation system. Xu et al. (2016) performed syntax-based statistical machine translation on the English/simplified English part of the Newsela Corpus.

Nisioi et al. (2017) introduced neural sequence-to-sequence models to automatic text simplification, performing experiments on both the Wikipedia dataset of (Hwang et al., 2015) and the Newsela Corpus for English, with automatic alignments derived from *CATS* (cf. Section 2.2). The authors used a Long Short-term Memory (LSTM) architecture (Hochreiter and Schmidhuber, 1997) as instance of Recurrent Neural Networks (RNNs).

Surya et al. (2019) proposed an unsupervised or partially supervised approach to text simplification. Their model is based on a neural encoder-decoder but differs from previous approaches by adding reconstruction, adversarial, and diversification loss, which allows for exploiting non-parallel data as well. However, the authors' results prove that some parallel data is still essential.

Finally, Palmero Aprosio et al. (2019) experimented with data augmentation methods for low-resource text simplification for Italian. Their unaugmented dataset is larger than the one presented in this paper but includes more low-quality simplifications due to automatic extraction of simplified sentences from the Web. Our work differs in that we benchmark and compare a wider variety of low-resource methods.

The most commonly applied automatic evaluation metrics for text simplification are BLEU (Papineni et al., 2002) and SARI (Xu et al., 2016). BLEU, the *de-facto* standard metric for machine translation, computes token n-gram overlap between a hypothesis and one or multiple references. A shortcoming of BLEU with respect to automatic text simplification is that it rewards hypotheses that do not differ from the input. By contrast, SARI was designed to punish such output. It does so by explicitly considering the input and rewarding tokens in the hypothesis that do not occur in the input but in one of the references (addition) and tokens in the input that are retained (copying) or removed (deletion) in both the hypothesis and one of the references.

SARI is generally used with multiple reference sentences, which are hard to obtain. Due to this limitation, human evaluation is often needed. This mostly consists of three types of ratings: how well the content or meaning of the standard-language text is preserved, how fluent or natural the simplified output is, and how much simpler the output is compared to the standard-language original. Each simplified unit (in most cases, a sentence) is typically rated on a 5-point scale with respect to each of the three dimensions.

4 Automatic Text Simplification for German

4.1 Training Data

All data used in our experiments was taken from the Austria Press Agency (*Austria Presse Agentur*, APA) corpus built by our group. At this press agency, four to six news items covering the topics of politics, economy, culture, and sports are manually simplified into two language levels, B1 and A2, each day following the *capito* guidelines introduced in Section 2.1. The subset of data used for the experiments reported in this paper contains standard-language news items along with their simplifications on level B1 between August 2018 and December 2019. The dataset will be described in more detail in a separate publication.

[2] https://www.capito.eu/ (last accessed: February 3, 2020)

[3] Note that while the CEFR was designed to measure foreign language skills, with simplified language, it is partly applied in the context first-language acquisition (Bredel and Maaß, 2016).

Original	*Jedes Kalb erhält spätestens sieben Tage nach der Geburt eine eindeutig identifizierbare Lebensnummer, die in Form von Ohrmarken beidseitig eingezogen wird.* ('At the latest seven days after birth, each calf is given a unique identification number, which is recorded on ear tags on both sides.')
B1	*In Österreich bekommt jedes Kalb spätestens 7 Tage nach seiner Geburt eine Nummer, mit der man es erkennen kann.* ('In Austria, at the latest 7 days after birth, each calf receives a number, with which it can be identified.')
Original	*US-Präsident Donald Trump hat in seiner mit Spannung erwarteten Rede zur Lage der Nation seine politischen Prioritäten betont, ohne große wirtschaftliche Initiativen vorzustellen.* ('In his eagerly awaited State of the Union address, U.S. President Donald Trump stressed his political priorities without presenting any major economic initiatives.')
B1	*US-Präsident Donald Trump hat am Dienstag seine Rede zur Lage der Nation gehalten.* ('U.S. President Donald Trump gave his State of the Union address on Tuesday.')
Original	*Sie stehe noch immer jeden Morgen um 6.00 Uhr auf und gehe erst gegen 21.00 Uhr ins Bett, berichtete das Guinness-Buch der Rekorde.* ('She still gets up at 6:00 a.m. every morning and does not go to bed until around 9:00 p.m., the Guinness Book of Records reported.')
B1	*Sie steht auch heute noch jeden Tag um 6 Uhr in der Früh auf und geht um 21 Uhr schlafen.* ('Even today, she still gets up at 6 every morning and goes to bed at 9.')

