Language Models for German Text Simplification: Overcoming Parallel Data Scarcity through Style-specific Pre-training

Miriam Anschütz, Joshua Oehms, Thomas Wimmer,
Bartłomiej Jezierski and **Georg Groh**
School for Computation, Information and Technology
Technical University of Munich, Germany
{miriam.anschuetz, joshua.oehms, thomas.m.wimmer, b.jezierski}@tum.de
grohg@in.tum.de

Abstract

Automatic text simplification systems help to reduce textual information barriers on the internet. However, for languages other than English, only few parallel data to train these systems exists. We propose a two-step approach to overcome this data scarcity issue. First, we fine-tuned language models on a corpus of German Easy Language, a specific style of German. Then, we used these models as decoders in a sequence-to-sequence simplification task. We show that the language models adapt to the style characteristics of Easy Language and output more accessible texts. Moreover, with the style-specific pre-training, we reduced the number of trainable parameters in text simplification models. Hence, less parallel data is sufficient for training. Our results indicate that pre-training on unaligned data can reduce the required parallel data while improving the performance on downstream tasks.

1 Introduction

Automatic text simplification (ATS) is the task of simplifying a text's lexical and structural complexity while preserving its original meaning. Easy-to-read texts can help people with learning deficiencies or non-native speakers gain access to texts that they could not understand otherwise. On the one hand, ATS can be used to create assisting tools for people with reading disabilities or professional translators (Suárez-Figueroa et al., 2022). On the other hand, ATS can be applied as a preprocessing step for other natural language processing tasks such as machine translation or information retrieval to improve their performances (Štajner and Popovic, 2016), making it an important field of study.

In German, there exist multiple levels of simplified language. In contrast to the underspecified simple language, the so-called Leichte Sprache (Easy Language) enforces a very strong simplification level and follows predefined structural rules (Netzwerk Leichte Sprache, 2013). These rules include conveying only one message per sentence (structural simplification), restriction to common words (lexical simplification), and usage of simplified grammar (syntactical simplification). This simplified grammar breaks with standard German grammar, for example, by using dative instead of genitive to indicate possession. We consider Easy Language as a standalone language style. Therefore, we refer to Easy Language data as monolingual data in the further course of the paper, even though it is German as well.

This work shows the benefits of fine-tuning language models for specific styles and characteristics. We publish and discuss a collection of causal language models fine-tuned for German Easy Language. As shown in previous work (Gururangan et al., 2020), pre-training language models for specific domains can benefit the performances of downstream tasks in the respective domain. We extend this analysis to the language style of Easy Language. In addition, the fine-tuned models can be used to generate text with the specificities of Easy Language, for example, in data augmentation applications. Finally, we present how these models can serve as plug-in-decoders in BART-like architectures (Lewis et al., 2020) to speed up and improve the training on sequence-to-sequence (seq2seq) tasks. Therefore, our contributions are the following:

- We publish five German Easy Language causal language models and extensively evaluate their language style adaptions.
- We assess the models' performance on the two downstream tasks of text complexity prediction and text simplification.
- We suggest an ATS training process that exploits our pre-trained language models. This process reduces the number of trained param-

Findings of the Association for Computational Linguistics: ACL 2023, pages 1147–1158
July 9-14, 2023

eters by over 90% while preserving state-of-the-art performance.

With the reduction of trainable parameters, less aligned data is needed to train an ATS system. Especially for languages other than English, where aligned data is sparse, pre-trained causal language models can improve ATS performance. We publish our code and results for further research and application[1].

2 Related work

Causal language models can complete text based on a prompt. In contrast to masked language models, where the models know about the context before and after a specific token, these causal language models rely only on the input and the previously outputted tokens. Therefore, they are called autoregressive models. The Generative Pre-trained Transformer (GPT) (Radford et al., 2019) is a prominent example of such an autoregressive language model. It was trained on a collection of web data and, thus, outputs text for general purposes. Previous work has fine-tuned GPT for multiple domains and tasks, such as the task of quest generation in games (Värtinen et al., 2022) or the medical domain (Schneider et al., 2021). In addition to domain adaption, GPT was tailored to specific text styles and characteristics. These style transfer approaches include fine-tuning for poem generation (Liao et al., 2019) or the reduction of non-normative clauses (Peng et al., 2020). Li et al. (2022) trained a GPT model to mimic the language of people with dementia. By calculating the perplexities of texts with the fine-tuned and original version, they could distinguish samples from healthy and diseased people.

