(Psycho-)Linguistic Features Meet Transformer Models for Improved Explainable and Controllable Text Simplification

Yu Qiao[1], **Xiaofei Li**[1], **Daniel Wiechmann**[2], **Elma Kerz**[1]
[1] RWTH Aachen University
[2] University of Amsterdam
{yu.qiao, xiaofei.li1}@rwth-aachen.de
d.wiechmann@uva.nl, elma.kerz@ifaar.rwth-aachen.de

Abstract

State-of-the-art text simplification (TS) systems adopt end-to-end neural network models to directly generate the simplified version of the input text, and usually function as a black-box. Moreover, TS is usually treated as an all-purpose generic task under the assumption of homogeneity, where the same simplification is suitable for all. In recent years, however, there has been increasing recognition of the need to adapt the simplification techniques to the specific needs of different target groups. In this work, we aim to advance current research on explainable and controllable TS in two ways: First, building on recently proposed work to increase the transparency of TS systems (Garbacea et al., 2021), we use a large set of (psycho-)linguistic features in combination with pre-trained language models to improve explainable complexity prediction. Second, based on the results of this preliminary task, we extend a state-of-the-art Seq2Seq TS model, ACCESS (Martin et al., 2020), to enable explicit control of ten attributes. The results of experiments show (1) that our approach improves the performance of state-of-the-art models for predicting explainable complexity and (2) that explicitly conditioning the Seq2Seq model on ten attributes leads to a significant improvement in performance in both within-domain and out-of-domain settings.

1 Introduction

Text simplification (henceforth TS) is a natural language generation task aimed at transforming a text into an equivalent that is more readable and understandable for a target audience, while preserving the original information and underlying meaning. It involves a number of transformations applied at different linguistic levels, including lexical, syntactic and discourse aimed at reducing the complexity of content for the purpose of accessibility and readability (see Siddharthan, 2011; Shardlow, 2014; Alva-Manchego et al., 2020; Al-Thanyyan and Azmi, 2021; Jin et al., 2022, for overviews). Simplification techniques have been shown to be beneficial as reading supports across a wide range of populations, from children (De Belder and Moens, 2010; Kajiwara et al., 2013), individuals with language disorders such as aphasia (Carroll et al., 1999; Devlin and Unthank, 2006), dyslexia (Rello et al., 2013a,b) or autism (Evans et al., 2014); language learners and non-native English speakers (Petersen and Ostendorf, 2007; Paetzold and Specia, 2016), and people with low literacy skills (Max, 2006; Candido Jr et al., 2009; Watanabe et al., 2009). Moreover, TS techniques have also been successfully employed as a preprocessing step to improve the performance of various downstream NLP tasks such as parsing (Chandrasekar et al., 1996), machine translation (Gerber and Hovy, 1998; Hasler et al., 2017), summarization (Beigman Klebanov et al., 2004; Silveira and Branco, 2012), semantic role labeling (Vickrey and Koller, 2008), and information extraction (Miwa et al., 2010). TS approaches typically learn simplification transformations using parallel corpora of matched original and simplified sentences and can be classified into six categories (for recent overviews see Alva-Manchego et al., 2020; Al-Thanyyan and Azmi, 2021): Early approaches relied on either (1) manually generated rules for splitting and reordering sentences (Candido Jr et al., 2009; Siddharthan, 2011) or (2) learned simple lexical simplifications, i.e., one-word substitutions (Devlin, 1998; Carroll et al., 1998). Subsequent work has introduced (3) phrase-based and syntax-based statistical machine translation techniques (Wubben et al., 2012; Xu et al., 2016), (4) grammar induction (Paetzold and Specia, 2013; Feblowitz and Kauchak, 2013), and (5) semantics-assistance, i.e., obtaining semantic representations of the original sentences (Narayan and Gardent, 2014; Štajner and Glavaš, 2017). More recently, TS tasks have been approached with (6)

Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022), pages 125 - 146
December 8, 2022 ©2022 Association for Computational Linguistics

neural machine translation methods, in particular sequence-to-sequence (Seq2Seq) models using an attention-based encoder-decoder architecture (Nisioi et al., 2017; Alva-Manchego et al., 2017; Zhang et al., 2017). While the performance of Seq2Seq TS models is impressive, most of these models are black-box models characterized by the lack of interpretability of their procedures (Alva-Manchego et al., 2020). In recent years, there have been growing calls to a move away from black-box models toward explainable (white-box) models (Loyola-Gonzalez, 2019; Qiao et al., 2020; Aguilar et al., 2022). Moreover, recent work in TS suggests that the performance of state-of-the-art TS systems can be improved by conducting explainable complexity prediction as a preliminary step (Garbacea et al., 2021).

Another important trend in current TS research is the growing recognition that the concept of 'text complexity' is not homogeneous for different target populations (Gooding et al., 2021). That is, rather than viewing TS as a general task where the same simplification is appropriate for everyone (one-fits-all approach), researchers are placing a greater emphasis on the need to develop TS systems that can flexibly adapt to the needs of different audiences: For example, while second language learners might struggle with texts with rare or register-specific vocabulary, aphasic patients might be overwhelmed by a high cognitive load associated with long, syntactically complex sentence structures. In response, recent TS research has begun to adopt methods proposed in controllable text generation research (see the 2 section for further discussion). Controllable text generation refers to the task of generating text according to a given controlled property of a text. More generally, the development of controllable text generation systems makes an important contribution to the general development of ethical AI applications. This requires the ability to avoid biased content such as gender bias, racial discrimination, and toxic words. In addition, it is widely seen as critical to the development of advanced text generation technologies that better address specific needs in real-world applications (Prabhumoye et al., 2020; Zhang et al., 2022). For example, the task of dialog response generation requires effective control over text attributes associated with emotions (Li et al., 2021) and persona (Zhang et al., 2018). In the context of TS, the relevant attributes involve various linguistic aspects of text complexity (Siddharthan, 2011). By combining multiple attributes, a natural language generation system can theoretically achieve not only greater controllability but also greater interpretability. This requires the inclusion not only of surface features, but also of more sophisticated features. Traditionally, TS has used readability measures that consider only surface features. For example, the Flesch Reading Ease Score (Flesch, 1948), a commonly used surface feature, measures the length of words (in syllables) and sentences (in words). While readability has been shown to correlate to some degree with such features (Just and Carpenter, 1980), there is general consensus that they are insufficient to capture the full complexity of a text.

In a nutshell, despite significant progress in data-driven text simplification, the development of explainable and controllable models for automatic text simplification remains a challenge. In this paper, we advance current research on explainable and controllable text simplification in two ways:

1. First, we use what is, to our knowledge, the most comprehensive set of (psycho-)linguistic features that goes beyond traditional surface measures and includes features introduced in the recent literature on human (native and non-native) language learning and processing. These encompass lexical, syntactic, register-specific ngram, readability and psycholinguistic features and are used in combination with pre-trained language models to improve explainable complexity prediction proposed in Garbacea et al. (2021).

2. Second, based on the results of this preliminary task, we extend a state-of-the-art Seq2Seq TS model, ACCESS (Martin et al., 2020), to provide explicit control over ten attributes so that simplifications can be adapted to the linguistic needs of different audiences.

The remainder of the paper is organized as follows: Section 2 provides a concise overview of related work in the field of explainable and controllable text generation with a focus on TS. Section 3 outlines the experimental setup including the description of three benchmark datasets used (Section 3.1), the type of features extracted from these datasets (Section 3.2), and the models performed to improve explainable and controllable TS (Sections 3.3-3.5). Section 4 presents and discusses the results of our experiments before presenting conclusions and future work in Section 5. Sections 6 and 7 address the limitations of the study and point out ethical considerations.

2 Related work

State-of-the-art systems for controllable text generation typically use a Sequence-to-Sequence (Seq2Seq) architecture. These systems follow either a learning-based or a decoding-based approach: In the learning-based approaches, the Seq2Seq model is conditioned on the attribute under consideration at training time and then used to control the output at inference time. Within this approach, controlled text generation can be achieved by disentangling the latent space representations of a variational autoencoder between the text representation and the controlled attributes (Hu et al., 2017). Decoding-based methods, on the other hand, are based on a Seq2Seq training setup that is modified to control specific attributes of the output text (Kikuchi et al., 2016; Scarton and Specia, 2018). For instance, Kikuchi et al. (2016) controlled the length of the text output in the encoder-decoder framework by preventing the decoder from generating the end-of-sentence token before the desired length was reached, or by selecting only hypotheses of a certain length during the beam search. Recently, Martin et al. (2020) adapted a discrete parameterization mechanism to the task of sentence simplification by conditioning on relevant attributes. Building on the earlier work of TS (Scarton and Specia, 2018), their model, called ACCESS – short for AudienCe-CEntric Sentence Simplification – provides explicit control of TS by conditioning the output returned by the model on specific attributes. These attributes and their values are prepended as additional inputs to the source sentences at train time as plain text 'parameter tokens'. Results of experiments on the WikiLarge corpus (Zhang and Lapata, 2017) show that with carefully chosen values of three attributes - (i) character length ratio between source sentence and target sentence, (ii) normalized character-level Levenshtein similarity between source and target, and (iii) WordRank, a proxy to lexical complexity, the ACCESS model outperformed previous TS systems on simplification benchmarks, achieving state-of-the-art at 41.87 SARI, corresponding to a +1.42 improvement over the best previously reported score.