Table 1: Examples from the Austria Press Agency (APA) corpus

We aligned the sentences from the original German news articles with the simplified articles using *CATS* (cf. Section 2.2). We chose the WAVG similarity strategy in conjunction with fastText embeddings (Bojanowski et al., 2017). fastText offers pretrained word vectors in 157 languages, derived from Wikipedia and Common Crawl (Grave et al., 2018).[4] As our alignment strategy, we dismissed the monotonicity restriction due to our observation that the order of information in a simplified-language text is not always preserved compared to that of the corresponding standard-language text.

CATS is built on the heuristic that every simplified-language sentence is aligned with one or several standard-language sentences. For 1-to-*n* and *n*-to-1 alignments, each of the *n* sentences forms a separate sentence pair with its counterpart, i.e., the single counterpart is duplicated. This leads to oversampling of some sentences and—as we will discuss in Section 4.4—poses a significant challenge for learning algorithms, but it is inevitable because we cannot assume that the order of information is preserved after simplification.[5] Sentence pairs with a similarity score of less than 90% were discarded (this threshold was established based on empirical evaluation of the tool on a different dataset), which resulted in a total of 3,616 sentence pairs. Table 1 shows examples, which are also representative of the wide range of simplifications present in the texts. Table 2 shows the number of German and simplified German sentences that we used for training and evaluation. The sets are all disjoint, i.e., there are no cross-alignments between any of them. Since the dataset is already very small

German	Simplified German	Alignment	Usage
3316	3316	1:1, 1:*n*, *n*:1	training
300	300	1:1	validation
	3316	–	data augmentation
50		–	evaluation

Table 2: Number of sentences from the Austria Press Agency (APA) corpus in our experiments

and the automatic alignments are not perfect, we decided not to use a parallel test set but to select models based on their best performance on the validation set and evaluate manually without a target reference. We chose the number of sentences for data augmentation to match the number of parallel sentences during training, in accordance with Sennrich et al. (2016a).

We applied the following preprocessing steps:

- In the simplified German text, we replaced all hyphenated compounds (e.g., *Premier-Ministerin* 'female prime minister') with their unhyphenated equivalents (*Premierministerin*), but only if they never occur in hyphenated form in the original German corpus.
- We converted all tokens to lowercase. This reduces the subword vocabulary and ideally makes morpheme/subword correspondences more explicit across different parts of speech, since nouns are generally capitalized in German orthography.
- We applied byte-pair encoding (BPE) (Sennrich et al., 2016b), trained jointly on the source and target text. BPE splits tokens into subwords based on the frequencies of their character sequences. This decreases the total vocabulary size and increases overlap between source and target.

[4] `https://fasttext.cc/docs/en/crawl-vectors.html` (last accessed: November 25, 2019)

[5] Another possibility to deal with 1-to-*n* and *n*-to-1 alignments would be to merge them into single alignments by concatenation. However, in our case, this would have resulted in many segments becoming too long to be processed by the sequence-to-sequence model.

4.2 Neural Models in Our Experiments

All models in our experiments are based on the Transformer encoder-decoder architecture (Vaswani et al., 2017). We used *Sockeye* version 1.18.106 (Hieber et al., 2017) for training and translation into simplified German. Unless otherwise stated, the hyperparameters are defaults defined by *Sockeye*. The following is an overview of the models:

BASE baseline model; embedding size of 256

BPE5K same as BASE but with less BPE merge operations (10,000 → 5,000) (Sennrich and Zhang, 2019)

BATCH1K same as BASE but with a smaller token-based batch size (4096 → 1024) (Sennrich and Zhang, 2019)

LINGFEAT same as BASE but extending embedding vectors with additional linguistic features (lemmas, part-of-speech tags, morphological attributes, dependency tags, and BIEO tags marking where subwords begin or end) (Sennrich and Haddow, 2016)

NULL2TRG same as BASE but with additional <*null*>-to-target sentence pairs generated from non-parallel simplified sentences, doubling the size of the training set (Sennrich et al., 2016a)

