Zero-Shot Crosslingual Sentence Simplification

Jonathan Mallinson[1] **Rico Sennrich**[2,1] **Mirella Lapata**[1]

[1]School of Informatics, University of Edinburgh
[2]Department of Computational Linguistics, University of Zurich
J.Mallinson@ed.ac.uk, Sennrich@cl.uzh.ch, mlap@inf.ed.ac.uk

Abstract

Sentence simplification aims to make sentences easier to read and understand. Recent approaches have shown promising results with encoder-decoder models trained on large amounts of parallel data which often only exists in English. We propose a zero-shot modeling framework which transfers simplification knowledge from English to another language (for which no parallel simplification corpus exists) while generalizing across languages and tasks. A shared transformer encoder constructs language-agnostic representations, with a combination of task-specific encoder layers added on top (e.g., for translation and simplification). Empirical results using both human and automatic metrics show that our approach produces better simplifications than unsupervised and pivot-based methods.

1 Introduction

Sentence simplification aims to reduce the linguistic complexity of a text whilst retaining most of its meaning. It has been the subject of several modeling efforts in recent years due to its relevance to various applications (Siddharthan, 2014; Shardlow, 2014). Examples include the development of reading aids for individuals with autism (Evans et al., 2014), aphasia (Carroll et al., 1999), dyslexia (Rello et al., 2013), and population groups with low-literacy skills (Watanabe et al., 2009), such as children and non-native speakers.

Modern approaches (Zhang and Lapata, 2017; Mallinson and Lapata, 2019; Nishihara et al., 2019; Dong et al., 2019) view the simplification task as monolingual text-to-text rewriting and employ the very successful encoder-decoder neural architecture (Bahdanau et al., 2015; Sutskever et al., 2014). In contrast to traditional methods, which target individual aspects of the simplification task, such as sentence splitting (Carroll et al. 1999; Chandrasekar et al. 1996, inter alia) or the substitution of complex words with simpler ones (Devlin, 1999; Kaji et al., 2002), neural models have no special purpose mechanisms for ensuring how to best simplify text. They rely on representation learning to *implicitly* capture simplification rewrites from data, i.e., examples of complex-simple sentence pairs.

While large-scale parallel datasets exist for English (Xu et al., 2015; Zhang and Lapata, 2017) and Spanish (Agrawal and Carpuat, 2019), there is a limited amount of simplification data for other languages. For example, Klaper et al. (2013) automatically aligned 7,000 complex-simple German sentences,[1] and Brunato et al. (2015) released 1,000 complex-simple Italian sentences. But data-driven approaches to simplification, in particular popular neural models, require significantly more training data to achieve good performance, making these datasets better suited for testing or development purposes. Unsupervised approaches (Surya et al., 2019; Artetxe et al., 2018) which forgo the use of parallel corpora are an appealing solution to overcoming the paucity of data. However, in this paper we argue that better simplification models can be obtained by taking advantage of existing complex-simple data in a high-resource language, and bilingual data in a low-resource language (i.e., a language for which no parallel simplification corpus exists).

Drawing inspiration from the success of machine translation (Firat et al., 2016b; Blackwood et al., 2018; Johnson et al., 2017), we propose a modeling framework which transfers simplification knowledge from English to another language while generalizing across language and task barriers during training. The backbone of our model is an encoder-decoder transformer (Vaswani et al., 2017) trained using multi-task learning to either translate, autoencode, simplify, or language model in both high-

[1]This dataset has not been publicly released.

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 5109–5126,
November 16–20, 2020. ©2020 Association for Computational Linguistics

and low-resource languages. Regardless of the task or language, we employ the same base encoder on top of which *task-specific* transformer layers are added, while *language-specific* transformer decoders are used to generate the output sequence. Since the same base encoder is used for all tasks and languages, the model learns task- and language-agnostic representations. A beneficial side-effect is that the proposed architecture can be trained using one language and tasked to simplify another.

As simplifications for multiple languages can be produced within the *same* model, our approach is more scalable compared to pivot-based methods (Mallinson et al., 2018; Conneau et al., 2018). The latter would first translate the complex sentence into a high-resource language, apply a monolingual simplification model, and then translate back the output to the original language. We avoid having to train multiple models and make *multiple* hops, where each hop can add noise and latency, and instead develop a *one-hop* crosslingual zero-shot approach. We evaluate our model using English as our high-resource language and German as our low-resource language on two test sets from different domains, and with different end-users in mind. These include TextComplexityDE (Naderi et al., 2019), a recently created corpus of German Wikipedia sentences deemed complex by second language German learners. We also release a second dataset which contains manual simplifications of articles taken from GEOlino[2], a popular children's magazine. Empirical results using both human and automatic metrics show that our approach produces better simplifications than both unsupervised and pivot-based methods.

Our contributions in this paper are threefold: (1) a cross-lingual architecture which allows the transfer of simplification knowledge from high- to low-resource languages, alleviating the paucity of training data for monolingual simplification; (2) a comprehensive evaluation framework using automatic metrics and human judges; and (3) the release of a dataset in German which we hope will facilitate further research in automatic simplification.[3]

2 Related Work

Simplification The majority of previous work has focused on English, using large-scale datasets like Newsela and Wikipedia (Xu et al., 2015). One of the first neural network approaches to simplification was presented in Zhang and Lapata (2017) who use an encoder-decoder LSTM, trained with reinforcement learning, to optimize for grammaticality, simplicity, and adequacy. Dong et al. (2019) use a Programmer-Interpreter (Reed and de Freitas, 2016), which receives the source sentence as an input, and applies a sequence of edit operations (add, delete, keep). Kriz et al. (2019) propose to rerank a diverse set of simplifications according to fluency, adequacy, and simplicity. Martin et al. (2020a) introduce a simplification model which allows the user to control the generated output and in follow-on work (Martin et al., 2020b) they create multilingual paraphrasing datasets for training their model. Palmero Aprosio et al. (2019) explore different ways to to incorporate non-parallel simplification data to expand small scale training data, including autoencoding and backtranslation.

Translation data, in the form of paraphrases, has also been incorporated into simplification models leading to significant improvements. Guo et al. (2018) use multi-task learning to augment the limited amount of simplification training data. In addition to training on complex-simple sentence pairs, their model employs paraphrases, created automatically using machine translation. Zhao et al. (2018) augment a Transformer-based simplification model with lexical rules obtained from Simple PPDB (Pavlick and Callison-Burch, 2016), a database of paraphrase rules, automatically annotated with simplicity scores.

Unlike previous approaches, we do not train models to create training data, either via backtranslation or extracting paraphrases. Instead, our model is able to train directly on existing datasets, saving computation power and time. In the future, it would be interesting to explore whether additional datasets or tasks improve simplification performance.

Crosslingual Generation Cross-lingual transfer learning-based approaches have originated in machine translation. Dong et al. (2015) translate from one source language to multiple target languages (one-to-many) adding a separate decoder for each. Follow-on work (Luong et al., 2016; Firat et al., 2016a) performs translation with multiple encoders and decoders (many-to-many). Johnson et al. (2017) and Ha et al. (2016) train multilingual models where all languages share encoder and de-

[2] https://www.geo.de/geolino

[3] Our code and dataset can be found at http://www.github.com/Jmallins/ZEST.

coder parameters, and language tags (prepended to the source sentence) are used to specify the target.