Sun and Wan (2022) adapted a language model for simple language by only masking easy-to-understand words in training. However, this model is a masked language model that can only fill in blanks and not generate text from scratch. Most similar to our work is the TransformerLM by Maruyama and Yamamoto (2019) trained for Japanese text simplification. The authors used a parallel corpus to directly fine-tune a GPT model for simplification. In contrast, our models are fine-tuned on monolingual Easy Language data. Therefore, they do not require alignments and can be used for a broader range of tasks.

[1] https://github.com/MiriUll/Language-Models-German-Simplification

2.1 German Text simplification

In contrast to the English language, automatic text simplification in German has seen little research. The first system for Easy Language was proposed by Suter et al. (2016) and consisted of a collection of hand-crafted rules, including sentence splitting and paraphrasing. Säuberli et al. (2020) published the first neural simplification approach based on the transformer architecture, together with an aligned corpus. They discussed multiple data augmentation strategies, but their results lacked fluency and content preservation. Based on an extended version of this dataset, Spring et al. (2021) built a controllable simplification system that can output different simplification levels based on the Common European Framework of References for Languages (CEFR), but not specifically Easy Language. Finally, Rios et al. (2021) proposed a modified mBART architecture for document-level simplification. In our paper, we adopted their architecture to evaluate our language models on the downstream task of ATS.

3 Datasets

Several sources are available in Easy Language; however, they mostly encompass news websites, and only a few are aligned with articles in standard German. In the following sections, we detail the information on the data used in our training, including the Easy Language monolingual corpus utilized for fine-tuning German language models and the parallel corpus for the downstream task of text simplification. The dataset utilized for the downstream task of text complexity prediction is publicly available as a part of the GermEval 2022 shared task (Mohtaj et al., 2022) (refer to Subsection 5.4). We published scrapers to recreate our sources for the use of the academic community[2]. We also provide an overview of available monolingual and parallel data sources for simplified German beyond our training data in Appendix A.

3.1 Monolingual corpus

An overview of the available monolingual data can be found in Table 1. The publicly available Easy Language datasets are very limited: The Simple German corpus published by Toborek et al. (2022) contains texts on health and medication, public administration, politics, information texts for disabled people, and news articles. The second publicly available resource is a small corpus published by

[2] https://github.com/brjezierski/scrapers

Siegel et al. (2019). It contains election programs, excerpts from the Bible, children's stories, and Red Cross documents.

Kurier, InfoEasy, and NDR are public broadcasting services in Austria, Switzerland, and northern Germany, respectively, and have specific columns in Easy Language. In addition, Hurraki and Lebenshilfe offer online dictionaries in Easy Language, while Einfachstars contains news articles about celebrities. These three data sources diversify our covered domains and styles of writing. More details about the data sources can be found in Table 8 in Appendix A. Our fine-tuning data combines all sources included in Table 1. The combined data was shuffled and randomly split into a training set containing 90% of the data and a validation set with 10% of the total.

Dataset	Sentences	Domain
Hurraki	56,785	lexicon
Lebenshilfe	7,144	lexicon
Einfachstars	129,674	news
Nachrichtenleicht	122,842	news
Kurier	67,827	news
NDR	60,749	news
InfoEasy	10,310	news
Siegel et al. (2019)	4,210	misc.
Toborek et al. (2022)	28,356	misc.
Total	**544,467**	

Table 1: Overview of the monolingual data used for language model fine-tuning.

3.2 Parallel corpus

For training the text simplification model, we used the publicly available 20 Minuten dataset[3]. The dataset consists of full articles paired with shortened, simplified summaries from the Swiss news magazine 20 Minuten. It comprises 17,905 article pairs in the training dataset and 200 pairs in the validation and test set each (Rios et al., 2021). The dataset's compression ratio (the reduction in the word count of simplified summaries) was estimated at 11%.