Another recently introduced line of research, on which the present work builds, explores how the transparency and explainability of the TS process can be facilitated by decomposing the task into several carefully designed subtasks. More specifically, Garbacea et al. (2021) propose that TS benefits from a preparatory task aimed at the explainable prediction of text complexity, which in turn is divided into two subtasks: (1) classifying whether a given text needs to be simplified or not (complexity prediction) and (2) highlighting the part of the text that needs to be simplified (complexity explanation). Garbacea et al. (2021) focuses on empirical analysis of the two subtasks of explicable prediction of text complexity. Specifically, they conduct experiments using a broad portfolio of deep and shallow classification models in combination with model-agnostic explanatory techniques, in particular LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). The results of their experiment show that a combination of a Long Short-Term Memory network at the word level and LIME explanations can achieve strong performance on datasets. As a next step, they conduct follow-up experiments with state-of-the-art controllable end-to-end text generation systems, including ACCESS. The results of these experiments suggest that the performance of state-of-the-art TS models can be significantly improved in out-of-sample text simplification simply by applying explainable complexity prediction as a preliminary step.

3 Experimental Setup

In this section, we first introduce the three datasets used in our experiments (Section 3.1) and the type of (psycho-)linguistic features used in our models (Section 3.2). We then describe the methods used to address the three subtasks, i.e., (1) complexity prediction, (2) complexity explanation, and (3) simplification generation. For subtask (1), we perform experiments with five complexity prediction models described in Section 3.3: (1) A word-level Long Short-Term Memory (LSTM) network, (2) a fine-tuned pre-trained BERT-based model, (3) and (4) two hybrid Bidirectional Long-Term Memory (BLSTM) classifiers that integrate GloVe word embeddings with (psycho-)linguistic features using different fusion methods, and (5) A hybrid classifier that integrates the those features with BERT representations. In subtask (2), we apply these five models to identify the complex parts of a given input set to facilitate model validation and evaluation (section 3.4). In Section 3.5, we turn to subtask (3) and introduce an extended ACCESS model, which we refer to as ACCESS-XL, containing a total of ten control features (parameter tokens) covering several dimensions of linguistic complexity.

3.1 Datasets

We conducted our experiments with three benchmark datasets and ground truth complexity labels that were also used in Garbacea et al. (2021): (1) the WikiLarge corpus Zhang et al. (2017), composed of parallel-aligned "Wikipedia-simple-Wikipedia" sentence pairs, (2) the Newsela corpus (Xu et al., 2015), comprised of news articles simplified by professional news editors, and (3) the Biendata dataset, comprising matches of research papers from different scientific disciplines with press releases describing them[1]. The size of the three datasets and their distribution among training, validation, testing datasets are shown in Table 1.

Dataset	Training	Validation	Test
Newsela	94,944	1,131	1,079
WiKiLarge	207,480	30,632	59,639
Biendata	29,710	4,244	8,490

Table 1: Number of aligned complex-simple sentence pairs by dataset

3.2 (Psycho-)Linguistic Features

The textual data of the three datasets were automatically analyzed using CoCoGen (short for Complexity Contour Generator), a computational tool that implements a sliding window technique to calculate sentence-level measurements for a given feature (for recent applications of the tool, see Kerz et al., 2020, 2022; Wiechmann et al., 2022). We extracted 107 features that fall into five categories: (1) measures of syntactic complexity (N=16), (2) measures of lexical richness (N=14), (3) register-based n-gram frequency measures (N=25), (4) readability measures (N=14), and (5) psycholinguistic measures (N=38). The first category comprises (i) surface measures that concern the length of production units, such as the mean length clauses and sentences, or (ii) measures of the type and incidence of embeddings, such as dependent clauses per T-Unit or verb phrases per sentence. These features are implemented based on descriptions in Lu (2010) using the Tregex tree pattern matching tool (Levy and Andrew, 2006) with syntactic parse trees for extracting specific patterns. The second category comprise several distinct sub-types, including (i) measures of lexical variation, i.e. the range of vocabulary as displayed in language use, captured by text-size corrected type-token ratio and (ii) lexical sophistication, i.e. the proportion of relatively unusual or advanced words in the learner's text. The operationalizations of these measures follow those described in Lu (2012) and Ströbel et al. (2016). The register-based n-gram frequency measures of the third category are derived from the five register sub-components of the Contemporary Corpus of American English (COCA, Davies, 2009): spoken, magazine, fiction, news and academic language (see Kerz et al., 2020, for details). The fourth category combine a word familiarity variable defined by pre-specified vocabulary resource to estimate semantic difficulty together with a syntactic variable, such as average sentence length. Examples of these measures are the Fry index (Fry, 1968) or the SMOG formula (McLaughlin, 1969). The psycholinguistic measures of the fifth category capture cognitive aspects of human language processing not directly addressed by the surface vocabulary and syntax features of traditional formulas. These measures include a word's average age-of-acquisition (Kuperman et al., 2012) or prevalence, which refers to the number of people knowing the word (Brysbaert et al., 2019; Johns et al., 2020). For an overview of all features, see Table 3 in the Appendix. Tokenization, sentence splitting, part-of-speech tagging, lemmatization and syntactic PCFG parsing were performed using Stanford CoreNLP (Manning et al., 2014).

3.3 Complexity prediction

For complexity prediction, i.e. the preliminary task of classifying whether a given text needs to be simplified or not, we performed experiments with five (hybrid) deep neural network architectures. Two of these prediction models are reimplementations of models used in Garbacea et al. (2021) and serve as baselines: The first model, LSTM, is a 2-layer word-level BLSTM classifier that uses GloVe word embeddings as input. The second baseline model is a 12-layer BERT model for sequence classification using a pre-trained BERT, with the first 8 layers frozen during fine-tuning.

We also conducted experiments with three hybrid models that integrate the (psycho-)linguistic features described in Section 3.2 into neural networks. GloVe-PSYLING A and GloVe-PSYLING B are hybrid BLSTM with attention models (Wu et al., 2019) that differ in how the integration was

[1] https://www.biendata.com/competition/ hackathon

performed: In model A, the linguistic features were concatenated with word embeddings before being fed into a BLSTM. In the B model, the linguistic features were concatenated with the last layer hidden state of the BLSTM. In the third hybrid model, referred to as BERT-PSYLING, we concatenated the linguistic features with the last layer output for [CLS] token from BERT. The vector representation of a sentence was then fed into a MLP classifier with ReLu as activation function. For all classifiers, Best hyperparameters were found by grid search: For BERT on WiKiLarge, the best results were obtained with a learning rate of 3×10^{-5} and a batch size of 64. For LSTM on Newsela, the learning rate was 2×10^{-4} and the batch size was 32. For BERT-PYSLING on Biendata, the learning rate was 2e-5 and the batch size was 32. We used Adam as the optimizer with $\beta = (0.9, 0.999)$ and $\epsilon = 10^{-8}$. Early stopping, where accuracy did not increase for more than 4 epochs, was used as the stopping criterion. All models were evaluated using precision, recall, F1, and classification accuracy on balanced training, validation, and testing datasets.

3.4 Complexity Explanation

The objective of the complexity explanation subtask is to highlight the part of the text that needs to be simplified. In Garbacea et al. (2021) this was achieved by quantifying the relative importance of the features in the of complexity prediction models (unigrams, bigrams, trigrams, GloVe word embeddings) using model-agnostic explanatory techniques, in particular LIME (Ribeiro et al., 2016) and SHAP (Lundberg and Lee, 2017). To afford complexity explanation of the five complexity prediction models described in section 3.3, we utilized BERT attention outputs: Since BERT uses byte-pair tokenization, we converted token attentions to word attentions by averaging the token attention weights per word. For a given attention head, the attention weights from the [CLS] token to other words at the first layer were considered as weights of those words for a given sentence. For each individual word, its final weight was the average of the weights from the 12 heads of BERT. The decision whether or not to highlight a particular word was based on a comparison of its final weight and the average of the final weights of all words in a given sentence: a word was considered complex, and thus highlighted, if its final weight fell below sentence average (see Figure 2 in the Appendix). We compare these complexity explanatory approaches with LSTM-LIME, random highlighting, and lexicon-based highlighting based on words that appear in the Age-of-Acquisition (AoA) lexicon Garbacea et al. (see 2021, for details on these basic methods). Following Garbacea et al. (2021), we evaluated the models using token-wise precision (P), recall (R), and translation edit rate (TER) (Snover et al., 2006), which assesses the minimum number of edits needed to the unhighlighted part of a source sentence so that it exactly matches the target sentence.

3.5 Simplification Generation

The original AudienCe-CEntric Sentence Simplification (ACCESS) model, introduced by Martin et al. (2020), provides explicit control of TS by conditioning the output returned by the model on specific attributes. The ACCESS model used four such parameter tokens as control features: (1) NbChars, the character length ratio between source sentence and target sentence, (2) LevSim, the normalized character-level Levenshtein similarity between source and target, which quantifies the amount of modification operated on the source sentence, (3) WordRank, a proxy to lexical complexity measured as the third-quartile of log-ranks of all words in a sentence. To get a ratio the WordRank of the target was divided by that of the source. The Seq2Seq model is parametrized on the control features by prepending a these attributes and their values as additional inputs to the source sentences as plain text 'parameter tokens'. The special token values are the ratio of this parameter token calculated on the target sentence with respect to its value on the source sentence. For example to control the number of characters of a generated simplification, the compression ratio between the number of characters in the source and the number of characters in the target sentence is computed. Ratios are discretized into bins of fixed width of 0.05 and capped to a maximum ratio of 2. Special tokens are then included in the vocabulary. At inference time, we the ratio is set a fixed value for all samples. For example, to generate simplifications that are 80% of the length of the source sentence, the token <NbChars 0.8> is prepended to each source sentence. As the Seq2Seq model, a Transformer model with a base architecture (Vaswani et al., 2017) was trained utilizing FairSeq toolkit (Ott et al., 2019a).