Multilingual models are also capable of translating between unpaired languages, thereby performing zero-shot translation (Firat et al., 2016b; Johnson et al., 2017; Ha et al., 2016). Blackwood et al. (2018) propose sharing all parameters but the attention mechanism, while Lu et al. (2018) develop a shared "interlingua layer" at the interface of language-specific encoders and decoders. Advances in unsupervised machine translation (Artetxe et al., 2018; Lample et al., 2018) have further spurred interest in modeling sequence-to-sequence problems without a parallel corpus. Surya et al. (2019) learn from unpaired simple and complex English sentences using a shared encoder, two decoders, denoising, backtranslation and discrimination-based losses. Zhao et al. (2020) propose a similar setup, they create a denoising objective by using simple PPDB, replacing simple phrases with complex phrases. Reinforcement learning is further used to reward the fluency, adequacy and simplicity.

While zero-shot approaches are effective for translating between unpaired languages, they do not consider the case where there exists *no parallel* data for a language. For simplification, we assume that there is no parallel corpus in the low-resource language (e.g., complex-simple German). Furthermore, preliminary results showed that zero-shot translation approaches (Johnson et al., 2017) which prepend a tag in the source sentence — this tag would indicate the simplification task in our case — perform poorly, basically resulting in the source sentence being copied over with no changes made. We circumvent this by replacing tags with task-specific transformer encoder layers which are added on top of the base encoder. This proposed architecture allows us to transfer supervision signals across languages and is potentially useful for other generation tasks, including question generation (Kumar et al., 2019) and sentence compression (Shen et al., 2018; Duan et al., 2019).

3 Zero-shot Simplification

We first define a basic encoder-decoder Transformer before adapting it for zero-shot crosslingual simplification with multi-task learning.

Task	Source Language	Target Language	Target Domain
Translate	HR	HR	complex
Translate	LR	LR	simple
Translate	HR	HR	simple
Translate	LR	LR	complex
Translate	HR	LR	complex
Translate	LR	HR	complex
LM	None	HR	complex
LM	None	HR	simple
LM	None	LR	complex
LM	None	LR	simple
Simplify	HR	HR	simple

Table 1: Training tasks and their instantiations.

3.1 Encoder-Decoder

Given a source sentence $X = (x_1, x_2, ..., x_{|X|})$, our model learns to predict target $Y = (y_1, y_2, ..., y_{|Y|})$, where Y could be a translation (e.g., from English to German) or a simplification (e.g., from complex to simple English). Inferring target Y given source X can be modeled as a sequence-to-sequence learning problem (Bahdanau et al., 2015). Our approach adopts the Transformer's multi-layer and multi-head attention encoder-decoder architecture (Vaswani et al., 2017). The Transformer encoder has n layers (denoted L_i for layer i), which transform the input sequentially, $X^{l+1} = L_i(X^l)$, to yield representations $X^N = L_{1:N}(X)$. For more details regarding the Transformer layer, we refer the reader to Vaswani et al. (2017). The decoder is composed of a stack of identical layers. In addition to self-attention the decoder attends to the source sentence X^N. Encoder and decoder stacks are trained to minimize the cross-entropy loss of Y given X:

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{|Y|} \log p(y_i | y_{<i}, X^N; \theta) \qquad (1)$$

3.2 Multi-task Learning

We define a multi-task crosslingual setup where the model is trained on four basic tasks; namely translation, autoencoding, language modeling, and simplification. We train on different instantiations of these tasks depending on the *source language* which can be high-resource (HR; e.g., English) or low-resource (LR; e.g., German), the *target language* (which is again HR or LR), and the *output domain* which can be simple or complex. We assume we only have monolingual simplification data in the high-resource language and that we have bilingual translation data only in the complex do-

main. Table 1 has a breakdown of the tasks we consider, with a more detailed description below.

Simplification is the backbone of the model and consists of a complex source sentence which must be transformed into a simple sentence, while still retaining the original meaning. We assume we only have parallel training data in the high-resource language (see last row in Table 1).

Translation consists of a source sentence, which must be translated into the target language while retaining the meaning of the source. By training on translation data, our model learns language-agnostic representations which are helpful for simplifying in the low-resource language.

Autoencoding refers to translating between the same language, as seen in Table 1. As it is trivial to autoencode with attention, we apply source token dropout, where randomly selected source tokens are replaced with a special DROP token (Lample et al., 2018). We apply this dropout to all tasks (translation, autoencoding, and simplification). Additionally, this task allows us to incorporate monolingual non-parallel simple data from the low-resource language.

Language Modeling has no source sentence; instead the decoder must learn to predict the next token based on its history. This task also allows us to incorporate monolingual non-parallel simple data from the low-resource language.

Domains in our case our two, the simple domain which consists of text that is easy to read and the complex domain where text has not been explicitly written for ease of reading. Introducing domains to the model allows us to further inject knowledge about monolingual non-parallel simple sentences from the low-resource language. We use the target audience of the data to determine if it is simple or complex (e.g., if the text comes from Simple Wikipedia or a children's book it is representative of simple language). In practice, there often exists only limited amounts of non-parallel simple sentences in the low-resource setting, highlighting the difficulty of the task.

3.3 Crosslingual Training

With the tasks defined, we explain how the model is able to switch among them. We propose a modular encoder, where different encoder layers are used for different tasks; an outline of this can be seen in Figure 1. For every task we use the same k base transformer encoder layers, where k is a hyperparameter. Each task $\mathcal{T}$ (simplification, translation, language modeling), has additional t dedicated transformer layers $L^{\mathcal{T}}_{1:t}$, which are applied to the top of the base k layers, $L^{\mathcal{T}}_{1:t}(L_{1:k}(X))$. Each domain $\mathcal{D}$ (simple/complex), also has d additional dedicated transformer layers $d^{\mathcal{D}}_{1:d}$ applied on top of the task specific layers. The final representation of the source sentence X is:

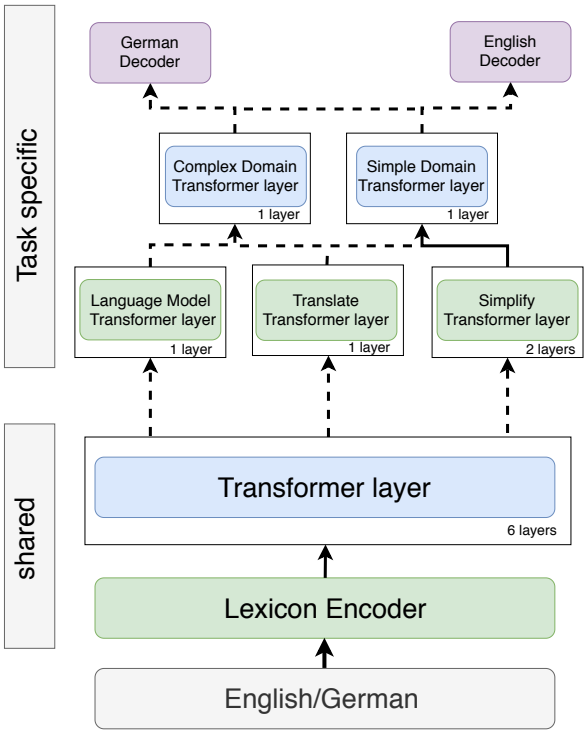

Figure 1: Architecture of our crosslingual encoder-decoder model. The *lexicon Encoder* transforms words into word embeddings. Solid lines indicate mandatory paths, dotted lines indicate possible paths.