3.3 Preprocessing pipeline

Analyzing the outputs of publicly available language models in standard German, we noticed that in many cases, especially for the news headline-like input, the output contained noise, such as HTML tags or URLs. For this reason, coupled with the fact that we obtained data from multiple sources using various formats, we built a shared preprocessing pipeline to standardize the input for the fine-tuning of the language models as well as the simplified parts in the aligned dataset. Our pipeline removed redundant tags and characters. Some Easy Language texts use bullet points to break down sentences. Since most of the data did not follow this guideline, we converted the existing bullet points into comma-separated phrases. Another feature of Easy Language is the hyphenation of compound nouns. We compiled a list of hyphenated nouns in the monolingual dataset and used it to replace equivalent non-hyphenated compound nouns.

4 Methodology

Our approach is divided into two parts. First, we fine-tuned generative language models for German Easy Language. Then, we used these models as plug-in decoders in a BART-based simplification task.

4.1 Fine-tuning language models

We selected five different pre-trained GPT-based models from Huggingface (Wolf et al., 2020) as the base for our language models, four German models, and one multilingual model. As shown in Table 2, the models differ in their original training data, initialization, and size. All German models use an embedding size of 1024, while mGPT has a size of 2048. To fine-tune the models, we used a NVIDIA A100 GPU. We trained for one epoch, with a learning rate of $1e^{-4}$, a weight decay of 0.01, and a batch size of eight together with a gradient accumulation of four. However, due to the large model size, we had to decrease the batch size to one for mGPT. The dropout parameters for the embedding, the attention mechanism, and the fully connected layers were set to 0.1 each.

Su et al. (2022) proposed a new learning objective for generative language models, the contrastive loss. This loss adds a similarity regularization to the cross entropy loss to enforce discriminative token representations. We used this loss function together with an AdamW optimizer for our fine-tuning.

[3] https://github.com/ZurichNLP/20Minuten

Model	Training data	Initialization	#Params
GerPT2 (Minixhofer, 2020)	CC-100 Corpus	English GPT2	163M
german-gpt2 (Schweter, 2020)	Wikipedia dump, EU Bookshop corpus, Open Subtitles, CommonCrawl, ParaCrawl and News Crawl	from scratch	124M
GPT2 Wechsel (Minixhofer et al., 2022)	OSCAR corpus, MUSE	English GPT2	124M
Oscar fine-tune (ml6team, 2021)	OSCAR corpus	*no info*	354M
mGPT (Shliazhko et al., 2022) *(multilingual)*	Wikipedia, Colossal Clean Crawled Corpus	from scratch	1417M

Table 2: Training setup and number of parameters for different German GPT2 models. These models were used as base for our Easy Language fine-tuning.

4.2 Text simplification

The simplification task can be considered as a translation-like seq2seq problem. Thus, we used an encoder-decoder architecture based on mBART's architecture (Liu et al., 2020). It consists of a BERT-like encoder and a GPT-like decoder. Additionally, mBART was pre-trained on multilingual data (including German) on a denoising objective and forms the current baseline for transformer-based German ATS (Rios et al., 2021). The baseline's mBART-encoder was modified to use sliding attention to be applied to article inputs. Thus, it was possible to use long input sequences efficiently. We adapted this architecture and replaced the mBART-decoder with our fine-tuned GPT models. For the target text, we used the same preprocessing used for fine-tuning the decoder models. As our language models already output text in the desired style, no further training of the decoder was necessary. Therefore, we only trained the encoder-decoder cross attention to align the encoding of the complex articles with our language models. This was proven successful for machine translation with pre-trained language models by Gheini et al. (2021). Training only the cross attention reduced the number of parameters to be updated, making the training of the simplification more efficient. In addition, the language models were not updated, and thus, we avoided catastrophic forgetting (Goodfellow et al., 2013) of their German language comprehension. We trained with the same hyperparameters as the baseline, except we set label smoothing to zero and added a contrastive part to the loss function (Su et al., 2022). We trained on a single NVIDIA TITAN X. Similar to the baseline, the training converged after 3 to 4 days according to validation loss, which means training for about 20 epochs. Due to hardware limitations, we trained with a batch size of one and a gradient accumulation of 32.