TRG2TRG same as BASE but with additional target-to-target sentence pairs (same simplified sentence in source as in target), doubling the size of the training set (Palmero Aprosio et al., 2019) (cf. Section 3)

BT2TRG same as BASE but with additional backtranslated-to-target sentence pairs (source sentence is machine-translated from target sentence), doubling the size of the training set (Sennrich et al., 2016a)

For LINGFEAT, all linguistic features were obtained with *ParZu* (Sennrich et al., 2013), using *clevertagger* (Sennrich et al., 2013) for part-of-speech tags and *Zmorge* (Sennrich and Kunz, 2014) for morphological analysis. The embedding sizes for these features are: 221 for lemmas, 10 each for part-of-speech, morphology, and dependency tags, and 5 for subword BIEO tags, thus extending the total embedding size to 512.

For the backtranslation system, we used the same architecture, the same method, and the same set of sentence pairs as in LINGFEAT, and the added non-parallel sentences were the same for all models trained with augmented data (NULL2TRG, TRG2TRG, BT2TRG).

Moreover, each model type was trained three times, with three different random seeds for shuffling and splitting the training and validation set, in order to reach statistical significance.

After running preliminary trainings, it became clear that all of these models overfit quickly. Validation perplexity regularly reached its minimum before sentences of any kind of fluency were produced, and BLEU scores only started to increase *after* this point. Therefore, we decided to optimize for the BLEU score instead, i.e., stop training when BLEU scores on the validation set reached the maximum. We will discuss more specific implications of this decision in Section 4.4.

4.3 Results of Our Simplication Experiments

We report case-insensitive BLEU and SARI on the validation set, calculated using *SacreBLEU* (Post, 2018). Since we optimized the models for the BLEU score, these values may be taken as a kind of "upper bound" rather than true indicators of their performance.

Figure 1 shows results for the models listed in Section 4.2. TRG2TRG is the only model whose improvements compared to the baseline reached high statistical significance ($p = 0.00014$ for BLEU, $p = 0.00050$ for SARI), although improvements by LINGFEAT look promising ($p = 0.10$ for BLEU, $p = 0.020$ for SARI). The low performance of BT2TRG is surprising, considering the significant BLEU score improvements we observed in a previous experiment with a different German dataset (Battisti et al., 2020). BPE5K and BATCH1K, both proposed as low-resource optimizations in machine translation, do not have much of an effect in this context, either.

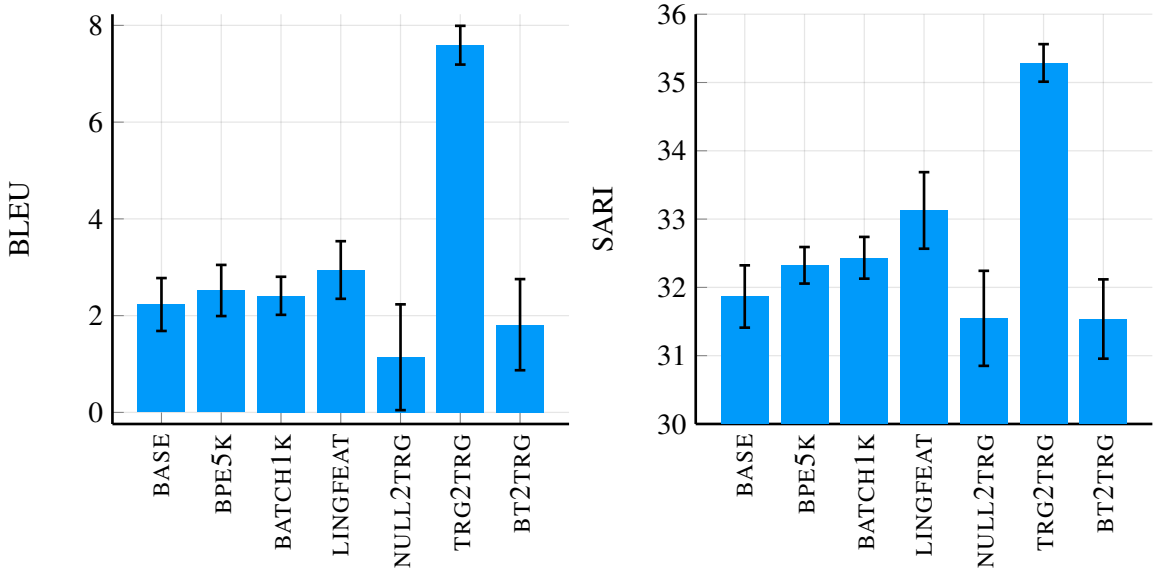

Figure 1: BLEU and SARI scores on the validation set (means and standard errors from three runs)