$$X^N = L^{\mathcal{D}}_{1:d}(L^{\mathcal{T}}_{1:t}(L_{1:k}(X))) \quad (2)$$

Our model is trained end-to-end to minimize cross entropy; for each minibatch we specify the task, domain and output language ($\mathcal{O}$):

$$\mathcal{L}_{\text{CE}} = -\sum_{i=1}^{|Y|} \log P(y|y_{<i}, X^N; \theta, \{\mathcal{D}, \mathcal{T}, \mathcal{O}\}) \quad (3)$$

$\mathcal{D}$ and $\mathcal{T}$ determine the choice of dedicated Transformer encoder layers. We use a dedicated Transformer decoder for each output language $\mathcal{O}$ to encourage the model to learn language-agnostic representations. All text is preprocessed using SentencePiece (Kudo and Richardson, 2018), resulting

in a shared vocabulary between LR and HR. This allows for word embeddings to be shared between the encoder and the decoders.

We further force representations to be language-agnostic, by employing a DISCriminator (Ganin and Lempitsky, 2015), a feed-forward network trained to distinguish HR and LR from the hidden representations. The encoder is then trained to perplex the discriminator. Specifically, we add two discriminators to our model; one determines the language of the source sentence ($\mathcal{I}$) using $L_{1:k}(X)$, and the other predicts the target language using the output of the encoder X^N. In this way we ensure the input to the simplification transformer layers is language-agnostic as well as the output. The discriminator is trained to minimize the binary cross-entropy loss (BCE) between its predictions and the ground truth:

$$\sum_{i=1}^{|X|} \text{BCE}(\mathcal{I}, \text{DISC}(L_{1:k}(X)_i; \theta_{d\mathcal{I}}) + \text{BCE}(\mathcal{O}, \text{DISC}(X_i^N); \theta_{d\mathcal{O}}) \quad (4)$$

where $\theta_{d\mathcal{I}}$ and $\theta_{d\mathcal{O}}$ are the parameters of the two discriminators. The encoder is trained using an adversarial loss, to perturb the discriminator:

$$\mathcal{L}_{\text{ADV}} = -\sum_{i=1}^{|X|} \text{BCE}(\mathcal{I}, \text{DISC}(L_{1:k}(X)_i; \theta) + \text{BCE}(\mathcal{O}, \text{DISC}(X_i^N); \theta) \quad (5)$$

The adversarial loss is combined, and optimized simultaneously, with the cross-entropy loss to produce the training objective of the entire model.

$$\mathcal{L} = \mathcal{L}_{\text{CE}} + \lambda \mathcal{L}_{\text{ADV}} \quad (6)$$

where λ moderates the degree to which the encoder should perturb the discriminators. A high value for λ can cause the encoder to not encode any information regarding the source input.

To perform simplification in the low-resource language at test time, the base encoder is used with the simplification stack which is subsequently decoded with the LR decoder. To perform crosslingual simplification, the decoder can simply be changed to the HR decoder.

4 Experimental Setup

Training Set Our training data is summarized in Table 2. For all experiments we assume that English is the high-resource language and German is the low-resource language. Simplification data in English is taken from WikiLarge (Zhang and Lapata, 2017), a fairly large corpus which consists of a mixture of three automatically-collated Wikipedia simplification datasets (Zhu et al., 2010; Woodsend and Lapata, 2011; Kauchak, 2013). English-German bilingual data is taken from the WMT19 news translation task. Complex monolingual non-parallel data uses one side of the WMT19 translation data. Simple English non-parallel data uses sentences extracted from simple Wikipedia, a simplified version of Wikipedia. Simple German non-parallel data uses sentences scraped from GEOLino (Hancke et al., 2012), a German general-interest magazine for children aged between 8–14.

	Source	Target	Size
WikiLarge	English$_C$	English$_S$	300K
WMT19	English$_C$	German$_C$	6.0M
GeoLino	—	German$_S$	200K
Wikipedia	—	English$_S$	1.4M

Table 2: Training data used in our experiments; monolingual corpora shown under Target; indices are shorthands for Complex and Simple language.

Test Set We evaluated our model on two German simplification datasets, each targeting different users. TextComplexityDE (Naderi et al., 2019) consists of sentences from Wikipedia, which were considered complex by second language German learners. These sentences were then simplified by a native German speaker. In addition, we created a test set from GEOlino. We extracted 20 articles[4] from three categories: nature, physics, and people. A trained German linguist then simplified the articles, sentence by sentence, to be understandable for children aged between 5–7 years. Our simplifying instructions can be found in the Appendix.

Table 2 shows various descriptive statistics on our test sets. GEOlino is larger and consists of both single and multiple source sentences. The FRE readability metric (see the description in the following section) shows that both the source and target sentence are very simple. We also see moderate amounts of sentence splitting (the number of sentences per instance increases in the simplified target). TextComplexityDE is more complex, with the source sentences having the lowest FRE score. The target simplifications, while noticeably simpler than the source, are still more complex than

[4] Articles were limited to 20 sentences. Half the articles were reserved for a validation set.

	TextComplexityDE		GEOlino	
	Source	Target	Source	Target
Length	28.66	29.23	15.68	15.05
Sents	1.09	2.17	1.13	1.55
FRE	28.53	49.3	62.87	68.73
Size	122		663	
TER	67.95		24.12	
Insertions	3.20		0.43	
Deletions	3.17		1.08	
Substitution	9.10		1.54	
Shifts	1.70		0.18	

Table 3: Descriptive statistics of test set, including: *Size*, number of instances; *Length*, average number of words; *Sents*, average number of sentences per instance; average *Flesch Reading Ease* (*FRE*; higher is simpler); *TER*, translation error rate measuring distance between source and target; it is composed of four parts: insertions, deletions, substitutions and shifts.

GEOlino. We also observe a significant amount of sentence splitting in this dataset. TextComplexityDE also has a significantly higher Translation Error Rate (TER), However, GEOLino approximately matches the TER of the WikiLarge test set (25.85). While both test sets use a large proportion of substitutions, TextComplexityDE has a much large proportion of insertions, which could be explained by the greater amount of sentence splitting.

Model Parameters During training, the base encoder stack consists of six transformer layers, the decoder stack six layers. The simplification stack consists of two weight tied transformer layers, we note that the simplicity level can be increased by applying the stack multiple times at test time. All other stacks consist of a single layer. Each layer has a hidden dimension of size 512 and an inner dimension size of 2,048. Word embeddings, size 512, were initialized randomly and shared between the encoder and both decoders. We used eight attentional heads. Dropout was set to 0.1; source word dropout was also set to 0.1. The discriminator consists of a four layer feedforward network with dropout set to 0.1. The networks were optimized using Adam (Kingma and Ba, 2014). Multi-tasking was performed by alternating batches of different tasks. Tasks varied in dataset sizes and had different difficulties. As we wished to do equally well with all tasks we select a minibatch from a task with a probability inversely proportional to the training loss of the task. One model was selected using the average FRE-BLEU score across both development sets.