5 Evaluation

This section describes four experiments to compare our fine-tuned (FT) models with their original (O) versions. First, we measured the models' perplexities on easy and normal texts and analyzed the readability of their outputs. In addition, the models were evaluated on two downstream tasks; text complexity prediction and automatic text simplification.

5.1 Perplexity scores

The perplexity describes how likely a specific model will produce a given text. A lower perplexity score indicates a better match between the model and text. We evaluated how well our models adapt to the style of Easy Language. Therefore, the fine-tuned and original models' perplexities on easy and normal texts were compared. The data was collected from the MDR, a public broadcasting service in Germany that publishes news articles in Easy Language. We manually aligned 100 paragraphs from the easy and original articles. To calculate the perplexity of the data, we used the tutorial code from Huggingface (transformers, 2022) that implements perplexity as a sliding window over the input data. We adapted the code for a sample-wise calculation and averaged the perplexity over all samples.

Perplexity is highly dependent on the tokenization and the length of the samples (Wang et al., 2022). Therefore, we cannot determine the best fine-tuned models by selecting the model with the

lowest perplexity. However, the fine-tuned and original versions of the models use the same tokenizers. Thus, we can compare their perplexities and assess the effects of fine-tuning.

Table 3 shows the average perplexity values for the easy and normal texts. No model has seen any of the data before in training. All fine-tuned models show a lower perplexity for the Easy Language samples. In contrast, except for one model, the original models perform better on the normal texts. This suggests that the fine-tuned models match the specificities and structure of Easy Language better and, thus, that they are more likely to produce similar texts.

Model	Easy text		Normal text	
	FT	O	FT	O
gerpt2	**25.35**	51.31	**53.74**	56.42
german_gpt	**31.81**	47.19	77.76	**31.49**
wechsel	**25.99**	38.98	69.29	**34.80**
oscar	**34.24**	59.31	112.75	**66.22**
mGPT	**24.93**	25.05	99.53	**19.18**

Table 3: Comparison of perplexity scores between easy and normal texts. Lower score means better match. The fine-tuned models fit easy German text better, while the original models favor normal texts.

5.2 Readability and Easy Language characteristics

To evaluate the readability of the models' outputs, we compared the Flesch Reading Ease (FRE) scores (Amstad, 1978) of sample outputs. We prompted the models with six different inputs: "Das"(*This*), "Heute"(*Today*), "Wir"(*We*), "Die Türkei"(*Turkey*), "Dieses Haus"(*This house*), and "Mein Vater"(*My father*). The models had to output 100 new tokens, and we set a repetition penalty to enforce novel content in the output. Moreover, three different decoding strategies (contrastive search, sampling, and beam search) were used, resulting in 18 output texts per model. Finally, the FRE score was calculated for each of the model outputs. This score considers the average sentence length and the average number of syllables per word, which favors concise sentences with short words. Therefore, a higher score indicates a more accessible text. Table 4 shows each model's average FRE score. The fine-tuned models achieve a higher score, which implies that their output is more readable than their original's. In addition, we counted the number of suggested newline (\n) tokens. As presented in Table 4, the fine-tuned models output this token more often. This shows that they adapted to the Easy Language characteristic of only writing one thought per line.

Model	Average FRE		\n tokens	
	FT	O	FT	O
gerpt2	**65.17**	51.09	**67**	34
german_gpt	**75.09**	70.89	**79**	74
wechsel	**70.72**	55.86	**69**	18
oscar	**68.21**	49.32	**61**	0
mGPT	**72.16**	55.30	**106**	29

Table 4: Flesch Reading Ease score averaged over different prompts and decoding strategies, and total number of \n tokens suggested. The fine-tuned models output more simple texts.

To further investigate this conformity with Easy Language, we gave the models the input sentence "Heute scheint die Sonne" (*Today sun is shining*) and let them predict the next token. As highlighted in Table 5, most of the fine-tuned models proposed to end the sentence, i.e., predicted a point or a modifier. In contrast, the original models added further information by continuing the sentence with a comma or an "and".