Our extended model, referred to here as ACCESS-XL, integrates ten of the 107 features examined in the complexity prediction step. These ten measures were selected to cover all feature groups. Within the lexical richness group, which is the largest of the five groups, features were selected to represent all subcategories of the group, i.e. length of production unit, lexical diversity, lexical sophistication, n-gram frequency, and both crowdsourcing-based and corpus-based word prevalence. Figure 4 in the Appendix shows the differences in mean standardized feature scores between 'normal' and 'simple' sentences in Wikipedia, highlighting in blue the features selected in our model. Following (Martin et al., 2020), we then trained a base transformer (Vaswani et al., 2017) using the FairSeq toolkit (Ott et al., 2019b). Both encoder and decoder consist of 6 layers. For the encoder, each of the 6 layers consists of an 8-head self-attention sub-layer and a position-wise fully connected sub-layer with a dimensionality of 2048. Each decoder layer has a similar structure, but with an additional 8-head self-attention layer that performs multi-head attention over the output of the encoder stack. The embedding size is 512. Dropout with a rate 0.2 was used for regularization. The optimizer used is the Adam optimizer with a learning rate of 0.00011, $\beta = (0.9, 0.999)$, $\epsilon = 10^{-8}$. Label smoothing with a uniform prior distribution of $\epsilon = 0.54$ was applied. Early stopping was used to prevent overfitting, with non-increasement of SARI score for more than 5 epochs as the stopping criterion. Sentencepiece with a vocabulary size of 10k was used as the tokenizer (Kudo and Richardson, 2018). Beam search with a beam size of 8 for searching for the best possible simplified sentence. A fixed combination of control tokens (a control feature along with its binned value) was used in text generation. To find the best combination, we applied the greedy forward select algorithm; we progressively added the control token from a candidate set that, in combination with the previously added control tokens, leads to the largest performance improvement in terms of SARI score on the validation set of WiKiLarge. After adding a control token to the combination, all control tokens sharing the same control feature with the newly added token were removed from the candidate set. The algorithm stopped when no control token led to an improvement in SARI score or no control token was left in the candidate set. The 5 most frequent control tokens from the WiKiLarge training set were used as the initial candidate set for each control feature, resulting in a reduction of the total search space from about 40^{10} to 5^{10}. We evaluated the output of the text simplification models using the FKGL (Flesch-Kincaid Grade Level) readability metric (Kincaid et al., 1975) to evaluate simplicity and SARI (Xu et al., 2016) as an overall performance metric, since FKGL does not take into account grammaticality and meaning preservation (Wubben et al., 2012)[2]. All scores were calculated using the EASSE python package for sentence simplification (Alva-Manchego et al., 2019)[3]. We selected the model with the best SARI on the validation set and report its score on the test set. The best combination of control tokens was as follows: MLS$_{0.50}$, Fry$_{0.85}$, FORCAST$_{0.90}$, WPCorp$_{0.95}$, WPCrowd$_{0.90}$, BigramNews$_{2.00}$, ANC$_{0.85}$, AoA$_{1.00}$, MLWs$_{0.90}$, CTTR$_{0.85}$.

4 Results

An overview of the results of the three subtasks – complexity prediction, complexity explanation and simplification generation – is presented in Table 2. We discuss the results of each subtask in turn.

Complexity prediction: Our best-performing models outperformed the classification accuracy of explainable model – the word-level LSTM - predicting complexity in all three benchmark datasets reported in Garbacea et al. (2021). Since the pattern of results is consistent across all evaluation metrics, we focus here on classification accuracy: On WikiLarge, we improve on the word-level LSTM presented in Garbacea et al. (2021) by +8.08% by extracting attention weights from the pre-trained BERT model. On Newsela, our GloVe-based LSTM model outperforms the word-level LSTM by +6.68%. On the Biendata dataset, our hybrid model that integrates BERT representations with linguistic features leads to an improvement of +4.43%. Overall, our results replicate the general pattern of results reported in Garbacea et al. (2021) in that the best-performing models achieve approximately 80% accuracy on the WikiLarge and Newsela datasets and much higher – approximately 95% accuracy – on the Biendata dataset. These results support the conclusion drawn in Garbacea et al. (2021) that complexity prediction is influenced by the application domain, with the distinction between scientific content and public domain press releases (Biendata) being much easier

[2] See Appendix for definitions and more details on these evaluation metrics.

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

	Complexity Prediction											
	WikiLarge				Newsela				Biendata			
Model	P	R	F1	Acc	P	R	F1	Acc	P	R	F1	Acc
LSTM (Garbacea et al., 2021)	-	-	-	0.716	-	-	-	0.733	-	-	-	0.898
BLSTM GloVe	0.731	0.710	0.721	0.725	0.867	**0.703**	**0.776**	**0.797**	0.923	0.889	0.906	0.907
BERT	**0.794**	**0.807**	**0.800**	**0.799**	**0.973**	0.572	0.720	0.778	0.934	**0.947**	0.940	0.940
GloVe-PSYLING A (ours)	0.766	0.781	0.773	0.771	0.929	0.609	0.736	0.781	0.930	0.915	0.922	0.923
GloVe-PSYLING B (ours)	0.762	0.783	0.772	0.769	0.925	0.604	0.731	0.778	0.924	0.928	0.926	0.926
BERT-PYSLING (ours)	0.779	**0.807**	0.793	0.789	0.972	0.580	0.727	0.782	**0.942**	0.945	**0.943**	**0.943**
	Complexity Explanation											
	WikiLarge				Newsela				Biendata			
	P	R	F1	TER↓	P	R	F1	TER↓	P	R	F1	TER↓
Random highlighting	0.410	0.463	0.457	1.084	0.550	0.488	0.504	1.029	0.803	0.424	0.550	1.011
AoA lexicon	0.407	0.549	**0.500**	1.026	0.550	0.620	0.572	0.858	0.770	0.629	0.678	0.989
LSTM+LIME	0.404	0.639	0.419	0.997	0.520	0.615	0.506	1.062	0.805	**0.826**	**0.796**	0.983
BERT	0.405	**0.660**	0.434	**0.936**	0.542	**0.729**	**0.597**	0.817	0.784	0.635	0.688	0.965
GloVe-PSYLING A (ours)	**0.454**	0.596	0.426	1.010	**0.579**	0.481	0.501	0.827	0.806	0.552	0.641	0.959
GloVe-PSYLING B (ours)	0.453	0.643	0.440	0.999	0.544	0.554	0.524	**0.816**	**0.813**	0.556	0.646	**0.951**
BERT-PSYLING (ours)	0.400	0.619	0.419	0.949	0.540	0.701	0.586	0.818	0.781	0.638	0.688	0.966
	Simplification Generation											
	WikiLarge (wd)				Newsela (OOD)				Biendata (OOD)			
	SARI↑		FK↓		SARI↑		FK↓		SARI↑		FK↓	
ACCESS	41.87		7.22		29.44		6.45		20.21		12.53	
ACCESS-XL (ours)	**43.34**		**4.39**		**34.91**		**3.96**		**27.25**		**10.71**	

Table 2: **Prediction**: Scores represent out-of-sample precision (P), recall (R), F1 and accuracy (Acc) scores. **Explanation**: P, R, and and F1 values represent token-level scores. TER scores represent Translation Edit Rates (Snover et al., 2006) **Simplification**: Scores represent out-of-sample SARI and Flesch-Kinkaid Grade Level (FK)

than the distinction between regular and simplified news articles (Newsela) or Wikepedia articles (WikiLarge).

On the WikiLarge dataset, the BERT model performed the best, with a +7.4% performance increase over the LSTM. On the Newsela dataset, however, the LSTM achieved the highest accuracy, outperforming both the BERT and BERT-HYBRID models by +1.9% and +1.5%, respectively. On the Biendata dataset, the highest performance was achieved by our BERT-HYBRID model, which improved the already high performance of the LSTM by +3.6%. Across all datasets, the GloVe word embedding-based models consistently ranked between the LSTM and BERT-based models, suggesting that the use of contextualized word embeddings of the BERT-based model may reduce the generalizability of the model, leading to variations in model performance across datasets.

Complexity explanation: The second part of Table 2 presents the results of the subtask designed to evaluate how well complexity classification can be explained, as measured by how accurately the complex parts of a sentence can be identified (highlighted). In general, all of our models showed better recall than precision, meaning that they were better at identifying words that were removed in the simplified version of a pair than words that were truly removed from the complex version. This pattern is opposite to what is reported in Garbacea et al. (2021), where precision is strongly favoured over recall. This may indicate that using average attention as a threshold may not be optimal: While this approach is the de facto standard in text style transfer research, recent work has pointed out the limitations of this approach, such as its inability of handling flat attention distributions (Lee et al., 2021)[4]. Future research may address this issue. As in the case of complexity prediction, we found that the performance of the models is dataset-specific and also varies with respect to the rank order across evaluation metrics: For WikiLarge, the BERT model achieved the best recall and TER scores, while precision was highest for the GloVE-based hybrid models (+4.5% compared to BERT). For Newsela, the BERT-based models outperformed the other models in terms of recall and F1, while the GloVe-based hybrid models achieved higher precision. All of our models significantly outperformed the three base models in terms of TER values, with the best performing model, Glove-PSYLING B, reducing the TER of the AoA method by 4.2% and that of the LSTM by as much as -24.4%. For Biendata, Glove-PSYLING B achieved the best values for precision and TER. However, the LSTM dominated the ranking in terms of recall and F1

[4] Figure 2 in the Appendix illustrates the differences in attention weight distributions among our models.

with improvements of the next best model (BERT-PSYLING) by up to 18%.