All text was preprocessed using the UDPipe tokenization script (Straka, 2018) and truecasing was applied. SentencePiece was subsequently applied to the text to split words into subwords, with a SentencePiece vocabulary size of 50,000 and a sampling size of $l = \infty$ and a smoothing parameter of $\alpha = 0.25$ (Kudo, 2018).

Evaluation As there is no single agreed-upon metric for simplification (Alva-Manchego et al., 2020; Sulem et al., 2018), we evaluate model output using a combination of four automatically-generated scores.[5] These metrics have been previously shown to correlate with human judgments of simplification quality (Xu et al., 2016) and essentially quantify: a) whether the output is similar to the gold standard reference (*Target*-based, *T*); b) whether the output is similar to the source (*Source*-based, *S*); and c) whether the output is simple on its own, with no regard to preserving the meaning of the original sentence (*Readability*-based, *R*). We indicate the type of each metric using superscripts.

BLEU^{T} (Papineni et al., 2002) assesses the degree to which generated simplifications agree with the gold standard references.[6]

I-BLEU^{T,S} (Sun and Zhou, 2012) combines self-BLEU and BLEU to reward systems with high overlap with the reference, and penalize those with high overlap to the source. Self-BLEU computes the BLEU score between the output and the source. It allows us to examine whether the models are making trivial changes to the input. Following Xu et al. (2016), we set the parameter which balances the contribution of the two metrics to $\alpha = 0.9$.

SARI^{T,S} (Xu et al., 2016) is calculated using the average of three rewrite operation scores: addition, copying, and deletion. It rewards addition operations when the system's output is not in the input but occurs in the references. Analogously, it rewards words deleted/retained if they are in both the system output and the references.[7]

FRE-BLEU^{T,S,R} is a modification of FKGL-BLEU (Xu et al., 2016), which combines the difference in FKGL of the source and the output and the I-BLEU score. FKGL is a shorthand for the Flesch-Kincaid Grade Level readability score which was originally developed for English but has *not* been ported to German. So instead we use the Flesch

[5] Our evaluation procedure can be found at `http://www.github.com/Jmallins/ZEST`

[6] We used `multi-bleu-detok.perl` to calculate corpus-level BLEU.

[7] We use corpus level SARI, using precision for deletion rewards and F1 for addition and copying.

Models	FRE-BLEU	I-BLEU	BLEU	SARI
ZEST	**36.04**	**12.99**	**21.11**	**41.12**
Pivot	28.44	8.09	11.50	38.64
U-SIMP	29.95	8.97	15.03	37.40
U-NMT	26.63	7.09	11.72	35.97

(a) TextComplexityDE

Models	FRE-BLEU	I-BLEU	BLEU	SARI
ZEST	62.37	44.72	58.68	39.09
Pivot	39.54	17.81	22.92	27.94
U-SIMP	59.53	**46.33**	**61.10**	**40.00**
U-NMT	**62.57**	39.50	52.02	35.22

(b) GEOlino

Table 4: Results using automatic evaluation metrics; best scores for each metric are **boldfaced**.

	TextComplexityDE		GEOlino	
Model	FRE-BLEU	SARI	FRE-BLEU	SARI
ZEST	36.04	41.11	**62.37**	39.09
−ADV	**36.81**	40.47	60.61	**40.98**
−LM	35.46	41.26	57.29	40.33
−AE	35.56	41.60	57.66	36.49
−LM−AE	35.39	**41.71**	55.37	35.42

Table 5: Ablation study examining the impact of removing the adversarial (ADV) loss, and then additionally removing the language modeling loss (LM), and autoencoding loss (AE), separately then together.

Reading Ease readability test which has been modified for German (FRE; Amstad 1978) and adapt FK-BLEU to use the difference in FRE.[8]

We also evaluated system output by eliciting human judgments via Amazon's Mechanical Turk. Native German speakers (self reported) were asked to rate simplifications on three dimensions: *Grammaticality* (is the output grammatical and fluent?), *Meaning Adequacy* (to what extent is the meaning expressed in the original sentence preserved in the output, with no additional information added?), and *Simplicity* (is the output a simpler version of the input?). Ratings were obtained using a five point Likert scale. We randomly sampled 100 source sentences from each test set (GEOlino and TextComplexityDE), each sample received five ratings, resulting in 500 judgments per test set.

5 Results

Automatic Evaluation Table 4 summarizes our automatic evaluation results. We compare our **ZE**ro-shot cro**S**slingual **S**entence simplifica**T**ion model, which we call ZEST, against multiple baselines, both unsupervised and supervised ones.

Previous work (Artetxe et al., 2018; Lample et al., 2018) demonstrates how an *unsupervised* neural MT model can be trained by optimizing two objectives: (1) *denoising*, where a source sentence is noised and then the corresponding decoder is tasked with reconstructing the original sentence and (2) *on-the-fly back-translation*, which translates the sentence in inference mode; this translation is then encoded and the task is to reconstruct the original sentence. This model can be easily adapted for simplification by considering simple German and complex German to be different languages (U-NMT). Surya et al. (2019) extend this approach further (U-SIMP) by adding two losses, which they show result in better simplifications: (1) an *adversarial loss* using a discriminator which tries to determine if the source sentence is complex or simple, and (2) a *diversification loss*, where a classifier is trained to determine if the source sentence was encoded using the complex or simple encoder. We trained both models using the code provided by Surya et al. (2019) and the same simple and complex *non-parallel* German data used to train our own model (see Table 2; WMT19 complex German and GEOlino simple German).

We additionally include a supervised baseline based on *pivoting*, which requires three independently trained models, consisting of over twice as many parameters: a complex source German sentence is first translated to English (de → en); it is then simplified (complex en → simple en), before translating it back to German (en → de). All three models consist of a transformer with eight encoder/decoder layers and were trained using the same data as employed in our approach (see Table 2; WMT19 and WikiLarge). On the WMT19 test set, the Pivot-based system obtained a BLEU score of 34.15/31.72 for the en → de/de → en directions. For comparison, ZEST achieved 32.11/30.90 for the same directions. With regard to English simplification (complex en → simple en), the pivot system achieved a SARI score of 36.30 on the WikiLarge test, and ZEST 37.78. On the same test set, Zhang and Lapata (2017), a standard baseline simplification system trained on WikiLarge obtains 37.26, and the state-of-the-art system achieves 41.70 (Martin et al., 2020a). It is possible to incorporate some of the improvements of these approaches (e.g., controlling the amount of compression, paraphrasing, lexical complexity) into our model, however, we leave this to future work.

[8] Calculated as FRE $= 180 - \text{ASL} - (58.5 \cdot \text{ASW})$ where ASL is the average sentence length and ASW the average number of syllables per word.

C	Das ist nur etwa das Doppelte **[des Weltenergiebedarfs]**[4], **[was]**[5] bedeutet, **[dass]**[5] **[Erdwärmenutzung]**[6] **[im]**[2] großen Stil immer auf eine lokale Abkühlung des Gesteins hinausläuft.
R	Das ist nur etwa das Doppelte **[des Energiebedarfs der Welt]**[4]. Das bedeutet, **[dass]**[5] die **[Benutzung]**[6] von Erdwärme immer dazu führt, **[dass]**[5] an **[sich]**[2] diesen Stellen das Gestein abkühlt.
P	Dabei handelt es sich nur um eine **[Verdoppelung]**[6] **[des weltweiten Energiebedarfs]**[5], **[was]**[5] **[bedeutet]**[2], **[dass]**[5] die großflächige **[geothermische]**[7] **[Nutzung]**[6] immer einer lokalen **[Kühlung]**[6] **[des Gesteins]**[4] entspricht.
Z	Das bedeutet, **[dass]**[5] Erdwärme im großen Stil immer auf eine lokale Abkühlung **[des]**[2] Gesteins hinausläuft.