Model	Suggested next token	
	FT	O
gerpt2	.	,
german_gpt	sehr (*very*)	,
wechsel	.	und (*and*)
oscar	.	,
mGPT	auf (*on*)	bei (*at*)

Table 5: Suggested next token for the input sentence "Heute scheint die Sonne" (*Today the sun is shining*). The original models propose to continue the sentence, while the fine-tuned models only put one thought per sentence.

5.3 Human grammar evaluation

Fine-tuning language models to a specific style can result in catastrophic forgetting (Goodfellow et al., 2013). To test if our fine-tuning for Leichte Sprache influences the output quality of the models, we asked human reviewers to rate the models' grammaticality. The reviewers were not paid for their

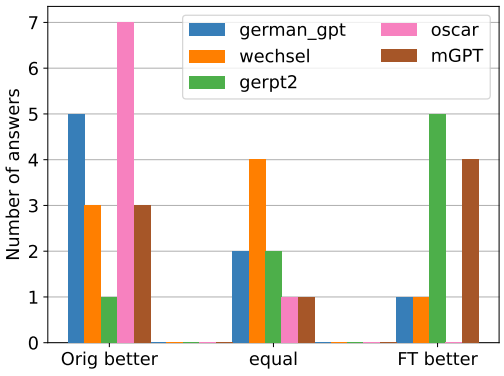

Figure 1: Human grammar evaluation with a ranking task. Participants selected which model output of the fine-tuned and original versions showed fewer grammatical mistakes.

review but participated voluntarily. We selected the outputs of the prompt "Dieses Haus"(*This house*) with decoding strategy contrastive from Section 5.2. Then, we presented the output of each original and its respective fine-tuned model side by side and asked the participants to select the candidate with fewer grammatical errors. Participants could also state that both models were equal. Overall, seven native speakers and one non-native speaker participated in the survey. The distribution of answers is shown in Figure 1. While most participants preferred the fine-tuned version of gerpt2 and mGPT, the fine-tuning of oscar decreased its grammar score. When averaging over all responses and models, the worsening of the grammaticality by fine-tuning the models on Leichte Sprache is neglectable.

5.4 Text complexity prediction

Fine-tuning models for a specific domain improves their performance on different tasks within this domain (Gururangan et al., 2020). To test if this applies to our models, we evaluated them on the downstream task of text complexity prediction. Therefore, we added a linear layer on top of the language model heads and fine-tuned the models for the respective task. The data for this task came from the GermEval 2022 shared task on text complexity assessment (Mohtaj et al., 2022). This shared task's goal was to predict a sentence's complexity on a continuous scale between 1 and 7. We split the shared task's training data into train, evaluation, and test subsets with a ratio of 80:10:10 and fine-tuned our models for ten steps with a batch size of eight, i.e., on 80 samples total. Table 6 reports the mean squared errors on the unseen test set after the few-shot fine-tuning. The first two models have a high error for both the fine-tuned and original models. As the model only performed ten training steps, the results highly depend on the initialization. For the other three models, however, the fine-tuned models clearly outperform the original models. This gives evidence that with the fine-tuning on Easy Language data, the models get a better understanding of text complexity and, thus, can better discriminate easy from normal texts.

Model	Mean squared error	
	FT	O
gerpt2	**2.36**	4.17
german_gpt	6.22	**4.25**
wechsel	**0.81**	1.79
oscar	**0.83**	1.65
mGPT	**0.92**	1.11

Table 6: Mean squared error after fine-tuning for continuous text complexity prediction on 80 sentences. Most of the fine-tuned models outperform their originals.

5.5 Text simplification

We used our pre-trained language models as plug-in decoders in a mBART simplification model. As the decoders already know how to output Easy Language, we only trained the encoder-decoder cross attention. Due to computational limitations, we could not test all our language models on the text simplification downstream task. Therefore, we selected the two most promising ones, gerpt2 and german_gpt. Table 7 shows how our simplification models perform on the 20 Minuten test dataset compared to the baseline by Rios et al. (2021). To generate the simplifications, we used a beam size of four and calculated the metrics with Huggingface evaluate. Our models outperform the baseline on the SARI metric; however, they fall behind when comparing ROUGE-L and BLEU scores. All of these metrics assess how well the proposed output overlaps with a reference simplification and do not consider synonyms. SARI is a score explicitly tailored to the task of simplification, while BLEU and ROUGE-L are general translation/seq2seq metrics. Herefore, a better SARI score may be an indication that our models do more rephrasing than the baseline model and, thus, yield better simplifications. To achieve this result, our models needed training

on only 7% of the trainable parameters of the baseline while preserving state-of-the-art performance.