	BASE		+TRG2TRG		+BT2TRG	
	BLEU	SARI	BLEU	SARI	BLEU	SARI
BASE	2.23 ± 0.55	31.87 ± 0.46	$\mathbf{7.59 \pm 0.40}$	$\mathbf{35.29 \pm 0.28}$	1.81 ± 0.94	31.54 ± 0.58
+LINGFEAT	2.94 ± 0.60	$\mathbf{33.13 \pm 0.56}$	$\mathbf{9.75 \pm 0.63}$	$\mathbf{36.88 \pm 0.67}$	3.11 ± 0.56	$\mathbf{32.96 \pm 0.59}$

Table 3: BLEU and SARI scores of final model configurations on the validation set (means and standard errors from three runs). Bold font indicates significant improvements ($p < 0.05$) with respect to BASE

We also trained additional models which combined the data augmentation methods (TRG2TRG and BT2TRG) with the linguistic features (LINGFEAT) to see if there was a combined effect. The validation scores of all six configurations are presented in Table 3. These results suggest that linguistic features are beneficial even with synthetic data, and that augmentation with target-to-target pairs is more effective than backtranslation.

In addition to automatic evaluation, we translated a test set of 50 sentences using the above models and manually evaluated the output. This was done by the first author, a native speaker of German, with reference to the original sentence along the three criteria shown in Table 4. These are based on Surya et al. (2019) but adapted to capture more specific weaknesses arising from the low-resource setting.

The results are in Figure 2. They provide a clearer picture of the strengths and weaknesses of the configurations. In general, the models have no difficulty producing fluent sentences. However, most of the time, these sentences have little in common with the original but are exact or partial copies of other sentences in the training set. In the worst cases, 60–80% of output sentences are exact copies from the training set. This is a direct consequence of overfitting. Only TRG2TRG (especially in combination with linguistic features) managed to preserve content in a significant portion of the cases. Very often, this was accompanied by decreased fluency in the produced sentences, as in the following examples from the test set, produced by TRG2TRG+LINGFEAT (non-words are marked with '*'):

(1) **Source:**
Die Österreichischen Lotterien rechnen mit rund acht Millionen Tipps auf etwa eineinhalb Millionen Wettscheinen.
('The Austrian lotteries are expecting around eight million games played on about one and a half million play slips.')
Output (TRG2TRG+LINGFEAT):
Die österreichischen loinnen und rechnen mit 8 Millionen Tipps auf nur auf einhalb Millionen Euro mitscheinen scheinen Millionen.
('The Austrian *loinnen and are expecting 8 million games played on only on a half million Euro *withslips slips millions.')

(2) **Source:**
Sobald die Temperaturen steigen und mit Schnee nicht mehr zu rechnen ist, sollten Sommerreifen aufgezogen werden.
('As soon as the temperatures rise and snow is no longer to be expected, summer tires should be mounted.')
Output (TRG2TRG+LINGFEAT):
Sobald die Temperaturen steigen und Schnee nicht mehr zu aufgehoben werden.
('As soon as the temperatures rise and snow no longer to be lifted.')

(3) **Source:**
Aber wir werden unser Öl weiter verkaufen, ... die Sanktionen brechen", sagte Präsident Hassan Rouhani am Montag.
('But we will continue selling our oil, ... break the sanctions," said president Hassan Rouhani on Monday')
Output (TRG2TRG+LINGFEAT):
Aber sie werden nicht Öl weiter verkaufen, sagte am Montag verhani.
('But they will not continue selling oil, said *verhani on Monday.')

In these cases, the system attempts sentence shortening and lexical simplification (note the numeral replacement in Example 1). Generally, the model copies less from training targets (about 10%) and tends more towards transferring tokens from the input.

The results for BT2TRG confirm that backtranslation was not effective in this setting. Given the low content preservation scores in our baseline model for backtranslating, this is not surprising.