Simplification Generation We establish the state-of-the-art at 43.34 SARI on the WikiLarge test set, an improvement of +1.47 over the best previously reported result. Our ACCESS-XL text simplification model consistently outperforms the original ACCESS model (Martin et al., 2020) on all datasets and performance metrics. The performance improvement was even greater in the out-of-domain settings – with a +5.47% increase in SARI in the Newsela dataset and +7.04% in the Biendata dataset – suggesting that increased controllability also leads to increased model robustness and generalizability. For FK readability, the performance gain is even more pronounced: in the within-domain setting (WikiLarge), the ACCESS-XL model achieves a Flesch-Kinkaid score of 4.39, an improvement of -2.88. To put this number in perspective, the original ACCESS model improved previous state-of-the-art models, SBMT+PPDB+SARI (Xu et al., 2016) and PBMT-R (Wubben et al., 2012), by only -0.07 and -1.11, respectively. As in the case of SARI, the improvement in FK performance extends to both out-of-domain settings with an improvement of -2.49 for Newsela and -1.82 for Biendata. To shed more light on the textual characteristics of the outputs of the two text simplification models, we compared their average scores on the ten parameter tokens. A visualization of the results along with the scores obtained for the target and source sentences of the testset for each dataset is shown in Figure 3 in the Appendix. The comparisons revealed several important facts about the behavior of the models as well as the training data: (1) For the WikiLarge dataset, on which the model was trained, we found that the differences in average scores between the 'complex' source sentences and the 'simple' target sentences varied in magnitude: On some measures, such as mean sentence length (MLS) – a proxy of syntactic complexity, the difference between simple and complex sentences is very pronounced ($\text{MLS}_{\text{simple}}$=14.9 words, $\text{MLS}_{\text{complex}}$=22.4 words). For others, e.g. LS.ANC – a measure of lexical sophistication, the difference between the standard versions and their simplified counterparts is minimal ($\text{LS.ANC}_{\text{simple}}$=0.411, $\text{LS.ANC}_{\text{complex}}$=0.414). These results are consistent with previous indications of limitations in the WikiLarge dataset related to the high proportion of inappropriate simplifications (Xu et al., 2016). We further observed (2) that the ACCESS-XL model successfully learned to control the attributes and achieved the desired effect on the generated simplifications: For example, its outputs are characterized by much lower MLS values ($\text{MLS}_{\text{ACCESS-XL}}$ = 10.8 words) compared to the source. We note that shorter MLS values were achieved by splitting the sentence (rather than simply deleting content), which has been shown to be a weakness of current seq2seq TS models (Maddela et al., 2020). This is illustrated in the sentence set in Table 5 in the Appendix. And (3) we found that the ACCESS-XL model was able to successfully generalize its ability to control the target attributes to out-of-domain settings. For example, the learned control over the MLS parameter led to the generation of Newsela simplifications that almost matched almost perfectly the mean value of the simple sentence targets in this dataset.

Lastly, we address the question of whether explainable prediction of text complexity is still a necessary preliminary step in the pipeline when using a strong, end-to-end simplification system. We found that for all datasets – and for both the original ACCESS model and the extended ACCESS-XL model – using of preliminary complexity prediction did not improve simplification performance (see Figure 6 in the Appendix): For both SARI and FKGL evaluation metrics the best performance was invariably achieved by a model without prior indication of what sentences should undergo simplification. These results stand in stark contrast to the results reported in Garbacea et al. (2021), where prior complexity prediction was found to improve the performance of the original ACCESS model. Rather than evaluating performance using SARI and FKGL, as was the case here and in the original ACCESS publication (Martin et al., 2020), Garbacea et al. (2021) evaluated model performance using edit distance (ED), TER, and Frechet Embedding Distance. For ED alone, the reported improvements ranged from 30% to 50%. Follow up experiments based on ED, conducted to determine if the discrepancy was related to the choice of evaluation metric only confirmed the pattern of results reported here for SARI and FKGL (see Tables 7 and 8 in the Appendix). Follow-up experiments based on ED, conducted to determine if the discrepancy was related to the choice of scoring metric, only confirmed the pattern of results reported here for SARI and FKGL (see Tables 7 and 8 in the Appendix). Garbacea et al. (2021) conclude that the ACCESS model – and also the DMLMTL presented in (Guo et al., 2018), which had the highest performance for Newsela (33.22 SARI) – tends to simplify even simple inputs. Moreover, (Garbacea et al., 2021) report that over 70% of the 'simple' sentences in the test data were modified (and thus oversimplified) by the ACCESS model. Note, however, that 'simple' here means that the input sentence in question was classified as such by a preliminary complexity prediction model. Since these classifiers in WikiLarge only achieve a classification accuracy of 80%, the true percentage of oversimplification cannot be accurately estimated.

5 Conclusion and Future Work

In this work, we have advanced research on explainable and controllable text simplification in two ways: First, we have shown that performance on a prior task of explainable complexity prediction can be significantly improved by the combined use of (psycho-)linguistic features and pre-trained neural language models. And second, by extending the AudienCe-CEntric sentence simplification model to explicitly control ten text attributes, we have achieved a new state of the art in text simplification in both within-domain and out-of domain settings. In future work, we plan to apply our modeling approach to another key text style transfer task, that of formality transfer, and evaluate it on existing benchmark datasets such as the GYAFC dataset (Rao and Tetreault, 2018). Moreover, we intend to explore the role of (psycho-)linguistic features for controllable TS in unsupervised settings using a variational auto-encoder and a content predictor in combination with attribute predictors (Liu et al., 2020).

6 Limitations

The current work relies exclusively on automatic evaluation metrics for text simplification. While such metrics provide a cost-effective, reproducible, and scalable way to gauge the quality of text generation results, they also have their own weaknesses. Human scoring is necessary to address some of the inherent weaknesses of automatic evaluation (for more details, see Jin et al., 2022)

Furthermore, the performance of the proposed text simplification methods was tested on informational texts in English. While we assume that the methods can be applied to other domains and languages, we have not tested this assumption experimentally and limit our conclusions to English and the types of language registers represented in the three datasets used in this work.

References

Diana Laura Aguilar, Miguel Angel Medina Perez, Octavio Loyola-Gonzalez, Kim-Kwang Raymond Choo, and Edoardo Bucheli-Susarrey. 2022. Towards an interpretable autoencoder: a decision tree-based autoencoder and its application in anomaly detection. *IEEE Transactions on Dependable and Secure Computing*.

Suha S Al-Thanyyan and Aqil M Azmi. 2021. Automated text simplification: A survey. *ACM Computing Surveys (CSUR)*, 54(2):1–36.

Fernando Alva-Manchego, Joachim Bingel, Gustavo Paetzold, Carolina Scarton, and Lucia Specia. 2017. Learning how to simplify from explicit labeling of complex-simplified text pairs. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 295–305.

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

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

Beata Beigman Klebanov, Kevin Knight, and Daniel Marcu. 2004. Text simplification for information-seeking applications. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 735–747. Springer.

Marc Brysbaert, Paweł Mandera, Samantha F McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 english lemmas. *Behavior research methods*, 51(2):467–479.

Arnaldo Candido Jr, Erick Galani Maziero, Lucia Specia, Caroline Gasperin, Thiago Pardo, and Sandra Aluisio. 2009. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for brazilian portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42.

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

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

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

Mark Davies. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics*, 14(2):159–190.

Jan De Belder and Marie-Francine Moens. 2010. Text simplification for children. In *Prroceedings of the SIGIR workshop on accessible search systems*, pages 19–26. ACM; New York.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Siobhan Devlin and Gary Unthank. 2006. Helping aphasic people process online information. In *Proceedings of the 8th International ACM SIGACCESS Conference on Computers and Accessibility*, pages 225–226.

Richard Evans, Constantin Orasan, and Iustin Dornescu. 2014. An evaluation of syntactic simplification rules for people with autism. Association for Computational Linguistics.

Dan Feblowitz and David Kauchak. 2013. Sentence simplification as tree transduction. In *Proceedings of the second workshop on predicting and improving text readability for target reader populations*, pages 1–10.

Rudolph Flesch. 1948. A new readability yardstick. *Journal of applied psychology*, 32(3):221.

Edward Fry. 1968. A readability formula that saves time. *Journal of reading*, 11(7):513–578.

Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable prediction of text complexity: The missing preliminaries for text simplification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1086–1097, Online. Association for Computational Linguistics.

Laurie Gerber and Eduard Hovy. 1998. Improving translation quality by manipulating sentence length. In *Conference of the Association for Machine Translation in the Americas*, pages 448–460. Springer.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 2021. Word complexity is in the eye of the beholder. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4439–4449. Association for Computational Linguistics.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. *arXiv preprint arXiv:1806.07304*.

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

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P Xing. 2017. Toward controlled generation of text. In *International conference on machine learning*, pages 1587–1596. PMLR.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Brendan T Johns, Melody Dye, and Michael N Jones. 2020. Estimating the prevalence and diversity of words in written language. *Quarterly Journal of Experimental Psychology*, 73(6):841–855.

Marcel A Just and Patricia A Carpenter. 1980. A theory of reading: from eye fixations to comprehension. *Psychological review*, 87(4):329.

Tomoyuki Kajiwara, Hiroshi Matsumoto, and Kazuhide Yamamoto. 2013. Selecting proper lexical paraphrase for children. In *Proceedings of the 25th Conference on Computational Linguistics and Speech Processing (ROCLING 2013)*, pages 59–73.

Elma Kerz, Yu Qiao, Daniel Wiechmann, and Marcus Ströbel. 2020. Becoming linguistically mature: Modeling English and German children's writing development across school grades. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Elma Kerz, Yu Qiao, Sourabh Zanwar, and Daniel Wiechmann. 2022. Pushing on personality detection from verbal behavior: A transformer meets text contours of psycholinguistic features. *arXiv preprint arXiv:2204.04629*.

Yuta Kikuchi, Graham Neubig, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. *arXiv preprint arXiv:1609.09552*.