Score	Baseline*	gerpt2 FT	german_gpt FT
ROUGE-L	**19.96**	18.52	17.93
SARI	33.29	42.25	**42.74**
BLEU	**6.29**	4.95	4.80
#Params trained	416M	**29M**	**29M**

Table 7: Text simplification performance on the 20 Minuten testset. For our models, only the cross attention was trained which reduced the number of trained parameters by far;
*: copied from the baseline paper (Rios et al., 2021).

6 Conclusion

With this paper, we have published a collection of causal language models for German Easy Language. These models mimic the style of Easy Language and favor short and precise sentences. In addition, they adapt to the conventions of only conveying one thought per sentence and putting a line break after every sentence. We exploited these pre-trained models in a sequence-to-sequence text simplification task. As the models were already fine-tuned to the desired output style, we only had to train the encoder-decoder cross attention and, thus, reduced the number of trainable parameters by 93%. With this, training a style-transfer system becomes feasible for settings with few aligned data or a lack of computational power.

Limitations

This paper focuses on the style transfer of Easy Language for German. Due to their word inflections and high average word length, languages like German are harder to learn for language models (Mielke et al., 2019). Therefore, the proposed approach may work even better on easier-to-model languages, but we did not test any other language. In addition, the style transfer of simplified language uses the same vocabulary as the original language and only reduces its diversity. Our approach has yet to be evaluated on other styles, for example, ones that introduce new words.

When evaluating the influence of fine-tunung on the grammaticality of the model outputs, we found that even the original models were not perfect and produced grammatical errors. One possible reason is relying on GPT2-based models that are relatively small and, thus, perform worse than state-of-the-art language models like PaLM (Chowdhery et al., 2022). In addition, the German base models are often already fine-tuned versions of English models, and thus, may already suffer from catastrophic forgetting due to fine-tuning.

Ethics Statement

ATS systems can provide more accessible versions of texts, however, a good text simplification is targeted to the knowledge and language level of its audience. Therefore, to utilize these systems for the target group directly, the systems need to be deployed in a controllable setting where the user can set the level of simplification or ask for additional explanations if necessary. Nevertheless, there are also applications where ATS systems can increase the amount of accessible information on the internt withput being used by the target group directly. For example, these systems can yield a draft simplification for professional translators or can be helpful for public state authorities that are forced by law to offer online information in Easy Language. Another problem is the possible stigmatization of users if they request a simplified version of the data (Hansen-Schirra, 2020). Finally, the availability of information in Easy Language is very sparse; thus, it is hard to fact-check material on the internet with other sources. This makes the target group of Easy Language highly vulnerable to misinformation and fake news. Hence, our generative models must be used with care as they do not provide hallucination control.

Among the sources of our dataset, there is a significant bias towards news articles as well as some regional bias due to the large proportion of articles related to Austria, Switzerland, and northern Germany. As all sources are from official website articles, and the dataset does not include user comments, we expect the data to be unoffensive and of high quality. Nevertheless, we find topical biases such as the COVID-19 pandemic due to the years from which the articles were scraped. In respect of any intellectual property laws, we published the scrapers used to obtain the data but not the data itself.

References

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Ph.D. thesis, Universität Zürich.

Dennis Aumiller and Michael Gertz. 2022. Klexikon: A german dataset for joint summarization and simplification. *arXiv preprint arXiv:2201.07198*.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Mozhdeh Gheini, Xiang Ren, and Jonathan May. 2021. Cross-attention is all you need: Adapting pretrained Transformers for machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1754–1765, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ian J. Goodfellow, Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. 2013. An empirical investigation of catastrophic forgetting in gradient-based neural networks.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Silvia Hansen-Schirra. 2020. Easy language, plain language, easy language plus: perspectives on comprehensibility and stigmatisation. *Easy language research: text and user perspectives*, 2:17.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Changye Li, David Knopman, Weizhe Xu, Trevor Cohen, and Serguei Pakhomov. 2022. GPT-D: Inducing dementia-related linguistic anomalies by deliberate degradation of artificial neural language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1866–1877, Dublin, Ireland. Association for Computational Linguistics.