4.4 Discussion

As reported in Section 4.2, we optimized our models for BLEU scores. This resulted in models which strongly favored fluency over content preservation by mainly reproducing training material exactly and thus acted more like translation memories. The fact that augmenting the data with simple-to-simple pairs was relatively successful shows that the main difficulty for the other models was finding relevant correspondences between source and target. In the augmented data, these correspondences are trivial to find, and apparently, the model partly succeeded in combining knowledge from this trivial copying job with knowledge about sentence shortening and lexical simplification, as demonstrated by Examples 1–3.

In higher-resource scenarios, a frequent problem is that neural machine translation systems used for text simplification tasks are "over-conservative" (Sulem et al., 2018; Wubben et al., 2012), i.e., they tend to copy the input without simplifying anything. One possible solution to this is to enforce a less probable output during decoding, which is more likely to contain some changes to the input (Štajner and Nisioi, 2018). However, in the present setting, it is

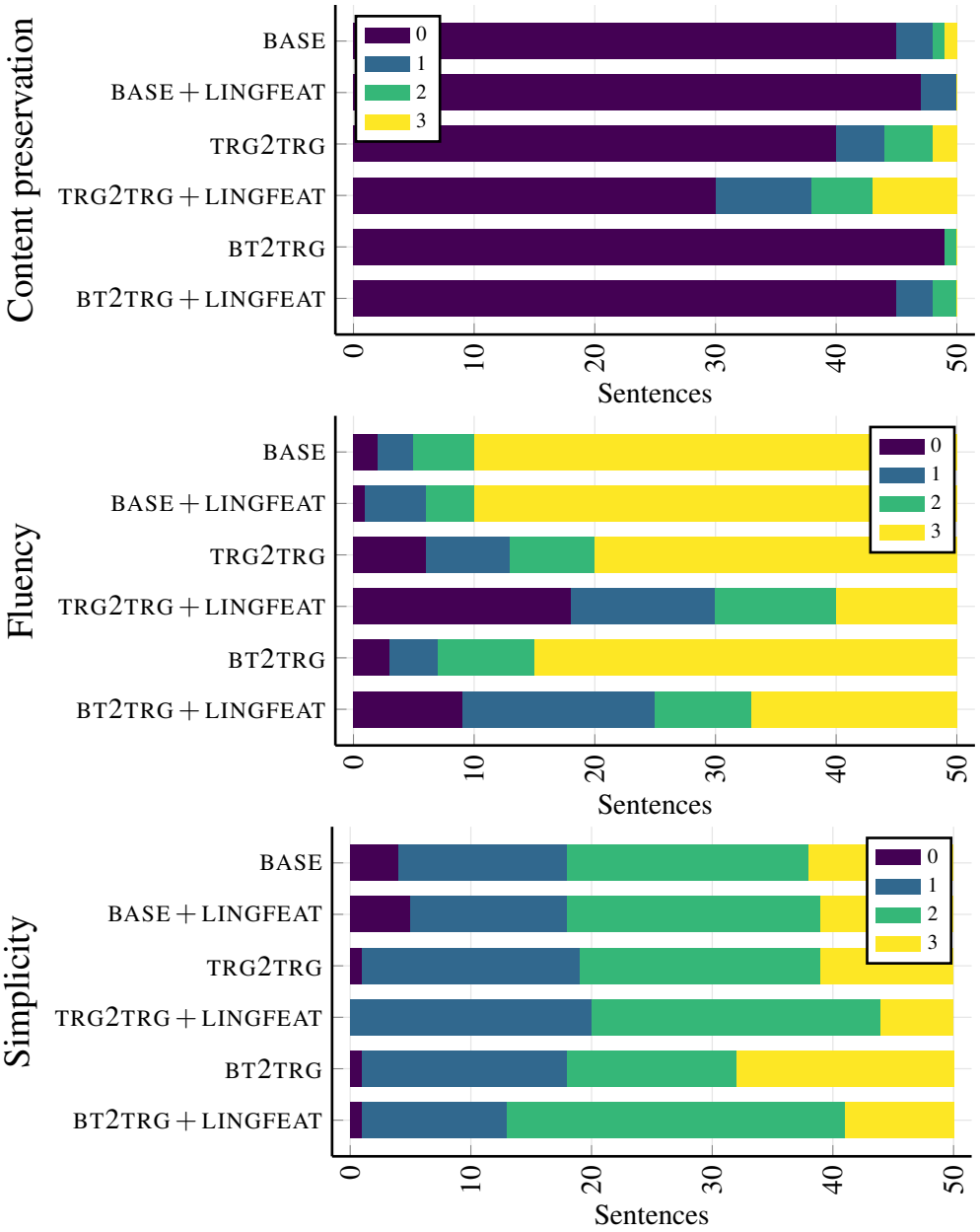

Figure 2: Human evaluation results

Criterion	**Values**	
content preservation	0	no content preserved
	1	general topic preserved, but wrong in specifics
	2	main statement recognizable, but wrong in details
	3	all relevant content preserved
fluency of output	0	gibberish, completely incomprehensible
	1	fluent in parts
	2	mostly fluent (modifying a word or two would make it acceptable)
	3	perfectly natural
relative simplicity	0	more complex than original
	1	equally complex
	2	somewhat simpler
	3	significantly simpler

Table 4: Criteria and values for human evaluation

quite the opposite: The models fail to reproduce most of the content, and adding simple-to-simple pairs can help in this case. However, as datasets grow larger, it may be challenging to balance the effects of real and synthetic data appropriately. To this end, approaches such as the semi-supervised one by Surya et al. (2019), where reconstruction of the input sequence is explicitly built into the model architecture, may be interesting to explore further.

When inspecting the model predictions in the test set, it also became clear that there was a considerable bias towards reproducing one of a handful of sentences in the training set. These are simplified sentences which occur more than once in training, because they are aligned with multiple original sentences. This suggests that including *n*-to-1 alignments in this way is a bad idea for sentence-to-sentence simplification.