J Peter Kincaid, Robert P Fishburne Jr, Richard L Rogers, and Brad S Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Taku Kudo and John Richardson. 2018. Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Victor Kuperman, Hans Stadthagen-Gonzalez, and Marc Brysbaert. 2012. Age-of-acquisition ratings for 30,000 english words. *Behavior research methods*, 44(4):978–990.

Dongkyu Lee, Zhiliang Tian, Lanqing Xue, and Nevin L Zhang. 2021. Enhancing content preservation in text style transfer using reverse attention and conditional layer normalization. *arXiv preprint arXiv:2108.00449*.

Roger Levy and Galen Andrew. 2006. Tregex and tsurgeon: Tools for querying and manipulating tree data structures. In *LREC*, pages 2231–2234. Citeseer.

Shifeng Li, Shi Feng, Daling Wang, Kaisong Song, Yifei Zhang, and Weichao Wang. 2021. Emoelicitor: an open domain response generation model with user emotional reaction awareness. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 3637–3643.

Dayiheng Liu, Jie Fu, Yidan Zhang, Chris Pal, and Jiancheng Lv. 2020. Revision in continuous space: Unsupervised text style transfer without adversarial learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8376–8383.

Octavio Loyola-Gonzalez. 2019. Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access*, 7:154096–154113.

Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, 15(4):474–496.

Xiaofei Lu. 2012. The relationship of lexical richness to the quality of esl learners' oral narratives. *The Modern Language Journal*, 96(2):190–208.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2020. Controllable text simplification with explicit paraphrasing. *arXiv preprint arXiv:2010.11004*.

Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of 52nd annual meeting of the association for computational linguistics: system demonstrations*, pages 55–60.

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

Aurélien Max. 2006. Writing for language-impaired readers. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 567–570. Springer.

G Harry McLaughlin. 1969. Clearing the smog. *J Reading*.

Makoto Miwa, Rune Saetre, Yusuke Miyao, and Jun'ichi Tsujii. 2010. Entity-focused sentence simplification for relation extraction. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 788–796.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *The 52nd annual meeting of the association for computational linguistics*, pages 435–445.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 2: Short papers)*, pages 85–91.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019a. fairseq: A fast, extensible toolkit for sequence modeling. In *NAACL*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019b. fairseq: A fast, extensible toolkit for sequence modeling.

Gustavo Paetzold and Lucia Specia. 2013. Text simplification as tree transduction. In *Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology*.

Gustavo Paetzold and Lucia Specia. 2016. Unsupervised lexical simplification for non-native speakers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Sarah E Petersen and Mari Ostendorf. 2007. Text simplification for language learners: a corpus analysis. In *Workshop on speech and language technology in education*. Citeseer.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. *arXiv preprint arXiv:2005.01822*.

Yu Qiao, Daniel Wiechmann, and Elma Kerz. 2020. A language-based approach to fake news detection through interpretable features and brnn. In *Proceedings of the 3rd international workshop on rumours and deception in social media (RDSM)*, pages 14–31.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Luz Rello, Ricardo Baeza-Yates, Stefan Bott, and Horacio Saggion. 2013a. Simplify or help? text simplification strategies for people with dyslexia. In *Proceedings of the 10th International Cross-Disciplinary Conference on Web Accessibility*, pages 1–10.

Luz Rello, Ricardo Baeza-Yates, and Horacio Saggion. 2013b. The impact of lexical simplification by verbal paraphrases for people with and without dyslexia. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 501–512. Springer.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. Model-agnostic interpretability of machine learning. *arXiv preprint arXiv:1606.05386*.

Carolina Scarton and Lucia Specia. 2018. Learning simplifications for specific target audiences. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 712–718.

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

Advaith Siddharthan. 2011. Text simplification using typed dependencies: A comparision of the robustness of different generation strategies. In *Proceedings of the 13th European Workshop on Natural Language Generation*, pages 2–11.

Sara Botelho Silveira and António Branco. 2012. Combining a double clustering approach with sentence simplification to produce highly informative multi-document summaries. In *2012 IEEE 13th International Conference on Information Reuse & Integration (IRI)*, pages 482–489. IEEE.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*, pages 223–231.

Sanja Štajner and Goran Glavaš. 2017. Leveraging event-based semantics for automated text simplification. *Expert systems with applications*, 82:383–395.

Marcus Ströbel, Elma Kerz, Daniel Wiechmann, and Stella Neumann. 2016. CoCoGen - complexity contour generator: Automatic assessment of linguistic complexity using a sliding-window technique. In *Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity (CL4LC)*, pages 23–31, Osaka, Japan. The COLING 2016 Organizing Committee.

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

David Vickrey and Daphne Koller. 2008. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*, pages 344–352.

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.

Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. 2022. Measuring the impact of (psycho-) linguistic and readability features and their spill over effects on the prediction of eye movement patterns. *arXiv preprint arXiv:2203.08085*.

Xing Wu, Tao Zhang, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. "mask and infill" : Applying masked language model to sentiment transfer.

Sander Wubben, Antal Van Den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024.

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.

Hanqing Zhang, Haolin Song, Shaoyu Li, Ming Zhou, and Dawei Song. 2022. A survey of controllable text generation using transformer-based pre-trained language models. *arXiv preprint arXiv:2201.05337*.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. *arXiv preprint arXiv:1703.10931*.

Yaoyuan Zhang, Zhenxu Ye, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. A constrained sequence-to-sequence neural model for sentence simplification. *ArXiv*, abs/1704.02312.

7 Appendix

Table 3: Overview of the 107 features investigated in the work

Feature group	Number of features	Features	Example/Description
Syntactic complexity	16	MLC	Mean length of clause (words)
		MLS	Mean length of sentence (words)
		MLT	Mean length of T-unit (words)
		C/S	Clauses per sentence
		C/T	Clauses per T-unit
		DepC/C	Dependent clauses per clause
		T/S	T-units per sentence
		CompT/T	Complex T-unit per T-unit
		DepC/T	Dependent Clause per T-unit
		CoordP/C	Coordinate phrases per clause
		CoordP/T	Coordinate phrases per T-unit
		NP.PostMod	NP post-mod (word)
		NP.PreMod	NP pre-mod (word)
		CompN/C	Complex nominals per clause
		CompN/T	Complex nominals per T-unit
		VP/T	Verb phrases per T-unit
Lexical richness	14	MLWc	Mean length per word (characters)
		MLWs	Mean length per word (sylables)
		LD	Lexical density
		NDW	Number of different words
		CNDW	NDW corrected by Number of words
		TTR	Type-Token Ration (TTR)
		cTTR	Corrected TTR
		rTTR	Root TTR
		AFL	Sequences Academic Formula List
		ANC	LS (ANC) (top 2000, inverted)
		BNC	LS (BNC) (top 2000, inverted)
		NAWL	LS New Academic Word List
		NGSL	LS (General Service List) (inverted)
		NonStopWordsRate	Ratio of words in NLTK non-stopword list
Register-based	25	Spoken ($n \in [1, 5]$)	Frequencies of uni-, bi-
		Fiction ($n \in [1, 5]$)	tri-, four-, five-grams
		Magazine ($n \in [1, 5]$)	from the five sub-components
		News ($n \in [1, 5]$)	(genres) of the COCA
		Academic ($n \in [1, 5]$)	

Feature group	Number of features	Features	Example/Description
Readability	14	ARI	Automated Readability Index
		ColemanLiau	Coleman-Liau Index
		DaleChall	Dale-Chall readability score
		FleshKincaidGradeLevel	Flesch-Kincaid Grade Level
		FleshKincaidReadingEase	Flesch Reading Ease score
		Fry-x	x coord. on Fry Readability Graph
		Fry-y	y coord. on Fry Readability Graph
		Lix	Lix readability score
		SMOG	Simple Measure of Gobbledygook
		GunningFog	Gunning Fog Index readability score
		DaleChallPSK	Powers-Sumner-Kearl Variation of the Dale and Chall Readability score
		FORCAST	FORCAST readability score
		Rix	Rix readability score
		Spache	Spache readability score
Psycholinguistic	38	WordPrevalence	See Brysbaert et al. (2019)
		Prevalence	Word prevalence list incl. 35 categories (Johns et al. (2020))
		AoA-mean	avg. age of acquisition (Kuperman et al. (2012))
		AoA-max	max. age of acquisition

Table 4: Means and standard deviations of all engineered languge features across the ‘normal’ and ‘simple’ sentences in the three benchmark datasets