Yi Liao, Yasheng Wang, Qun Liu, and Xin Jiang. 2019. Gpt-based generation for classical chinese poetry.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8(0):726–742.

Takumi Maruyama and Kazuhide Yamamoto. 2019. Extremely low resource text simplification with pre-trained transformer language model. In *2019 International Conference on Asian Language Processing (IALP)*, pages 53–58.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. Association for Computational Linguistics.

Benjamin Minixhofer. 2020. GerPT2: German large and small versions of GPT2.

Benjamin Minixhofer, Fabian Paischer, and Navid Rekabsaz. 2022. WECHSEL: Effective initialization of subword embeddings for cross-lingual transfer of monolingual language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3992–4006, Seattle, United States. Association for Computational Linguistics.

ml6team. 2021. German finetuned gpt2. `https://huggingface.co/ml6team/gpt2-medium-german-finetune-oscar`.

Salar Mohtaj, Babak Naderi, and Sebastian Möller. 2022. Overview of the GermEval 2022 shared task on text complexity assessment of German text. In *Proceedings of the GermEval 2022 Workshop on Text Complexity Assessment of German Text*, pages 1–9, Potsdam, Germany. Association for Computational Linguistics.

Das Netzwerk Leichte Sprache. 2013. Die regeln für leichte sprache.

Xiangyu Peng, Siyan Li, Spencer Frazier, and Mark Riedl. 2020. Reducing non-normative text generation from language models. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 374–383, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Annette Rios, Nicolas Spring, Tannon Kew, Marek Kostrzewa, Andreas Säuberli, Mathias Müller, and Sarah Ebling. 2021. A new dataset and efficient baselines for document-level text simplification in German. In *Proceedings of the Third Workshop on New Frontiers in Summarization*, pages 152–161, Online and in Dominican Republic. Association for Computational Linguistics.

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.

Elisa Terumi Rubel Schneider, João Vitor Andrioli de Souza, Yohan Bonescki Gumiel, Claudia Moro, and Emerson Cabrera Paraiso. 2021. A gpt-2 language model for biomedical texts in portuguese. In *2021 IEEE 34th International Symposium on Computer-Based Medical Systems (CBMS)*, pages 474–479.

Stefan Schweter. 2020. German gpt-2 model.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2022. mgpt: Few-shot learners go multilingual.

Melanie Siegel, Dorothee Beermann, and Lars Hellan. 2019. Aspects of linguistic complexity: A german – norwegian approach to the creation of resources for easy-to-understand language. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3.

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

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)*.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 230–242.

Yixuan Su, Tian Lan, Yan Wang, Dani Yogatama, Lingpeng Kong, and Nigel Collier. 2022. A contrastive framework for neural text generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 21548–21561. Curran Associates, Inc.

Mari Carmen Suárez-Figueroa, Isam Diab, Edna Ruckhaus, and Isabel Cano. 2022. First steps in the development of a support application for easy-to-read adaptation. *Universal Access in the Information Society*, pages 1–13.

Renliang Sun and Xiaojun Wan. 2022. Simplebert: A pre-trained model that learns to generate simple words.

Julia Suter, Sarah Ebling, and Martin Volk. 2016. Rule-based automatic text simplification for german. In *13th Conference on Natural Language Processing (KONVENS 2016)*. s.n.

Vanessa Toborek, Moritz Busch, Malte Boßert, Christian Bauckhage, and Pascal Welke. 2022. A New Aligned Simple German Corpus. *arXiv preprint arXiv:2209.01106*.

Huggingface transformers. 2022. Perplexity of fixed-length models.

Susanna Värtinen, Perttu Hämäläinen, and Christian Guckelsberger. 2022. Generating role-playing game quests with gpt language models. *IEEE Transactions on Games*, pages 1–12.