(a) TextComplexityDE

C	Von hier aus erhaltet ihr einen **[eindrucksvollen]**[1] Rundum-Blick über die ganze Schlucht **[hinweg]**[2] bis hin zu ihren etwa **[5000]**[3] Meter hohen Kraterwänden.
R	Von hier aus erhaltet ihr einen Rundum-Blick über die ganze Schlucht. Ihr seht hier bis hin zu ihren etwa **[5000]**[3] Meter hohen Kraterwänden.
P	Von hier genießen Sie einen **[beeindruckenden]**[1] Rundumblick über die gesamte Schlucht bis **[zu]**[2] den 500 m hohen Kraterwänden.
Z	Von hier aus erhaltet ihr einen Rundum-Blick auf die ganze Schlucht.

(b) GEOlino

Table 6: Examples of system output, Source (**C**), Reference (**R**), Pivot (**P**), ZEST (**Z**) and simplification violations: (1) word has 13+ letters; (2) sentence has 12+ words; (3) high number; (4) genitive; (5) subordinate clauses; (6) abstract words; (7) difficult/foreign words.

The results in Table 4 show that ZEST obtains the highest results for all metrics on TextComplexityDE. U-SIMP achieves the second best FRE-BLEU score, while Pivot achieves the second best SARI. Overall, U-NMT produces the worst results. Results on GEOlino are more mixed, with no model achieving the highest score across all metrics. ZEST does well across all metrics, scoring the second highest for every metric, whereas the scores for U-SIMP and U-NMT spike on different metrics. U-NMT achieves the best FRE-BLEU score, however, on other metrics it is the second lowest. In contrast, U-SIMP has a low FRE-BLEU score but for all other metrics it scores the highest. Pivot receives the lowest scores across all metrics. Example output is shown in Table 10 and the Appendix.

We further examined the impact different loss functions have on the performance of ZEST, and these results are presented in Table 5. We see that training only on simplification and translation data (−LM−AE) significantly damages the performance of the model, producing the lowest FRE-BLEU scores and the lowest SARI score on

Models	Mean	Gram	Simp	AVG	Min
Reference	4.35^{**}	4.54^{**}	3.81^{*}	4.23^{**}	3.60^{**}
U-SIMP	2.67^{**}	2.87^{**}	2.80^{**}	2.78^{**}	2.22^{**}
Pivot	3.65^{**}	4.13	**3.67**	3.82^{*}	3.18
ZEST	**4.05**	**4.15**	3.63	**3.94**	**3.23**

(a) TextComplexityDE

Models	Mean	Gram	Simp	AVG	Min
Reference	4.73^{**}	4.75^{**}	3.79^{**}	4.42^{**}	3.69^{**}
U-SIMP	4.19^{*}	4.30^{**}	3.22^{*}	3.90^{*}	3.08^{**}
Pivot	3.69^{**}	3.76^{**}	3.25^{*}	3.45^{**}	2.83^{**}
ZEST	**4.38**	**4.57**	**3.44**	**4.13**	**3.24**

(b) GEOlino

Table 7: Mean ratings given to simplifications by human participants; highest ratings for each system are **boldfaced**. Models significantly different from ZEST are marked with $^{*}(p < 0.05)$ and $^{**}(p < 0.01)$. Significance tests were performed using a student *t*-test.

GEOlino. We note that by removing both the language modelling loss and autoencoding loss we are removing the non-parallel simple German data (GEOlino), which could explain the performance drop on the GEOLino test set. While in the full model ZEST has access to GEOlino data, the GEOLino test set is simpler than the GEOLino non-parallel training set, as it was further simplified. Additionally, the ability to incorporate extra data is a strength of our approach, as there is no obvious way to include it within the Pivot-based model.

We observed that removing the autoencoding loss (−AE) led to sentences which strayed too far from the source sentence, thereby losing meaning; whereas removing the language modeling loss (−LM) led to sentences being too close to the source sentence, resulting in too little simplification. The inclusion of the adversarial loss (−ADV) showed a small overall increase in FRE-BLEU and a small decrease in SARI.

Human Evaluation Table 7 summarizes the results of the human evaluation. We elicited judgments for three systems, namely ZEST, U-SIMP, and the Pivot-based approach. We also included the gold standard Reference as an upper bound (see the Appendix for examples of sentence pairs shown to crowdworkers). We report mean ratings for Meaning adequacy, Grammaticality and Simplicity, their combined average (AVG), and their (average) Minimum value. We include Minimum because we argue that a simplification is only as good as its weakest dimension. We note that it is trivial to produce a sentence that is perfectly adequate and fluent, by simply repeating the source

Model	lex	SC	RC	pas	suj	gen	spl
Reference	38.7	11.9	10.5	6.8	16.2	12.2	35.5
U-SIMP	41.2	**18.7**	4.3	4.6	7.7	8.0	**3.2**
Pivot	44.9	17.3	7.8	**6.7**	11.6	**14.6**	**3.2**
ZEST	**51.9**	8.3	**11.8**	4.9	**13.0**	5.8	2.3

Table 8: Proportion of simplifications on 100 sentences including lexical (lex), subordinate clause (SC), relative clause (RC), passive voice (pas) subjunctive (suj), genitive (gen), and sentence splitting (spl).

sentence. It is also easy to produce a simple grammatical sentence if we do not care about adequacy.

On TextComplexityDE, ZEST is significantly better than the unsupervised approach across all dimensions. It is on par with Pivot in terms of Grammaticality, Simplicity, and Minimum (ratings are not significantly different). However, ZEST is significantly better in terms of Meaning adequacy, and on average. On GEOlino, ZEST is significantly better against all comparison models on all dimensions. Perhaps unsurprisingly, across datasets, participants perceive gold standard simplifications as superior to the output of all comparison models.

Error analysis We further analysed the types of simplifications produced by each system. We sampled 100 source sentences (50 from each dataset) and elicited judgments from annotators. The annotators were asked to indicate the types of simplification which occurred, including: lexical substitutions, passive to active voice, splitting a sentence into multiple sentences, and rewriting it to avoid subordinate clauses, relative clauses, the subjunctive mood, and the genitive case. The results in Table 8 show that ZEST performs a wide variety of simplification and produces the largest number of lexical simplifications. While all models produce more lexical substitutions than the references, the references split sentences frequently, whereas in all cases, the models split the sentence minimally. The Pivot model simplifies genitives the most while U-SIMP simplifies subordinate clauses most. ZEST produces the largest number of lexical simplifications, and simplifications related to relative clauses and subjunctives.