	Biendata				Newsela				WikiLarge			
	normal		simple		normal		simple		normal		simple	
Feature	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
LexDens	0.73	0.1	0.76	0.12	0.58	0.1	0.58	0.11	0.58	0.12	0.6	0.17
CTTR	3.9	0.66	3.53	0.56	4.69	0.83	3.94	0.67	4.33	0.92	3.71	1.06
RTTR	2.65	0.42	2.39	0.37	3.21	0.57	2.69	0.44	2.96	0.63	2.54	0.7
TTR	0.97	0.05	0.99	0.04	0.91	0.07	0.95	0.06	0.88	0.1	0.92	0.1
MLWc	6.59	1.09	5.87	1.03	4.89	0.61	4.67	0.67	4.98	0.83	4.95	1.19
MLWs	2.02	0.36	1.73	0.35	1.47	0.2	1.39	0.21	1.52	0.25	1.49	0.37
Prev.AllAP	6.25	1.07	7.12	0.74	7.3	0.59	7.38	0.68	6.54	1.25	6.51	1.5
Prev.AllBP	7.48	1.31	8.58	0.95	8.98	0.75	9.11	0.87	8	1.56	7.96	1.86
Prev.AllCD	9.43	1.73	10.65	1.4	11.98	1.13	12.25	1.3	10.66	2.14	10.6	2.59
Prev.AllSD	7.55	1.35	8.79	1	9.14	0.77	9.32	0.89	8.14	1.57	8.14	1.89
Prev.AllSDAP	3.63	0.69	4.23	0.5	4.44	0.38	4.51	0.44	3.95	0.77	3.93	0.93
Prev.AllSDBP	5.06	0.98	5.91	0.75	6.34	0.57	6.47	0.66	5.61	1.12	5.59	1.36
Prev.AllWF	10.03	1.85	11.18	1.51	12.74	1.22	13.01	1.41	11.39	2.3	11.31	2.79
Prev.FemAP	5.58	1.02	6.45	0.71	6.67	0.55	6.75	0.64	5.95	1.15	5.93	1.38
Prev.FemBP	6.72	1.26	7.81	0.92	8.26	0.71	8.4	0.83	7.32	1.45	7.3	1.74
Prev.FemCD	8.79	1.69	9.98	1.39	11.37	1.1	11.65	1.27	10.09	2.05	10.04	2.49
Prev.FemSD	6.96	1.31	8.18	0.99	8.6	0.74	8.79	0.86	7.64	1.49	7.64	1.8
Prev.FemSDAP	3.01	0.62	3.56	0.46	3.79	0.34	3.86	0.39	3.34	0.67	3.33	0.81
Prev.FemSDBP	4.35	0.91	5.16	0.72	5.63	0.53	5.76	0.61	4.94	1.02	4.93	1.24
Prev.FemWF	9.2	1.78	10.32	1.48	11.91	1.18	12.19	1.36	10.62	2.18	10.55	2.66
Prev.MaleAP	5.69	0.97	6.47	0.67	6.63	0.53	6.7	0.62	5.95	1.13	5.92	1.36
Prev.MaleBP	6.99	1.23	8.01	0.89	8.38	0.7	8.51	0.81	7.48	1.45	7.45	1.74
Prev.MaleCD	9.01	1.67	10.18	1.36	11.5	1.09	11.76	1.26	10.23	2.06	10.18	2.5
Prev.MaleSD	7.23	1.3	8.41	0.97	8.79	0.74	8.96	0.86	7.82	1.51	7.82	1.82
Prev.MaleSDAP	2.92	0.56	3.39	0.4	3.57	0.3	3.62	0.35	3.18	0.62	3.16	0.75
Prev.MaleSDBP	4.45	0.87	5.18	0.66	5.59	0.51	5.7	0.58	4.95	0.99	4.93	1.2
Prev.MaleWF	9.48	1.78	10.56	1.46	12.11	1.18	12.37	1.36	10.84	2.2	10.75	2.68
Prev.UKAP	4.97	0.9	5.73	0.63	5.93	0.49	6	0.56	5.31	1.02	5.29	1.23
Prev.UKBP	6.22	1.16	7.2	0.85	7.61	0.66	7.73	0.76	6.78	1.33	6.75	1.6
Prev.UKCD	8.26	1.59	9.38	1.33	10.72	1.05	10.99	1.21	9.52	1.94	9.47	2.36
Prev.UKSD	6.46	1.22	7.6	0.93	7.97	0.69	8.15	0.81	7.07	1.38	7.08	1.67
Prev.UKSDAP	2.42	0.5	2.85	0.38	3.05	0.28	3.1	0.32	2.71	0.54	2.7	0.66
Prev.UKSDBP	3.79	0.79	4.47	0.63	4.89	0.47	5.01	0.54	4.32	0.89	4.31	1.08
Prev.UKWF	8.72	1.7	9.75	1.43	11.33	1.14	11.59	1.31	10.11	2.09	10.03	2.55
Prev.USAAP	5.84	1.01	6.68	0.7	6.86	0.55	6.94	0.64	6.14	1.18	6.11	1.41
Prev.USABP	7.08	1.27	8.15	0.92	8.56	0.72	8.69	0.84	7.61	1.49	7.58	1.78
Prev.USACD	9.12	1.7	10.33	1.39	11.67	1.11	11.95	1.29	10.38	2.09	10.33	2.54
Prev.USASD	7.25	1.33	8.49	0.99	8.86	0.75	9.04	0.88	7.87	1.52	7.88	1.84
Prev.USASDAP	3.24	0.63	3.8	0.46	4.01	0.35	4.07	0.4	3.54	0.7	3.53	0.85
Prev.USASDBP	4.67	0.93	5.49	0.72	5.94	0.54	6.06	0.63	5.23	1.06	5.21	1.28
Prev.USAWF	9.55	1.81	10.68	1.48	12.24	1.19	12.51	1.38	10.93	2.23	10.85	2.71
AFL	0	0	0	0.01	0	0.01	0	0.01	0	0.01	0	0.01
ANC	0.53	0.15	0.46	0.17	0.32	0.12	0.29	0.14	0.42	0.16	0.42	0.22
BNC	0.7	0.12	0.67	0.14	0.53	0.11	0.51	0.14	0.6	0.14	0.62	0.18
NAWL	0.07	0.08	0.05	0.08	0.01	0.03	0.01	0.03	0.02	0.04	0.01	0.05
NGSL	0.43	0.16	0.29	0.16	0.22	0.12	0.19	0.13	0.35	0.18	0.35	0.23

	Biendata				Newsela				WikiLarge			
	normal		simple		normal		simple		normal		simple	
Feature	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
ngram1acad	100.3	45.99	82.6	30.9	218.21	97.56	134.33	54.24	191.58	106.77	134.35	89.59
ngram1fic	80.26	39.83	73.47	29.2	211.26	93.65	132.17	53.26	179.99	101.04	127.55	85.76
ngram1mag	94.13	43.86	82.86	30.58	222.63	98.35	137.77	54.8	191.85	106.55	135.32	89.95
ngram1news	86.11	43.24	78.63	30.42	222.82	98.5	137.91	54.84	190.63	105.75	134.58	89.44
ngram1spok	84.23	42.43	77.21	30.28	218.49	97.09	136.38	54.85	183.53	103.33	130.18	87.59
ngram2acad	11.07	12.58	8.13	9.15	41.46	30.02	27.72	21.32	32.37	28.85	24.59	24.74
ngram2fic	3.55	5.37	4.26	6.69	33.83	25.95	25.47	20.46	22.31	21.55	18.68	19.98
ngram2mag	7.87	9.44	8.24	9.36	45.95	31.26	31.95	22.9	32.18	27.69	25.37	24.67
ngram2news	6.25	8.25	6.35	8.02	47.49	32.39	32.88	23.57	31.52	27.44	24.99	24.53
ngram2spok	5.45	7.44	6.04	7.99	42.87	31	31.11	23.43	26.57	24.46	22.08	22.77
ngram3acad	0.82	1.97	0.56	1.5	3.81	5.23	2.89	4.53	3.12	5.06	2.72	4.67
ngram3fic	0.15	0.65	0.24	0.96	2.58	4.17	2.37	4.07	1.4	2.68	1.44	2.88
ngram3mag	0.47	1.3	0.58	1.57	4.52	5.8	3.65	5.25	2.87	4.53	2.67	4.41
ngram3news	0.36	1.12	0.42	1.26	4.91	6.19	3.96	5.61	2.85	4.62	2.68	4.53
ngram3spok	0.28	1	0.36	1.22	3.87	5.68	3.41	5.33	1.95	3.58	2.05	3.86
ngram4acad	0.09	0.42	0.06	0.32	0.41	1.06	0.34	1.02	0.35	1.04	0.32	0.97
ngram4fic	0.01	0.13	0.02	0.2	0.24	0.76	0.24	0.82	0.12	0.42	0.13	0.5
ngram4mag	0.05	0.26	0.07	0.35	0.52	1.21	0.45	1.21	0.31	0.89	0.31	0.91
ngram4news	0.04	0.23	0.04	0.26	0.57	1.29	0.5	1.28	0.3	0.92	0.29	0.94
ngram4spok	0.03	0.19	0.04	0.25	0.41	1.13	0.4	1.17	0.19	0.69	0.21	0.79
ngram5acad	0.01	0.16	0.01	0.09	0.07	0.33	0.05	0.35	0.05	0.3	0.05	0.29
ngram5fic	0	0.03	0	0.06	0.03	0.18	0.03	0.2	0.01	0.1	0.02	0.14
ngram5mag	0.01	0.09	0.01	0.1	0.09	0.38	0.07	0.38	0.05	0.25	0.04	0.24
ngram5news	0	0.07	0.01	0.08	0.09	0.37	0.08	0.38	0.05	0.28	0.04	0.28
ngram5spok	0	0.06	0	0.07	0.06	0.31	0.06	0.32	0.03	0.18	0.03	0.21
NonStopW	0.74	0.1	0.78	0.12	0.6	0.1	0.59	0.12	0.63	0.12	0.64	0.17
AoA.max	12.64	2.47	10.89	2.53	10.19	2.33	8.4	2.16	10.36	2.78	8.96	3.25
AoA.mean	7.43	1.34	6.8	1.31	5.55	0.72	5.22	0.74	5.73	1.16	5.45	1.68
WordPrev	1.62	0.42	2.04	0.29	1.99	0.28	2.01	0.33	1.62	0.49	1.59	0.58
KolDef	0.85	0.12	0.93	0.12	0.77	0.12	0.89	0.13	0.8	0.23	0.93	0.35
NPPostMod	6.41	5.6	2.8	3.3	3.99	5.76	2.03	3.17	5.64	6.44	3.58	4.73
NPPreMod	1.27	1.14	1.02	0.88	1.03	0.86	0.91	0.73	1.21	1.01	1.04	0.87
CpS	0.31	0.5	0.77	0.69	2.11	1.23	1.58	0.86	1.45	1.01	1.19	0.93
CpT	0.27	0.47	0.66	0.66	1.88	1.07	1.49	0.8	1.28	0.8	1.08	0.77
CompNompC	0.67	1.23	1.01	1.08	1.57	1.2	1.09	0.9	1.97	1.51	1.3	1.24
CompNompT	0.8	1.29	1.15	1.13	2.65	1.85	1.52	1.18	2.5	1.85	1.61	1.52
CompTpT	0.02	0.13	0.1	0.3	0.53	0.49	0.37	0.48	0.27	0.44	0.2	0.39
CoordPpC	0.12	0.36	0.06	0.24	0.32	0.52	0.18	0.39	0.45	0.68	0.28	0.52
CoordPpT	0.15	0.41	0.07	0.26	0.51	0.72	0.23	0.47	0.56	0.79	0.34	0.61
DCpC	0.04	0.17	0.12	0.29	0.3	0.29	0.2	0.27	0.16	0.27	0.12	0.24
DCpT	0.02	0.14	0.1	0.32	0.77	0.9	0.45	0.66	0.34	0.62	0.24	0.53
MLC	3.71	6.28	5.83	4.96	12.23	6.89	9.4	4.45	14.63	8.87	10.5	7.5
MLS	12.76	4.46	9.62	2.86	22.76	9.77	13.74	5.12	21.13	10.6	14.67	9.03
MLT	4.62	6.69	6.67	5.02	20.54	10.24	12.96	5.47	18.77	11.09	12.95	9.35
TpS	0.36	0.48	0.7	0.48	1.06	0.39	1	0.27	0.99	0.44	0.88	0.47
VPpT	0.41	0.64	0.93	0.82	2.46	1.42	1.87	1.04	1.56	1.08	1.26	0.98