Yequan Wang, Jiawen Deng, Aixin Sun, and Xuying Meng. 2022. Perplexity from plm is unreliable for evaluating text quality.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

A Overview of available data for Easy Language

Dataset	Articles	Sentences	Description
Hurraki[4]	3,911	56,785	Wikipedia-style dictionary
Lebenshilfe[5]	396	7,144	Dictionary for people with intellectual disabilities
Einfachstars[6]	6,488	129,674	News about celebrities
Nachrichtenleicht[7]	7,709	122,842	News published by Deutschlandfunk
Kurier[8]	4,519	67,827	News for Austria
NDR[9]	1,817	60,749	News for the states of Lower Saxony, Mecklenburg-Vorpommern, and Schleswig-Holstein
InfoEasy[10]	163	10,310	News for Switzerland
Siegel et al. (2019)	44	4,210	Compilation of election programs, excerpts from the Bible, children's stories, and Red Cross documents

Table 8: Overview of the available monolingual data in Easy Language.

Dataset	Articles	Sentences	Description
Kurier[8]	3,476	-	Article-aligned news data from Austria
BrandEins[11]	212	-	Paragraph-aligned data from a business journal
Wahlprogramm: Die Grünen[12]	-	100	Sentence-wise manually-aligned data from the election program of the Green party
MDR news[13]	-	100	Sentence-wise manually-aligned data from the news for the states of Thuringia, Saxony, and Saxony-Anhalt
MDR dictionary[14]	-	100	Manually-aligned data of dictionary entries between MDR Easy Language entries and German Wikipedia articles
Rios et al. (2021)	18,305	-	Full articles paired with simplified summaries from the Swiss news magazine 20 Minuten
Säuberli et al. (2020)	-	19,724	Sentence-aligned news data from Austria Press Agency aligned using CATS (Štajner et al., 2018)
Toborek et al. (2022)	708	5,942	Both article and sentence-aligned compilation of texts on health and medication, public administration, politics, information texts for disabled people, and news articles (has some overlap with some sources listed in Table 8)
Aumiller and Gertz (2022)	2,898	-	German online encyclopedia for children, called Klexikon (it contains simplified concepts rather than Easy Language)

Table 9: Overview of the parallel data in simplified German and Easy Language.

[4] https://hurraki.de/
[5] https://www.lebenshilfe.de/woerterbuch
[6] https://einfachstars.info/
[7] https://www.nachrichtenleicht.de/
[8] https://kurier.at/einfache-sprache
[9] https://www.ndr.de/fernsehen/barrierefreie_angebote/leichte_sprache
[10] https://infoeasy-news.ch/
[11] https://www.brandeins.de/themen/rubriken/leichte-sprache
[12] https://www.gruene-bw.de/wahlen/landtagswahl-2021/wahlprogramm/wahlprogramm-in-leichter-sprache/
[13] https://www.mdr.de/nachrichten-leicht/index.html
[14] https://www.mdr.de/nachrichten-leicht/woerterbuch/index.html

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Unnumbered Section 7 after the conclusion

☑ A2. Did you discuss any potential risks of your work?
Ethical considerations after conclusion

☑ A3. Do the abstract and introduction summarize the paper's main claims?
First page of paper

☑ A4. Have you used AI writing assistants when working on this paper?
Only Grammarly for language and plagiarism checks on the full paper

B ☑ Did you use or create scientific artifacts?

Sections 4 and 5

☑ B1. Did you cite the creators of artifacts you used?
All sections

☒ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Code is available on Github: https://github.com/MiriUll/Language-Models-German-Simplification, https://github.com/brjezierski/scrapers

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Sections 3-5

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 8 ethical considerations, no steps were taken as data comes from trustworthy public broadcasting services

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 3

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 3-5

C ☑ Did you run computational experiments?

Sections 4,5

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Sections 4,5

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4,5 but we only report the chosen parameters, no explicit search was performed

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 5

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Sections 4,5

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 5.3

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
We report the question asked in section 5.3 but do not report the questionaire in our paper as the text samples are to long. Nevertheless, all sample texts and results are published in our Github repository.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 5.3

☒ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Only used model outputs as data, and thus, no consent needed.

☒ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
The review only focused on grammar, and hence no ethical issues arised

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 5.3 we reported if the annotators were native speakers. Other characteristics were not asked.