Overall, even with a limited quantity of data, our models were able to learn essential features of simplified language, such as lexical substitutions, deletion of less relevant words and phrases, and sentence shortening. Although the performance of the models is not yet mature, these observations give a first idea about which types of texts are important in different settings. In particular, transformations of more complex syntactic structures require substantial amounts of data. When aiming for higher-quality output in low-resource settings, for example, it may be advisable to filter the texts to focus on lexical simplification and deletion, in order not to confuse the model with phenomena it will not learn anyway, and use the discarded sentences for data augmentation instead.

5 Conclusion

This paper introduces the first parallel corpus for data-driven automatic text simplification for German. The corpus consists of 3,616 sentence pairs. Since simplification of Austria Press Agency news items is ongoing, the size of our corpus will increase continuously.

A parallel corpus of the current size is generally not sufficient to train a neural machine translation system that produces both adequate and fluent text simplifications. However, we demonstrated that even with the limited amount of data available, our models were able to learn some essential features of simplified language.

6 Acknowledgments

The authors are indebted to *Austria Presse Agentur* (APA) and *capito* for providing the parallel corpus of standard-language and simplified-language news items.

7 Bibliographical References

Aluisio, S. M. and Gasperin, C. (2010). Fostering Digital Inclusion and Accessibility: The PorSimples project for Simplification of Portuguese Texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53, Los Angeles, CA.

Barlacchi, G. and Tonelli, S. (2013). ERNESTA: A Sentence Simplification Tool for Children's Stories in Italian. In *Proceedings of the 14th Conference on Intelligent Text Processing and Computational Linguistics (CICLing)*, pages 476–487, Samos, Greece.

Battisti, A., Pfütze, D., Säuberli, A., Kostrzewa, M., and Ebling, S. (2020). A Corpus for Automatic Readability Assessment and Text Simplification of German. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC)*, Marseille, France.

Bojanowski, P., Grave, E., Joulin, A., and Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Bott, S. and Saggion, H. (2012). Automatic simplification of Spanish text for e-Accessibility. In *Proceedings of the 13th International Conference on Computers Helping People with Special Needs (ICCHP)*, pages 527–534, Linz, Austria.

Bredel, U. and Maaß, C. (2016). *Leichte Sprache: Theoretische Grundlagen. Orientierung für die Praxis*. Duden, Berlin.

Brouwers, L., Bernhard, D., Ligozat, A., and Francois, T. (2014). Syntactic Sentence Simplification for French. In *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)*, pages 47–56, Gothenburg, Sweden.

