Exploring German Multi-Level Text Simplification

Nicolas Spring, Annette Rios, Sarah Ebling
Department of Computational Linguistics, University of Zurich, Switzerland
{spring,arios,ebling}@cl.uzh.ch

Abstract

We report on experiments in automatic text simplification (ATS) for German with multiple simplification levels along the Common European Framework of Reference for Languages (CEFR), simplifying standard German into levels A1, A2 and B1. For that purpose, we investigate the use of source labels and pretraining on standard German, allowing us to simplify standard language to a specific CEFR level. We show that these approaches are especially effective in low-resource scenarios, where we are able to outperform a standard transformer baseline. Moreover, we introduce copy labels, which we show can help the model make a distinction between sentences that require further modifications and sentences that can be copied as-is.

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 people with cognitive impairment and autism spectrum disorder, prelingually deaf and functionally illiterate people, and sometimes also foreign language learners and children (Bredel and Maaß, 2016).

Automatic text simplification (ATS), 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 has since then been approached by means of rule-based and statistical approaches. As part of the rule-based paradigm, the operations carried out typically include replacing complex lexical and syntactic units by simpler ones. The statistical paradigm so far has mainly conceptualized the simplification task as a case of monolingual (sentence-based) machine translation (MT), i.e., as one of converting standard-language into simplified-language sentences using MT techniques (Specia, 2010). However, while in bilingual parallel texts used for MT, the relation between source and target sentences is mostly 1:1, ATS usually requires n:m alignments with unaligned parts in-between.

ATS research has been documented for English (Zhu et al., 2010), Spanish (Saggion et al., 2015), Portuguese (Aluisio and Gasperin, 2010), French (Brouwers et al., 2014), Italian (Barlacchi and Tonelli, 2013), and other languages. Research on German is still sparse but has gained momentum in recent years due to a number of legal and political developments in German-speaking countries, such as the introduction of a set of regulations for accessible information technology (*Barrierefreie-Informationstechnik-Verordnung*, BITV 2.0) in Germany, the approval of rules for accessible information and communication (*Barrierefreie Information und Kommunikation*, BIK) in Austria, and the ratification of the United Nations Convention on the Rights of Persons with Disabilities (UN CRPD) in Switzerland.

In this paper, we report on work in automatic simplification of standard German into three separate simplification levels (A1, A2, B1) using a sentence-based MT approach. We show that the use of source-side labels indicating the targeted level of simplification benefited performance. Furthermore, pretraining the encoder and decoder on standard German also improved the performance of the ATS models. In our experiments, we noticed that MT models have a tendency to copy the source segments. While copies are sometimes desirable, we want to avoid this in cases where the original segment could benefit from further simplification.

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

Proceedings of Recent Advances in Natural Language Processing, pages 1339–1349
Sep 1–3, 2021.
https://doi.org/10.26615/978-954-452-072-4_150

We show that the use of special copy labels at training time can positively influence such behavior.

In particular, the contributions of the paper at hand are the following:

- We demonstrate the use of source-side Common European Framework of Reference for Languages (CEFR) (Council of Europe, 2009) labels and a fine-tuning approach to boost text simplification performance for certain CEFR levels.

- We investigate the use of source-side copy labels to reduce the copying behaviour of text simplification models in situations where copying is not desirable.

The remainder of this paper is structured as follows: Section 2 describes existing datasets for text simplification for a variety of languages as well as established approaches to ATS. Section 3 describes our approach to multi-level text simplification for German. We discuss our experiments in Section 4 and conclude in Section 5 with further thoughts on improving ATS for German and current challenges to overcome.

2 Previous Work: Automatic Text Simplification

2.1 Data

ATS with sentence-based MT models relies on pairs of standard-language/simplified-language texts aligned at the sentence level. 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 ATS 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 about 108,000 sentence pairs. Application of the corpus has been subject to criticism 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 with more parallel data, additional monolingual-only data (in simplified German), and new information on text structure (e.g., paragraphs, lines), typography (e.g., font type, font style), and images (content, position, and dimensions).[2] The corpus is compiled from Web sources in Germany, Austria, and Switzerland. The sources mostly represent websites of governments, specialized institutions, and non-profit organizations. The documents cover a wide range of topics, such as politics (e.g., instructions for voting), health (e.g., what to do in case of pregnancy), and culture (e.g., introduction to art museums). The corpus contains 6,217 documents (5,461 monolingual documents plus 378 documents for each side of the parallel data). The vocabulary of the simplified German texts is smaller than that of the German texts by 51% (33,384 vs. 16,352 types), which is comparable to the rate of reduction reported for the Newsela Corpus (50.8%).