	Biendata				Newsela				WikiLarge			
	normal		simple		normal		simple		normal		simple	
Feature	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
ARI	15.98	5.01	11.02	4.62	12.98	5.66	7.45	3.85	12.63	6.14	9.27	6.19
Coleman	54.6	26.05	35.59	16.63	111.87	57.51	58.53	30.06	102.55	62.16	64.31	52.94
DaleChall	10.16	2.13	8.9	2.7	6.2	1.96	5.44	2.3	7.58	2.49	7.48	3.37
DC.PSK	11.41	1.53	10.31	1.95	9.06	1.53	8.01	1.68	9.99	1.8	9.56	2.37
FK Grade	13.23	4.27	8.53	4.05	10.61	4.55	6.16	3.15	10.56	4.97	7.72	5.16
FK Read	22.95	30.03	51.06	29.3	59.52	19.89	75.39	18.56	56.99	23.43	65.85	31.32
FORCAST	13.23	2.14	11.86	2.62	9.79	1.67	9.23	1.96	10.2	1.93	10.08	2.91
Fry.x	202.05	36.22	172.58	35.25	146.84	19.97	138.88	21.31	151.77	25.29	149.05	37.25
Gunning	510.4	178.2	385.0	114.4	910.4	390.9	549.6	204.7	846.5	423.0	587.4	361.1
Lix	61.4	14.53	48.27	16.85	48.26	14.61	35.23	12.96	48.96	15.67	41.45	19.55
Rix	5.9	2.98	5.33	2.51	15.49	6.96	9.91	4.12	13.96	7.55	10.08	6.6
SMOG	8.78	1.64	7.18	2.18	6.26	1.55	5.49	1.86	6.45	1.71	5.88	2.23
Spache	2.25	0.54	1.86	0.34	3.41	1.17	2.33	0.61	3.24	1.27	2.47	1.08

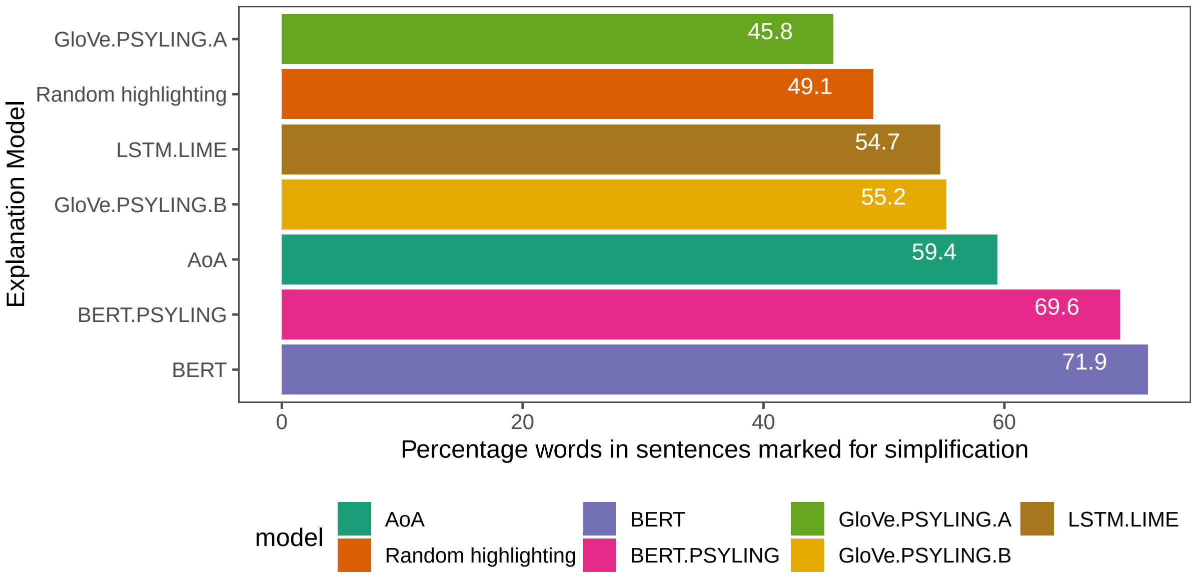

Figure 1: **Complexity explanation:** Differences in mean percentages of highlighted words across the five explanation models compared along with the two baselines: 'Random highlighting' and highlighting based on AoA (=age of acquisition) lexicon (Kuperman et al., 2012).

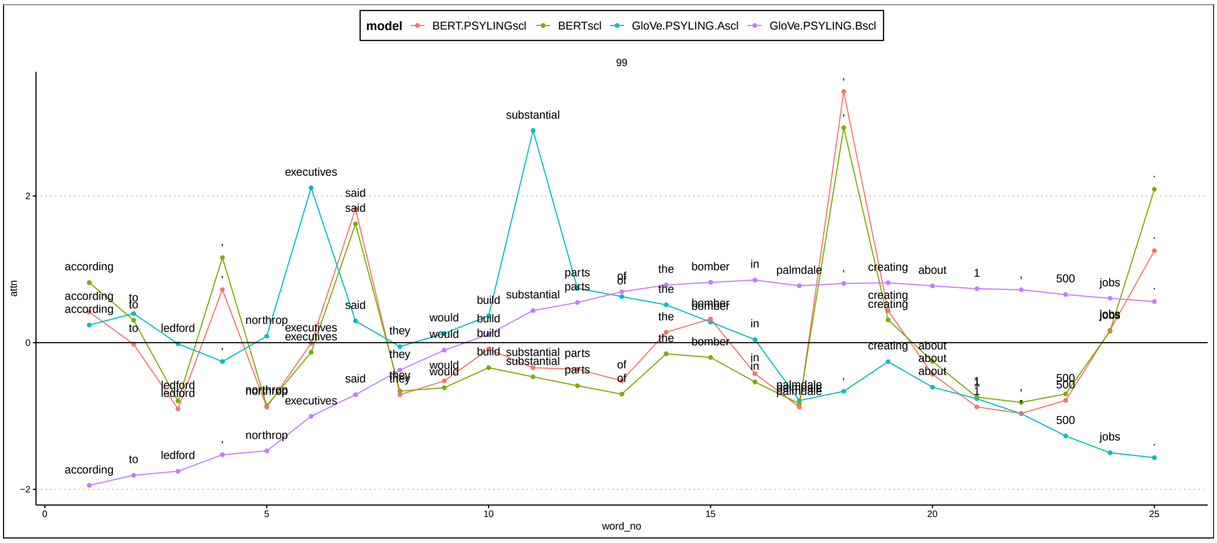

Figure 2: **Complexity explanation:** Distributions of attention weights over words in a randomly selected sentence.

Table 5: **Simplification Generation:** Example pair from WikiLarge corpus (normal, simplified) and source sentence simplified by ACCESS model (including four parameter tokens) and ACCESS-XL (including ten parameter tokens).

Type	Sentence
Source (Wikipedia)	One side of the armed conflicts is composed mainly of the Sudanese military and the Janjaweed, a Sudanese militia group recruited mostly from the Afro-Arab Abbala tribes of the northern Rizeigat region in Sudan.
Target (WikiSimple)	One side of the armed conflicts is made of Sudanese military and the Janjaweed, a Sudanese militia recruited from the Afro-Arab Abbala tribes of the northern Rizeigat region in Sudan.
ACCESS	One side of the armed conflict is made up of the Sudanese military and the Janjaweed, a Sudanese militia group brought mostly from the Afro-Arab Abbala tribes of the northern Rizeigat region in Sudan.
ACCESS-XL	The army of the armed conflicts is mainly made of the Sudanese military and the Janjaweed, a Sudanese militia group. They recruited mostly from the Afro-Arab Abbala tribes of the northern Rizeigat region in Sudan.