Crosslingual Simplification We next explore how different tasks can be combined with no additional training data. We illustrate how our model can be used to tackle the tasks of both simplifying *and* translating. We now assume that the source complex sentence is in English and the simplified output sentence is in German. As there currently exist no crosslingual German simplification test sets, for evaluation purposes we hand-translated 100 complex sentences from each of the German test sets into English. Results[9] can be seen in Table 9 and example output in the Appendix. For comparison, we provide the results of Pivot, which requires two independently-trained models: a complex source English sentence is first simplified (complex en → simple en), and then translated into German (en → de). While the results show that ZEST and Pivot are comparable, the fact that we can train our model on single tasks and then recombine task-specific layers to allow zero-shot transfer to unseen task combinations opens up exciting new opportunities for future work.

Models	FRE-BLEU	I-BLEU	BLEU	SARI
ZEST	31.82	10.26	14.29	41.11
Pivot	32.72	10.71	15.19	41.60

(a) TextComplexityDE

Models	FRE-BLEU	I-BLEU	BLEU	SARI
ZEST	43.65	19.17	25.00	34.62
Pivot	42.61	18.29	23.78	34.43

(b) GEOlino

Table 9: Crosslingual, simplifying English into German, automatic results.

6 Conclusions

In this paper we developed a general approach for transferring generation data from high- to low-resource languages. Experimental results on transferring simplification knowledge from English to German showed that our model was able to produce significantly better German simplifications than unsupervised and pivot-based approaches. In addition to zero-shot simplification, we showed that our model can generate German simplifications given English input, without any additional training. In the future, we plan to explore this approach with other language pairs and other generation tasks.

Acknowledgments The authors gratefully acknowledge the support of the European Research Council (award number 681760; Lapata) and the Swiss National Science Foundation (MUTAMUR, no. 176727; Sennrich). We thank all those who helped with German: Sabine Webber, Ivana Balažević, Denis Emelin, Frank Keller, and Steven Kleinegesse.

[9]Both SARI and FRE-BLEU are monolingual evaluation metrics, as such we use the original German source sentence.

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. 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 1549–1564, Hong Kong, China.

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

Toni Amstad. 1978. *Wie verständlich sind unsere Zeitungen?* Studenten-Schreib-Service.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018. Unsupervised neural machine translation. In *Proceedings of the 6th International Conference on Learning Representations, ICLR 2018*, Vancouver, Canada.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, California.

Graeme Blackwood, Miguel Ballesteros, and Todd Ward. 2018. Multilingual neural machine translation with task-specific attention. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3112–3122, Santa Fe, New Mexico.

Dominique Brunato, Felice Dell'Orletta, Giulia Venturi, and Simonetta Montemagni. 2015. Design and annotation of the first Italian corpus for text simplification. In *Proceedings of The 9th Linguistic Annotation Workshop*, pages 31–41, Denver, Colorado.

John Carroll, Guido Minnen, Darren Pearce, Yvonne Canning, Siobhan Devlin, and John Tait. 1999. Simplifying text for language-impaired readers. In *Proceedings of the 9th Conference of the European Chapter of the Association for Computational Linguistics*, pages 269–270, Bergen, Norway.

R. Chandrasekar, Christine Doran, and B. Srinivas. 1996. Motivations and methods for text simplification. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium.

Siobhan Devlin. 1999. *Simplifying Natural Language for Aphasic Readers*. Ph.D. thesis, University of Sunderland.

Daxiang Dong, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang. 2015. Multi-task learning for multiple language translation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1723–1732, Beijing, China.

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.

Xiangyu Duan, Mingming Yin, Min Zhang, Boxing Chen, and Weihua Luo. 2019. Zero-shot cross-lingual abstractive sentence summarization through teaching generation and attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3162–3172, Florence, Italy.

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

Orhan Firat, Kyunghyun Cho, and Yoshua Bengio. 2016a. Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California.

Orhan Firat, Baskaran Sankaran, Yaser Al-onaizan, Fatos T. Yarman Vural, and Kyunghyun Cho. 2016b. Zero-resource translation with multi-lingual neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 268–277, Austin, Texas.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1180–1189.

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

Thanh-Le Ha, Jan Niehues, and Alexander H. Waibel. 2016. Toward multilingual neural machine translation with universal encoder and decoder. *CoRR*, abs/1611.04798.

Julia Hancke, Sowmya Vajjala, and Detmar Meurers. 2012. Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012*, pages 1063–1080, Mumbai, India. The COLING 2012 Organizing Committee.

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.

Nobuhiro Kaji, Daisuke Kawahara, Sadao Kurohashi, and Satoshi Sato. 2002. Verb paraphrase based on case frame alignment. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 215–222, Philadelphia, Pennsylvania.

David Kauchak. 2013. Improving text simplification language modeling using unsimplified text data. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1537–1546, Sofia, Bulgaria.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

David Klaper, Sarah Ebling, and Martin Volk. 2013. Building a German/simple German parallel corpus for automatic text simplification. In *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, pages 11–19, Sofia, Bulgaria.

Reno Kriz, João Sedoc, Marianna Apidianaki, Carolina Zheng, Gaurav Kumar, Eleni Miltsakaki, and Chris Callison-Burch. 2019. Complexity-weighted loss and diverse reranking for sentence simplification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3137–3147, Minneapolis, Minnesota.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium.

Vishwajeet Kumar, Nitish Joshi, Arijit Mukherjee, Ganesh Ramakrishnan, and Preethi Jyothi. 2019. Cross-lingual training for automatic question generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4863–4872, Florence, Italy.

Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Unsupervised machine translation using monolingual corpora only. In *6th International Conference on Learning Representations, ICLR 2018,*, Vancouver, Canada.

Yichao Lu, Phillip Keung, Faisal Ladhak, Vikas Bhardwaj, Shaonan Zhang, and Jason Sun. 2018. A neural interlingua for multilingual machine translation. In *Proceedings of the 3rd Conference on Machine Translation: Research Papers*, pages 84–92, Brussels, Belgium.

Minh-Thang Luong, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser. 2016. Multi-task sequence to sequence learning. In *Proceedings of the 4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico.

Jonathan Mallinson and Mirella Lapata. 2019. Controllable sentence simplification: Employing syntactic and lexical constraints. *CoRR*, abs/1910.04387.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2018. Sentence compression for arbitrary languages via multilingual pivoting. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2453–2464, Brussels, Belgium.

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

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2020b. Multilingual unsupervised sentence simplification. *arXiv preprint arXiv:2005.00352*.

Babak Naderi, Salar Mohtaj, Kaspar Ensikat, and Sebastian Möller. 2019. Subjective assessment of text complexity: A dataset for german language. *arXiv preprint arXiv:1904.07733*.

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.

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*, pages 311–318, Philadelphia, Pennsylvania.

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

Scott E. Reed and Nando de Freitas. 2016. Neural programmer-interpreters. In *4th International Conference on Learning Representations, ICLR 2016*, San Juan, Puerto Rico.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013. Simplify or help?: text simplification strategies for people with dyslexia. In *International Cross-Disciplinary Conference on Web Accessibility, W4A '13*, pages 15:1–15:10, Rio de Janeiro, Brazil.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications*, 4(1):58–70.

Shiqi Shen, Yun Chen, Cheng Yang, Zhiyuan Liu, and Maosong Sun. 2018. Zero-shot cross-lingual neural headline generation. *IEEE/ACM Trans. Audio, Speech & Language Processing*, 26(12):2319–2327.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 685–696.

Hong Sun and Ming Zhou. 2012. Joint learning of a dual SMT system for paraphrase generation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–42, Jeju Island, Korea.