Säuberli et al. (2020) introduced a corpus of news items from the Austria Press Agency (*Austria Presse Agentur*, APA). At APA, four to six news items per day are manually simplified into two language levels, B1 and A2, following guidelines by *capito*, the largest provider of simplification services (translations and translators' training) in Austria, Germany, and Switzerland.[3] These news

[2] The importance of the latter type of information has been stressed repeatedly, e.g. for automatic readability assessment (Arfé et al., 2018; Bock, 2018; Bredel and Maaß, 2016).

[3] `https://www.capito.eu/`. *capito* distinguishes between three levels along the CEFR: A1, A2, and B1. Each level is linguistically operationalized, i.e., specified with respect to linguistic constructions permitted or not permitted at the respective level. Note that while the CEFR was designed to measure foreign language skills, with simplified language, it is partly applied in the context of first-language acquisition (Bredel and Maaß, 2016).

items cover the topics of politics, economy, culture, and sports.

A number of tools exist for sentence alignment of parallel documents in the context of sentence simplification; among them are CATS (Štajner et al., 2018), MASSAlign (Paetzold et al., 2017), and LHA (Nikolov and Hahnloser, 2019). Spring et al. (2021) evaluated these alignment methods for German text simplification, together with SBERT (Reimers and Gurevych, 2020) and Vecalign (Thompson and Koehn, 2019). Both of the latter tools were originally designed in the context of multilingual alignment. Evaluation against a human-created gold standard showed that LHA yielded the most accurate sentence alignments.

2.2 Approaches

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

Nisioi et al. (2017) pioneered neural machine translation (NMT) models for ATS, performing experiments with LSTMs on both the Wikipedia dataset of Hwang et al. (2015) and the Newsela Corpus for English, with automatic alignments derived from CATS (cf. Section 2.1).

More recent contributions to ATS include explicit edit operation modeling (Dong et al., 2019), graded simplification (Nishihara et al., 2019), weakly supervised (Palmero Aprosio et al., 2019), and unsupervised approaches (Surya et al., 2019; Kumar et al., 2020).

Suter et al. (2016) introduced a rule-based ATS system for German. Their rules are based on linguistically motivated guidelines and their simplification system yielded outputs with a syntactic complexity comparable to a human translation.

Battisti et al. (2020) presented an approach to German ATS using recurrent neural networks with attention and incorporated back-translation (Sennrich et al., 2016) to generate additional synthetic training data from the monolingual part of their corpus.

Säuberli et al. (2020) presented the first approach to ATS for German using (sentence-based) NMT models. As data, they used the APA Corpus introduced in Section 2.1, amounting to approximately 3,500 sentence pairs.

Other contributions that are relevant to our work originate from the field of MT. Source-side labels have previously been employed in a variety of tasks such as domain adaption (Kobus et al., 2017), multilingual translation (Johnson et al., 2017), and to improve training with back-translated data (Caswell et al., 2019).

2.3 Evaluation

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 for automatic evaluation of MT, computes token n-gram overlap between a hypothesis and one or multiple references. A shortcoming of BLEU with respect to ATS is that it does not punish hypotheses that are identical to the input. In contrast, SARI was introduced specifically for ATS and is designed to punish excessive copying behaviour. SARI considers the input and rewards tokens in the hypothesis that do not occur in the input but in one of the references (addition), as well as tokens in the input that are correctly retained (copying) or removed (deletion) in the hypothesis.[4]

Table 1 displays scores for previous sentence-level ATS systems for different languages.

3 Text Simplification Along CEFR Levels for German

3.1 Data

The data used for the experiments reported in this paper consists of two collections:

The first part comprises an expanded version of the Austria Press Agency (*Austria Presse Agentur*, APA) corpus described by Säuberli et al. (2020) (cf. Section 2.1). Our updated version of this corpus consists of standard-language news items with their corresponding simplifications between August 2018 and April 2021. We extracted simplified German documents along with their standard German counterparts. This extraction yielded 2,410 document pairs for B1 and 2,347 for A2. As human text simplification work at the APA is ongoing, this corpus is expected to grow with time.