Carroll, J., Minnen, G., Canning, Y., Devlin, S., and Tait, J. (1998). Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI'98 Workshop on Integrating aI and Assistive Technology*, pages 7–10.

Chandrasekar, R., Doran, C., and Srinivas, B. (1996). Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 1041–1044, Copenhagen, Denmark.

Coster, W. and Kauchak, D. (2011). Learning to simplify sentences using Wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation (MTTG)*, pages 1–9, Portland, OR.

Council of Europe. (2009). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press, Cambridge.

Gasperin, C., Maziero, E., and Aluisio, S. M. (2010). Challenging Choices for Text Simplification. In *Computational Processing of the Portuguese Language. Proceedings of the 9th International Conference, PROPOR 2010*, pages 40–50, Porto Alegre, Brazil.

Grave, E., Bojanowski, P., Gupta, P., Joulin, A., and Mikolov, T. (2018). Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Hieber, F., Domhan, T., Denkowski, M., Vilar, D., Sokolov, A., Clifton, A., and Post, M. (2017). Sockeye: A toolkit for neural machine translation. *arXiv preprint arXiv:1712.05690*, December.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural computation*, 9(8):1735–1780.

Hwang, W., Hajishirzi, H., Ostendorf, M., and Wu, W. (2015). Aligning Sentences from Standard Wikipedia

to Simple Wikipedia. In *Proceedings of NAACL-HLT*, pages 211–217.

Klaper, D., Ebling, S., and Volk, M. (2013). Building a German/Simple German Parallel Corpus for Automatic Text Simplification. In *ACL Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria.

Nisioi, S., Štajner, S., Ponzetto, S. P., and Dinu, L. P. (2017). Exploring Neural Text Simplification Models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 85–91, Vancouver, Canada, July.

Palmero Aprosio, A., Tonelli, S., Turchi, M., Negri, M., and Di Gangi, M. A. (2019). Neural text simplification in low-resource conditions using weak supervision. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 37–44, Minneapolis, Minnesota, June.

Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA.

Post, M. (2018). A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October.

Saggion, H., Štajner, S., Bott, S., Mille, S., Rello, L., and Drndarević, B. (2015). Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14.

Sennrich, R. and Haddow, B. (2016). Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany, August.

Sennrich, R. and Kunz, B. (2014). Zmorge: A German morphological lexicon extracted from Wiktionary. In *LREC*, pages 1063–1067.

Sennrich, R. and Zhang, B. (2019). Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy, July.

Sennrich, R., Volk, M., and Schneider, G. (2013). Exploiting synergies between open resources for german dependency parsing, pos-tagging, and morphological analysis. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 601–609.

Sennrich, R., Haddow, B., and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Sennrich, R., Haddow, B., and Birch, A. (2016b). Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August.

Specia, L. (2010). Translating from Complex to Simplified Sentences. In *Computational Processing of the Portuguese Language. Proceedings of the 9th International Conference, PROPOR 2010*, pages 30–39, Porto Alegre, Brazil.

Štajner, S. and Nisioi, S. (2018). A detailed evaluation of neural sequence-to-sequence models for in-domain and cross-domain text simplification. In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan.

Štajner, S., Franco-Salvador, M., Rosso, P., and Ponzetto, S. (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)*, pages 3895–3903, Miyazaki, Japan.

Sulem, E., Abend, O., and Rappoport, A. (2018). 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, Melbourne, Australia, July.

Surya, S., Mishra, A., Laha, A., Jain, P., and Sankaranarayanan, K. (2019). Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy, July.

Suter, J., Ebling, S., and Volk, M. (2016). Rule-based Automatic Text Simplification for German. In *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, pages 279–287, Bochum, Germany.

Suter, J. (2015). Rule-based text simplification for German. Bachelor's thesis, University of Zurich.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wubben, S., van den Bosch, A., and Krahmer, E. (2012). Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea, July. Association for Computational Linguistics.

Xu, W., Callison-Burch, C., and Napoles, C. (2015). Problems in Current Text Simplification Research: New Data Can Help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., and Callison-Burch, C. (2016). Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Zhu, Z., Bernhard, D., and Gurevych, I. (2010). A monolingual tree-based translation model for sentence simplification. In *Proceedings of the International Conference on Computational Linguistics*, pages 1353–1361, Beijing, China.