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.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 3104–3112, Montréal, Canada.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017*, pages 5998–6008, Long Beach, California.

Willian Massami Watanabe, Arnaldo Candido Junior, Vinícius Rodriguez Uzêda, Renata Pontin de Mattos Fortes, Thiago Alexandre Salgueiro Pardo, and Sandra Maria Aluísio. 2009. Facilita: reading assistance for low-literacy readers. In *Proceedings of the 27th ACM International Conference on Design of Communication*, pages 29–36, New York, New York.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK.

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.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark.

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

Yanbin Zhao, Lu Chen, Zhi Chen, and Kai Yu. 2020. Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, pages 9668–9675, New York, New York.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

A Simplification Instructions

This annotation experiment is concerned with simplification. You will be presented with a document. Your task is to read each sentence and simplify it such that children aged between 5 and 7 can understand it. The simplified version should be grammatical and retain all the important information of the original sentence.

In producing simplifications, you are free to delete words, add new words, substitute them, or reorder them. In addition, you might find it useful to change a complex sentence into multiple simple sentences.

To help you with the simplification task, we have produced a set of guidelines which you can follow. However, not all guidelines will always be applicable, so if you believe you can produce a simpler version,n then you may ignore the guidelines. We split the guidelines into two sections: word-level and sentence-level guidelines.

A.1 Word-level Guidelines

1. Special characters are not allowed, with the exception of: full stops, question marks, exclamation marks, quotation marks, and Mediopunkts (used to indicate compound splitting).

2. Numbers should be written as digits and not words.

3. The word *ein* ('one') should only be written with a 1 when it represents a number, not when it takes the role of an indefinite article.

4. Roman numerals must be avoided.

5. Large numbers, percentages and year dates should be used sparsely.

6. Use easy, short and well-known words. In case a difficult word is needed, it should be explained using simple words. For a list of simple words, please consult this dictionary: https://hurraki.de/wiki/Hauptseite.

7. Technical terms, foreign words and abbreviations should be avoided. Common acronyms like CD or *WC* may be used if their full forms (compact disc, water closet) are less common.

A.2 Sentence-level Guidelines

1. Coordinate and subordinate clauses are forbidden and should be transformed into independent main clauses. Main clauses should preferably contain active voice, and present, or past perfect tense. The subject-verb-object (SVO) word order should be chosen, unless another word order is more understandable.

2. Nominalizations and passive constructions are forbidden.

3. Attributive genitives should also be avoided. If possible, the genitive attribute should be transferred into a prepositional phrase using *von* ('of').

4. Negation should be avoided. If needed, it is better to formulate a sentence with *nicht* ('not') instead of *kein* ('no').

5. Transparent metaphors like leichte Sprache may be used if they can be easily understood. More complex metaphors and idioms should be replaced by literal expressions.

6. Split complex sentences into multiple simple sentences at semicolons and dashes. Also split sentences after colons if the segment after the colon is a complete sentence and not just an enumeration.

7. If a subordinate conjunction is found, split the sentence at the conjunction; edit and rephrase both resulting segments to form independent sentences. Add suitable connectives that express the intended rhetorical relation and restore word order.

8. Rephrase concessive clauses with subjunctions like *obwohl* ('although') the connective *trotzdem* ('however').

9. Analogously, rephrase consecutive clauses starting with *sodass* ('so that') using *deshalb* ('therefore').

10. Rephrase final clauses using the modal verb *wollen* ('want') and the connective *deshalb* ('therefore'). Since the subject is not mentioned overtly in German final clauses containing *um zu* ('in order to'), it has to be retrieved from the main clause.

11. Split coordinate clauses at coordinating conjunctions (e.g., *und* ('and'), *oder* ('or'), *aber* ('but'), *dennoch* ('however')). The second clause can start with *und* ('and') and *oder* ('or') to emphasize that they are linked to the previous sentence.

12. Replace appositions by sentences in which the noun phrase referred to by the apposition forms the subject (X) and the apposition itself becomes the predicative noun (Y), yielding an X is Y structure.

A.3 Final Remarks

The annotation will proceed on a document-by-document basis. In simplifying individual sentences you should ensure that:

- You have preserved all important information in the original sentence.
- The sentences are understandable to children aged 5 to 7.
- You did not render the resulting document incoherent or unreadable.
- You have preserved the grammaticality of the simplified sentences.

B System Output

In Table 10 we present examples of simplifications from both GEOlino and TextComplexityDE. We show the input Complex sentence, the Reference simplification, and the output of our model, ZEST, and two comparison systems U-SIMP and Pivot (see Section 5 in the main paper for more details). To provide some further insight on what the models are doing we have annotated words and phrases in the examples which constitute violations of simple language according to our guidelines above and those provided in https://hurraki.de/pruefung/pruefung.htm.

Table 11 contains additional examples, without annotation, while Table 12 presents crosslingual simplicity examples. Specifically, we show model output in German (DE ZEST) when the input is complex English (EN Source), and for comparison German output (DE ZEST) when the input is complex German (DE Source).

Complex	Von hier aus erhaltet ihr einen **[eindrucksvollen]**[1] Rundum-Blick über die ganze Schlucht **[hinweg]**[2] bis hin zu ihren etwa **[5000]**[3] Meter hohen Kraterwänden.
Reference	Von hier aus erhaltet ihr einen Rundum-Blick über die ganze Schlucht. Ihr seht hier bis hin zu ihren etwa **[5000]**[3] Meter hohen Kraterwänden.
U-SIMP	Von hier eraus ihr haltet einen umfassenden Rundum-Blick über die ganze bis **[hinweg]**[2] hinweg zu hin zu ihren **[5000]**[3] Meter hohen Kraterwände.
Pivot	Von hier genießen Sie einen **[beeindruckenden]**[1] Rundumblick über die gesamte Schlucht bis **[zu]**[2] den 500 m hohen Kraterwänden.
ZEST	Von hier aus erhaltet ihr einen Rundum-Blick auf die ganze Schlucht.

(a) GEOlino

Complex	Das ist nur etwa das Doppelte **[des Weltenergiebedarfs]**[4], **[was]**[5] bedeutet, **[dass]**[5] **[Erdwärmenutzung]**[6] **[im]**[2] großen Stil immer auf eine lokale Abkühlung des Gesteins hinausläuft.
Reference	Das ist nur etwa das Doppelte **[des Energiebedarfs der Welt]**[4]. Das bedeutet, **[dass]**[5] die **[Benutzung]**[6] von Erdwärme immer dazu führt, **[dass]**[5] an **[sich]**[2] diesen Stellen das Gestein abkühlt.
U-SIMP	Das ist nur etwa das Doppelte **[des Weltenergiebedarfs]**[4], **[was]**[5] bedeutet, **[dass]**[5] Erdwärmemer **[im]**[2] großen Stil immer auf eine andere Abkühlung des Gesteins**[]**[7].
Pivot	Dabei handelt es sich nur um eine **[Verdoppelung]**[6] **[des weltweiten Energiebedarfs]**[5], **[was]**[5] **[bedeutet]**[2], **[dass]**[5] die großflächige **[geothermische]**[8] **[Nutzung]**[6] immer einer lokalen **[Kühlung]**[6] **[des Gesteins]**[4] entspricht.
ZEST	Das bedeutet, **[dass]**[5] Erdwärme im großen Stil immer auf eine lokale Abkühlung **[des]**[2] Gesteins hinausläuft.