The second part of our data consists of the *capito* corpus. As a provider of simplification services,

[4] A copy or deletion is considered correct if the token is copied or deleted in at least one of the references.

Author(s)	Language	Approach	Scores
Specia (2010)	Portuguese	SMT	60.75 BLEU
Coster and Kauchak (2011)	English	SMT	60.46 BLEU
Wubben et al. (2012)		PBMT	34.07 SARI (Nisioi et al., 2017) 67.79 BLEU (Nisioi et al., 2017)
Xu et al. (2016)	English	SBMT	38.59 SARI (Nisioi et al., 2017) 73.62 BLEU (Nisioi et al., 2017)
Nisioi et al. (2017)	English	NMT	87.50 BLEU
Štajner and Nisioi (2018)	English	NMT	Newsela: 89.49 BLEU 36.48 SARI PWKP: 84.69 BLEU 35.78 SARI
Säuberli et al. (2020)	German	NMT	9.75 BLEU 36.88 SARI

Table 1: Automatic evaluation scores for sentence-level ATS systems (PBMT: phrase-based SMT; SBMT: syntax-based MT).

capito produces a high number of professional simplifications for a variety of documents and text genres. This includes, but is not limited to, booklets, information texts, websites and legal texts, which are manually simplified into one or more levels following the *capito* guidelines. The simplification levels in this corpus include B1, A2 and A1. We extracted simplified German documents along with their standard German counterparts and metadata. Currently, the corpus contains 1,245 document pairs for B1, 1,885 for A2 and 879 for A1, however, since *capito* provides ongoing translation services, the number of documents is constantly increasing.

3.2 Sentence Alignment

Sentence alignment for ATS includes some phenomena that do not occur in this form in sentence alignment for translation. Whereas in translation, the standard case is often a simple 1:1 correspondence, alignment for text simplification can be considered n:m, meaning that a single alignment can consist of a varying number of segments on each side. This is due to phenomena such as sentence splitting and compression, additional explanations, as well as the fact that the order of information can change.

Alignment	Dataset	# Sentences
OR-B1	APA	10,268
OR-B1	capito	54,224
OR-A2	APA	9,456
OR-A2	capito	136,582
OR-A1	capito	10,952

Table 2: Parallel corpus extracted from the different datasets and simplification levels (OR: original, i.e., complex German).

The results of NMT experiments are highly dependent on the available data. We extracted sentence alignments from our corpora using the LHA alignment method (Nikolov and Hahnloser, 2019), which was shown to yield the best results for simplified German (Spring et al., 2021) (cf. Section 2.1). For calculating the alignments, we used the Sentence Alignment Tools Evaluation Framework (SATEF),[5] which yields n:m alignments, meaning that a single alignment can consist of a varying number of segments on each side. We aligned our documents in the direction from complex to simple.

The number of documents differs considerably

[5] Code is available from: https://github.com/kostrzmar/SATEF

depending on the CEFR level and the dataset, see Section 3.1. Furthermore, the APA Corpus does not contain any data for level A1. This manifests itself in the number of sentence alignments we were able to extract on this level. The largest number of alignments in our parallel corpus are for level A2, followed by about half as many for level B1. With 10,952 sentence alignments, A1 is the simplification level with the smallest amount of data available for model training, see Table 2 for an overview.

The sentence alignments with LHA on the APA data are publicly available.[6]

3.3 Text Simplification

All the models we trained for our experiments shared the same architecture and hyperparameters. We trained transformer models (Vaswani et al., 2017) with five layers, four attention heads, 512 hidden units in the transformer layers, and a feed forward layer size of 2048. Embedding dropout and label smoothing were set to 0.3. We used early stopping according to BLEU on a held-out development set with a patience of 10. All models shared a 20k vocabulary between source and target. All our experiments were carried out in sockeye (Hieber et al., 2018).

We trained baseline models where we combined all available training data across all levels. These models had no explicit method to determine the desired level of simplification on the target side.

The diverse dataset allowed us to treat text simplification as a number of subtasks, where the model learns to simplify into different complexity levels, ranging from B1 to A1 according to the CEFR. To allow a model to make a distinction between the different levels of simplification, we used source-side labels indicating the desired CEFR level of the target segment (`<b1>`, `<a2>`, and `<a1>`). To better understand the copying behavior of our models, we trained our labeled models in two versions: 1) using a simple source-side label indicating the target CEFR level and 2) additionally using an explicit `<copy>` label instead of the CEFR level for all segments where source and target were identical. Apart from these modifications to the training data, all model hyperparameters were identical to the baseline models. We will refer to these models as "APA+capito multi" and "APA+capito multi copy", respectively. Note that copying the source segment to the target is not wrong *per se* and that there are many cases where no further modification of a segment is needed, especially at higher CEFR levels. We hypothesize that the addition of these copy labels at training time allows the model to better recognize these cases even when they are not present at test time. At test time, all segments were translated with their CEFR label and no `<copy>` labels were present. We observed that training a simplification model with explicit `<copy>` labels reduced the number of untranslated segments where source and reference are not identical. We treated these types of copies as undesired.

All experiments previously described were performed in two variations. In the first variation, we trained the models from scratch on the simplification data. The second variation involved pretraining the encoder and decoder on a DE→EN or EN→DE translation task, respectively. This was motivated by the relatively low number of aligned segments we could use for our parallel training data. We trained two translation models with the same hyperparameters as the simplification models, but we used separate source and target vocabularies for encoder and decoder (German only). The parallel DE↔EN data for pretraining the NMT models (cf. Section 3.3) consisted of Europarl v10, Common Crawl, News Commentary v15, and the Tilde Rapid Corpus from WMT20.[7] For the simplification models, we then initialized the parameters of the encoder with the encoder parameters of the DE→EN model. Likewise, we initialized the decoder parameters of the simplification models with the decoder parameters of the EN→DE model.[8] The DE→EN, EN→DE and all simplification models used the same German subword vocabulary. We then fine-tuned these pretrained models on our text simplification data. We append the tag "fine-tuned" to the name of these models.

Finally, for the purpose of reproducibility, we trained a labeled simplification model on the publicly available APA alignments[9] described in Section 3.2, referred to as "APA multi".[10]

For evaluation, we used a test set that consists of

[6] `https://zenodo.org/record/5148163`

[7] `http://www.statmt.org/wmt20/translation-task.html`

[8] The cross-attention heads were initialized randomly.

[9] See footnote 6.

[10] Preprocessing and training scripts are available from: `https://github.com/ZurichNLP/RANLP2021-German-ATS`

500 parallel segments per level (A1, A2 and B1), randomly sampled from the combined corpus. The "APA multi" model was evaluated on a different test set, consisting exclusively of APA data.

3.4 Results

Our results can be found in Tables 3 and 4. In general, the models with target labels for the simplification levels performed better than the baseline both in terms of BLEU and SARI, with the notable exception of BLEU at level A2. Note that A2 was by far the most common simplification level in our dataset. The simple labeling approach of "APA+capito multi" was already effective and outperformed both baseline models on A1 and B1. But it was in turn outperformed by its fine-tuned counterpart "APA+capito multi fine-tuned" on B1, and by the fine-tuned model with copy labels, "APA+capito multi copy fine-tuned", on A1. Pretraining the level-agnostic baseline model yielded improvements in terms of BLEU for A1 and A2 and only for A2 in terms of SARI.

When evaluating our labeled models with SARI, we could see improvements over the performance of the baselines for all models. Generally, SARI scores suggest that pretraining is especially effective when combined with labeling on levels A2 and B1. On A1, the simple labeling approach of "APA+capito multi" remained the most effective. It yielded an improvement of 6.86 points over the baseline. The best model for A2 and B1 was "APA+capito multi fine-tuned", which yielded SARI scores improved between 5.49 (A2) and 7.55 (B1).

An analysis of the copying behaviour of the different models can be found in Table 5. The standard fine-tuned model generally had the strongest tendency to copy the source, however, the addition of copy labels significantly reduced the number of copied segments. This was also true for the non-pretrained model variants. Furthermore, the number of undesired copies (where source and reference are not identical) decreased with the use of copy labels (percentage decreases as indicated in Table 5). This was true for both the generic and the pretrained models. In general, the models tended to produce more copies for higher CEFR levels, which was consistent with the training data: In B1, or even A2, shorter segments are more often identical to their standard German counterparts than in A1.