(b) TextComplexityDE

Table 10: Examples of system output and simplification violations (https://hurraki.de/pruefung/pruefung.htm): (1) word has 13+ letters; (2) sentence has 12+ words; (3) high number; (4) genitive; (5) subordinate clauses; (6) abstract words; (7) special characters; (8) difficult/foreign words.

Source	Zusätzlich kommt in der Forschung ein breites Spektrum verschiedener Isotopenanalysen zum Einsatz, deren jüngste Entwicklungen eine bis vor kurzem unerreichbare Messgenauigkeit ermöglichen.
Reference	Außerdem werden in der Forschung viele verschiedene Isotopenanalysen eingesetzt. Ihre jüngste Entwicklung ermöglichen ermöglichen eine bis vor kurzem unerreichbare Messgenauigkeit.
U-SIMP	Zusätzlich kommt in der Forschung ein großes verschiedene verschiedenen Gitopenempfindliche zum Einsatz, deren jüngsten Jahren eine bis vor kurzem brennende Messoberfläche ermöglichen.
Pivot	Darüber hinaus wird in der Forschung ein breites Spektrum unterschiedlicher Isotopanalysen eingesetzt.
ZEST	Zusätzlich kommt in der Forschung ein breites Spektrum verschiedenster Isotopen-Analysen zu Einsatz.
Source	Das ist nur etwa das Doppelte des Weltenergiebedarfs, was bedeutet, dass Erdwärmenutzung im großen Stil immer auf eine lokale Abkühlung des Gesteins hinausläuft.
Reference	Das ist nur etwa das Doppelte des Energiebedarfs der Welt. Das bedeutet, dass die Benutzung von Erdwärme immer dazu führt, dass an sich diesen Stellen das Gestein abkühlt.
U-SIMP	Das ist nur etwa das Doppelte des Weltenergiebedarfs, was bedeutet, dass Erdwärmemer im großen Stil immer auf eine andere Abkühlung des Gesteins).
Pivot	Dabei handelt es sich nur um eine Verdoppelung des weltweiten Energiebedarfs, was bedeutet, dass die großflächige geothermische Nutzung immer einer lokalen Kühlung des Gesteins entspricht.
ZEST	Das bedeutet, dass Erdwärme im großen Stil immer auf eine lokale Abkühlung des Gesteins hinausläuft.

(a) TextComplexityDE

Source	Tiere tauschen mittels Duftmarken viele verschlüsselte Botschaften untereinander aus.
Reference	Tiere tauschen mit ihrem Geruch viele Botschaften untereinander aus.
U-SIMP	Tiere tauschen Hilfe Duftmarken viele verschlüsselte Botschaften untereinander aus.
Pivot	Tiere tauschen viele verschlüsselte Nachrichten mit Duftmarken aus.
ZEST	Tiere tauschen mit Duftmarken viele verschlüsselte Botschaften aus.
Source	Der wiederum war überlebenswichtig für alle Landwirtschaft betreibenden Kulturen.
Reference	Der war wichtig für alle Kulturen, die Landwirtschaft betreiben.
U-SIMP	Der wiederum war überlebenswichtig für alle Landwirtschaft ben Kulturen.
Pivot	Sie war wiederum lebenswichtig für alle landwirtschaftlichen Kulturen.
ZEST	Der wiederum war für alle Landwirtschaft wichtig.

(b) GEOlino

Table 11: Simplification examples from TextComplexityDE and GEOlino.

EN Source	The mountain is the watershed on whose flanks the catchment areas of the Pacific Ocean, the Atlantic Ocean over the Gulf of Mexico, and the Arctic Ocean over Hudson Bay, meet.
DE Source	Der Berg ist der Wasserscheidepunkt an dessen Flanken sich die Einzugsgebiete des Pazifischen Ozeans, des Atlantischen Ozeans über den Golf von Mexiko und des Arktischen Ozeans über die Hudson Bay berühren.
Reference	Der Berg markiert die Grenze zwischen den Gebieten des Pazifischen Ozeans, des Atlantischen Ozean und des Arktischen Ozeans.
EN ZEST	Der Berg ist der Weckschatz, auf dessen Flanken die Fanggebiete des pazifischen Ozeans, des Atlantischen Ozeans über dem Golf von Mexiko, und des Arktischen Ozeans über Hudson Bay, treffen.
DE ZEST	Der Berg ist der Wasserscheidepunkt an dem sich die Einzugsgebiete des Pazifiks, des Atlantischen Ozeans, des Golfs von Mexiko und des Arktischen Ozeans über die Hudson Bay treffen.

(a) TextComplexityDE

EN Source	Without the radiation energy of the sun, plant photosynthesis would not work.
DE Source	Ohne die Strahlungsenergie der Sonne würde die pflanzliche Photosynthese nicht funktionieren.
Reference	Ohne die Energie der Sonne würde die Photosynthese von den Pflanzen nicht funktionieren.
EN ZEST	Ohne die Strahlungsenergie der Sonne, Pflanzen Photosynthese würde nicht funktionieren.
DE ZEST	Ohne die Strahlungsenergie der Sonne würde die Pflanze nicht funktionieren.

(b) GEOlino

Table 12: Examples of crosslingual simplification (EN Source → DE ZEST); for comparison, we also show the output of a monolingual system (DE Source → DE ZEST).

Parameter	Values
No. Base layers (k)	[4, **6**, 8]
No. Domain layer	[**1**, 2]
No. Task layers (t)	[**1**, 2]
ADV loss (λ)	[0, **1**, 5]
No. Discriminator layers	[2, **4**]
Word Dropout	[0, **10%**]
No. Decoder layers	[**8**]
No. Decoders	[1, **2**]
Batch size	**4000**

Table 13: Hyperparmeter bounds. Bold indicates final value.

C Reproducibility

We include additional details for reproducibility in this section.

Average Runtime for Each Approach Run time results were calculated using a batch size of 30 on a Nvidia Tesla K40. Inference speed on 100 sentences was 34s for ZEST and 60s for the Pivot model (time includes loading the models).

Hyperparameter Configurations and Bounds See Table 13 and section 4 of the main paper for more details. If not mentioned then we used the recommendation from OpenNMT-py https://opennmt.net/OpenNMT-py/FAQ.html#how-do-i-use-the-transformer-model. Hyperparameter bounds are also shown in Table 13 with selection done using the validation set and FRE-BLEU.

Explanation of Data Preprocessing See section 4 of the main paper for more details. In addition, training data was excluded if it exceeded 80 tokens. Scrapped training data (GEOlino / simple wikipedia) was excluded if it began with a special character, was less than 5 words long, or did not end in punctuation.

Links to Downloadable Version of the Data

- *Simplification*: We followed instructions from https://github.com/XingxingZhang/dress.
- *Translation*: http://www.statmt.org/wmt19/translation-task.html
- *TextComplexityDE*: https://github.com/Jmallins/TextComplexityDE
- *GEOlino test set*: https://github.com/Jmallins/ZEST
- *GEOlino training set*: Contact authors (Hancke et al., 2012). Scrapping scripts can be found https://github.com/Jmallins/ZEST.
- *Simple Wikipedia*: https://dumps.wikimedia.org/simplewiki/latest/