Table 6 shows two examples of how the models with copy labels can avoid source copies. For both samples, "APA+capito multi fine-tuned" simply copies the input. "APA+capito multi fine-tuned copy" avoids this by using two different strategies. In the first example, it produces a segment with different structure and content, which is related to the source segment thematically. Such outputs are common across all models and can be seen as a result from the many-to-many nature of alignment for ATS and the elaborations that are common in text simplification. In the second example, the model produces a shorter simplification by removing some of the information present in the source segment. This ellipsis is another common phenomenon in text simplification.

The performance and copying behaviour of the "APA multi" model cannot be directly compared to the other models because it was trained on different data (exclusively the APA corpus) and uses a different test set.

4 Discussion

Comparing our results to Säuberli et al. (2020), whose experiments are similar to ours in terms of scope and data, it is clear that our baseline models are already quite strong. We find that source-side labels for target language levels generally improve BLEU and SARI scores and that the same is true for pretraining and fine-tuning. Interestingly, combining labels and pretraining results in lower gains in both metrics on A1 for the model without copy labels, indicating that these two approaches cannot always simply be combined. We also note that the scores on CEFR level A2 did not profit as much from the different strategies and we were not able to improve over the pretrained baseline model in terms of BLEU by using labels. We attribute this to the relatively large amount of training data for A2, meaning that specifically dealing with this low resource setting was not needed for this CEFR level. On the other hand, A1 and B2, for both of which there was substantially less data, benefit from labeling and pretraining.

Regarding the copying behaviour, we note higher numbers of direct source copies on the higher CEFR levels. We attribute this to the fact that simplifications on these higher levels typically tend to be closer to the original text in terms of lexical and syntactic complexity. This means that there are more standard language segments without any

Model	**BLEU A1**	**BLEU A2**	**BLEU B1**
APA multi		*15.2*	*12.3*
Baseline	13.4	14.4	16.3
Baseline fine-tuned	13.5	**14.9**	15.7
APA+capito multi	14.2	14.1	17.2
APA+capito multi copy	14.0	14.0	15.2
APA+capito multi fine-tuned	13.9	14.2	**17.5**
APA+capito multi copy fine-tuned	**14.3**	12.4	13.9

Table 3: BLEU scores of the different models. The APA multi model was trained and evaluated on different data and is not comparable to the rest of the models.

Model	**SARI A1**	**SARI A2**	**SARI B1**
APA multi		*42.04*	*40.73*
Baseline	36.26	36.11	34.53
Baseline fine-tuned	36.21	36.99	33.98
APA+capito multi	**43.12**	41.53	41.81
APA+capito multi copy	43.11	41.52	40.68
APA+capito multi fine-tuned	42.88	**41.60**	**42.08**
APA+capito multi copy fine-tuned	42.86	40.86	40.48

Table 4: SARI scores of the different models. The APA multi model was trained and evaluated on different data and is not comparable to the rest of the models.

need for changes. Furthermore, copy labels are effective in reducing the number of copies overall as well as specifically reducing undesired copies where the reference differs from the source. Since the model never sees A1, A2 or B1 target labels where the target is identical to the source, it is less inclined to produce a copy for these labels at test time. The qualitative analysis of the copying behaviour showed that models with copy labels can avoid copies by creating thematically related output or by leaving out information present in the source. It is important to note that with the exception of BLEU on A1, the copy-labeled models did not clearly perform better than their counterparts in terms of BLEU and SARI. This could be partly due to the operations just mentioned, which result in fewer words in common with the source and conceivably also the reference. Automatic metrics provide a good estimate for the quality of the simplification, however, for a more accurate analysis, e.g. of the copying, we plan to conduct a human evaluation in collaboration with the experts at *capito*. Also, while SARI was introduced to punish excessive copying behaviour, it is not clear how suitable it is for the evaluation of our methods.

5 Conclusion and Outlook

We were able to demonstrate the advantages of different approaches to German multi-level ATS. We established strong baselines on a generic simplification task across all CEFR levels and were able to further boost model performance for specific levels of simplification using source-side labels and a pretraining/fine-tuning strategy. We tested fine-tuning with labeled multi-level models. These approaches were generally more effective on the CEFR levels where we had more limited data, suggesting that they are especially useful in low-resource scenarios. We also investigated the use of copy labels at training time to mark segments where source and target segments are identical. At test time, this resulted in a lower number of copies overall and especially in the number of instances where the references differs from the source. This suggests that copy labels are a valid tool to reduce undesired copying behaviour in text simplification, though their influence on the quality of the output can likely only be determined by human evaluation.

Further work will be conducted on a more advantageous combination of the two approaches of

Model	# A1	# A1*	% Undesired A1	# A2	# A2*	% Undesired A2	# B1	# B1*	% Undesired B1
APA multi				*4*	*3*	*75.00%*	*2*	*1*	*50.00%*
APA+capito multi	58	47	81.03%	60	44	73.34%	99	75	75.76%
APA+capito multi copy	39	31	79.49%	36	25	69.45%	69	52	75.36%
APA+capito multi fine-tuned	62	50	80.65%	74	59	79.73%	107	83	77.57%
APA+capito multi copy fine-tuned	35	**27**	77.14%	34	**24**	70.59%	57	**42**	73.68%

Table 5: The number of source copies produced by the models. Columns marked with an asterisk only count copies where the source is different from the reference (i.e., undesired copies). The APA multi model was trained and evaluated on different data and is not comparable to the rest of the models.

Model	German	English
Source	Lernen von der Natur!	Learning from nature!
APA+capito multi fine-tuned	Lernen von der Natur!	Learning from nature!
APA+capito multi fine-tuned copy	Wie funktioniert der Austausch von Wissen?	How does the exchange of knowledge work?
Source	Die praktische Fahrradprüfung findet im Grazer Verkehrsgarten im Stadtpark statt.	The practical bicycle test takes place in the Graz traffic garden in the city park.
APA+capito multi fine-tuned	Die praktische Fahrradprüfung findet im Grazer Verkehrsgarten im Stadtpark statt.	The practical bicycle test takes place in the Graz traffic garden in the city park.
APA+capito multi fine-tuned copy	Die praktische Fahrradprüfung findet im Grazer Verkehrsgarten statt.	The practical bicycle test takes place in the Graz traffic garden.

Table 6: Model simplification examples (level A2) comparing the two fine-tuned models with labels in cases where "APA+capito multi fine-tuned" simply copies the source segment. The source is identical to the simplification produced by "APA+capito multi fine-tuned" shown here.

labeling and pretraining with the goal of moving away from systems specialized in simplifying a single CEFR level and arriving at a single system with state-of-the-art performance in all CEFR levels. We will also conduct further experiments aimed at refining the ability of an NMT model to make a distinction between desired and undesirable copies.

Acknowledgements

The authors want to thank the Austria Presse Agentur and the CFS GmbH for providing the parallel corpora of standard German documents with their simplified counterparts. The authors are grateful to Marek Kostrzewa for his help in creating the sentence alignments for the corpus and to the anonymous reviewers for their valuable input and feedback. This research was funded by the Austrian Research Promotion Agency (Österreichische Forschungsförderungsgesellschaft, FFG) General Programme under grant agreement number 881202.

References

Sandra Maria Aluisio and Caroline Gasperin. 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.

Barbara Arfé, Lucia Mason, and Inmaculada Fajardo. 2018. Simplifying informational text structure for struggling readers. *Reading and Writing*, 31(9):2191–2210.

Gianni Barlacchi and Sara Tonelli. 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.

Alessia Battisti, Dominik Pfütze, Andreas Säuberli, Marek Kostrzewa, and Sarah Ebling. 2020. A corpus for automatic readability assessment and text simplification of german. In *Proceedings of The 12th Language Resources and Evaluation Confer-*

ence, pages 3295–3304, Marseille, France. European Language Resources Association.

Bettina M. Bock. 2018. "Leichte Sprache" – Kein Regelwerk. Sprachwissenschaftliche Ergebnisse und Praxisempfehlungen aus dem LeiSA-Projekt. Technical report, Universität Leipzig.

Stefan Bott and Horacio Saggion. 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.

Ursula Bredel and Christiane Maaß. 2016. *Leichte Sprache: Theoretische Grundlagen. Orientierung für die Praxis*. Duden, Berlin.

Laetitia Brouwers, Delphine Bernhard, Anne-Laure Ligozat, and Thomas François. 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.

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 the AAAI'98 Workshop on Integrating aI and Assistive Technology*, pages 7–10.

Isaac Caswell, Ciprian Chelba, and David Grangier. 2019. Tagged back-translation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)*, pages 53–63, Florence, Italy. Association for Computational Linguistics.

Raman Chandrasekar, Christine Doran, and Bangalore Srinivas. 1996. Motivations and methods for text simplification. In *Proceedings of the 16th Conference on Computational Linguistics*, pages 1041–1044, Copenhagen, Denmark.

William Coster and David Kauchak. 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.

Yue Dong, 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, Florence, Italy. Association for Computational Linguistics.

Caroline Gasperin, Erick Maziero, and Sandra M. Aluisio. 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.

Felix Hieber, Tobias Domhan, Michael Denkowski, David Vilar, Artem Sokolov, Ann Clifton, and Matt Post. 2018. The sockeye neural machine translation toolkit at AMTA 2018. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 200–207, Boston, MA. Association for Machine Translation in the Americas.

William Hwang, Hannaneh Hajishirzi, Mari Ostendorf, and Wei Wu. 2015. Aligning Sentences from Standard Wikipedia to Simple Wikipedia. In *Proceedings of NAACL-HLT*, pages 211–217.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

David Klaper, Sarah Ebling, and Martin Volk. 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.

Catherine Kobus, Josep Crego, and Jean Senellart. 2017. Domain control for neural machine translation. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 372–378, Varna, Bulgaria. INCOMA Ltd.

Dhruv Kumar, 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, ACL 2020, Online, July 5-10, 2020*, pages 7918–7928. Association for Computational Linguistics.

Nikola I. Nikolov and Richard Hahnloser. 2019. Large-scale hierarchical alignment for data-driven text rewriting. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 844–853, Varna, Bulgaria. INCOMA Ltd.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable Text Simplification with Lexical Constraint Loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. 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*, pages 85–91, Vancouver, Canada.

Gustavo Paetzold, Fernando Alva-Manchego, and Lucia Specia. 2017. Massalign: Alignment and annotation of comparable documents. In *Proceedings of the IJCNLP 2017, Tapei, Taiwan, November 27 - December 1, 2017, System Demonstrations*, pages 1–4. Association for Computational Linguistics.

Alessio Palmero Aprosio, Sara Tonelli, Marco Turchi, Matteo Negri, and Mattia A. Di Gangi. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318, Philadelphia, PA.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarević. 2015. Making it Simplext: Implementation and evaluation of a text simplification system for Spanish. *ACM Transactions on Accessible Computing (TACCESS)*, 6(4):14.

Andreas Säuberli, Sarah Ebling, and Martin Volk. 2020. Benchmarking data-driven automatic text simplification for German. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 41–48, Marseille, France. European Language Resources Association.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving Neural Machine Translation Models with Monolingual Data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 86–96, Berlin, Germany.

Lucia Specia. 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.

Nicolas Spring, Dominik Pfütze, Marek Kostrzewa, Alessia Battisti, Annette Rios, and Sarah Ebling. 2021. Comparing Sentence Alignment Methods for Automatic Simplification of German Texts. Presentation given at the 1st International Easy Language Day Conference (IELD), Germersheim, Germany.

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)*, pages 3895–3903, Miyazaki, Japan.

Sanja Štajner and Sergiu Nisioi. 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.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised Neural Text Simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy.

Julia Suter, Sarah Ebling, and Martin Volk. 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.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. 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)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

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 (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. 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).

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

A Model Hyperparameters

Hyperparameter	Value
architecture	transformer
seed	1
patience	10
optimized metric	BLEU
batch type	word
batch size	2048
update frequency	2
optimizer	adam
max length	95:95
label smoothing	0.3
vocab	20k
layers	5:5
model size	512:512
heads	4:4
ff	2048
dropout attention	0.1
dropout-act	0.0
dropout-prepost	0.1
embedding dropout	0.3
positional embeddings	fixed
initial lr	0.0002
learning-rate-reduce-factor	0.9
learning-rate-scheduler	plateau-reduce
init	xavier
init-scale	3.0
init-xavier-factor-type	avg

Table 7: Model hyperparameters.