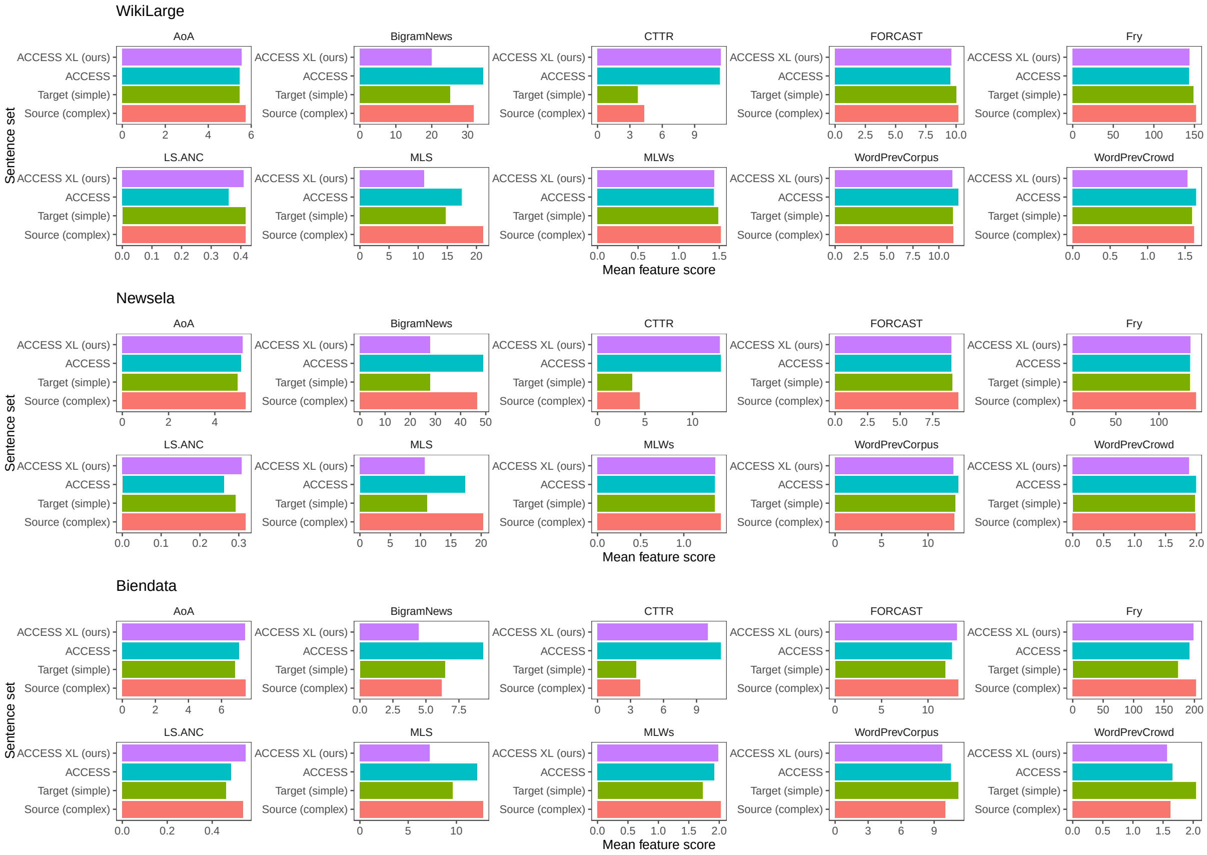

Figure 3: **Simplification Generation:** Mean values of the ten parameter tokens (engineered language features) across sentences sets.

Table 6: ACCESS model performance with prior complexity prediction using different complexity prediction models.

		ACCESS			
		Ours		Martin	
Dataset	Filter	SARI	FKGL	SARI	FKGL
WiKiLarge	BERT	43.01	5.14	40.97	7.21
	BERT_PSYLING	42.84	5.06	40.97	7.17
	GloVe-PSYLING-a	41.38	5.19	39.54	7.24
	GloVe-PSYLING-b	41.53	5.03	39.72	7.22
Biendata	BERT	26.92	10.85	19.93	12.61
	BERT_PSYLING	26.87	10.86	19.87	12.63
	GloVe-PSYLING-a	26.16	11.17	19.31	12.78
	GloVe-PSYLING-b	26.87	10.90	19.89	12.62
Newsela	bert	33.44	5.27	27.33	6.78
	BERT_PSYLING	33.13	5.19	27.30	6.75
	GloVe-PSYLING-a	34.88	3.96	29.41	6.45
	GloVe-PSYLING-b	34.90	3.96	29.43	6.45

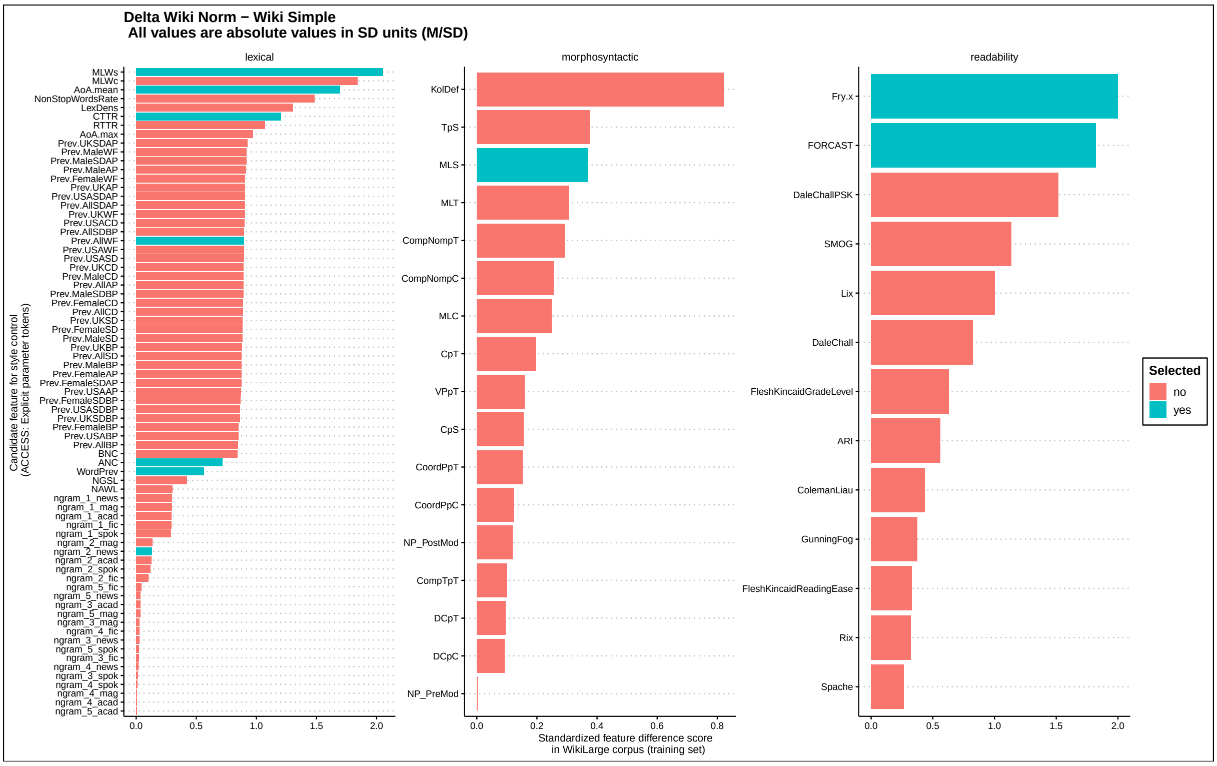

Figure 4: **Simplification Generation:** Differences in mean feature scores (standardized) between ‘normal’ and ‘simple’ sentences in WikiLarge corpus. Features in blue were selected for controllable sentence simplification in the ACCESS-XL model.

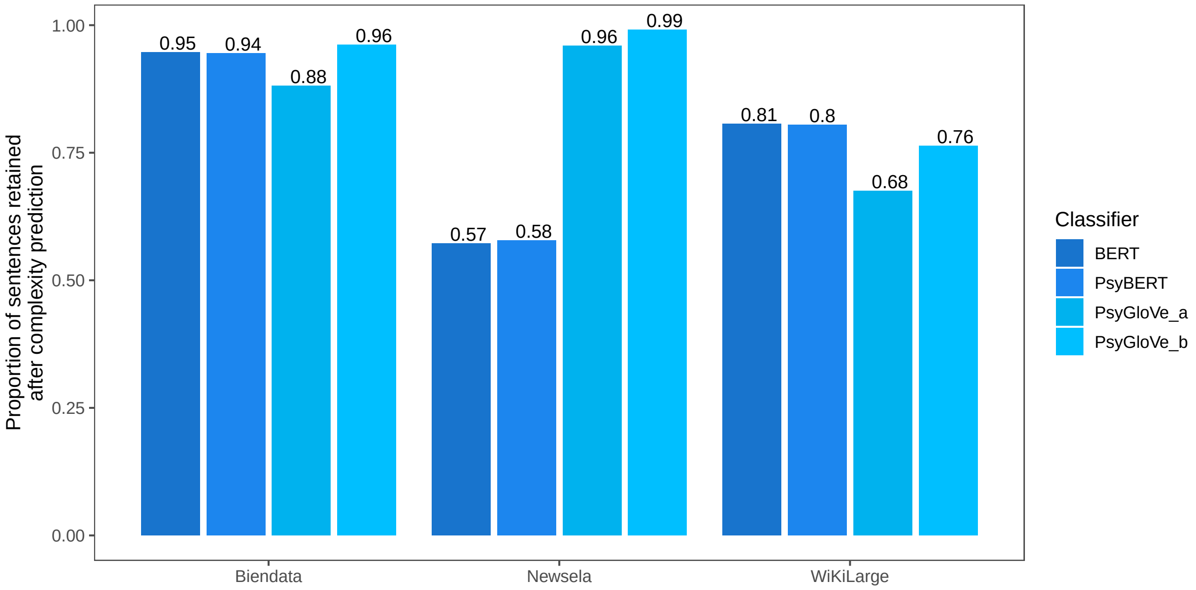

Figure 5: **Simplification Generation:** Proportion of sentences retained after complexity prediction after complexity prediction (step 1) across prediction model and dataset

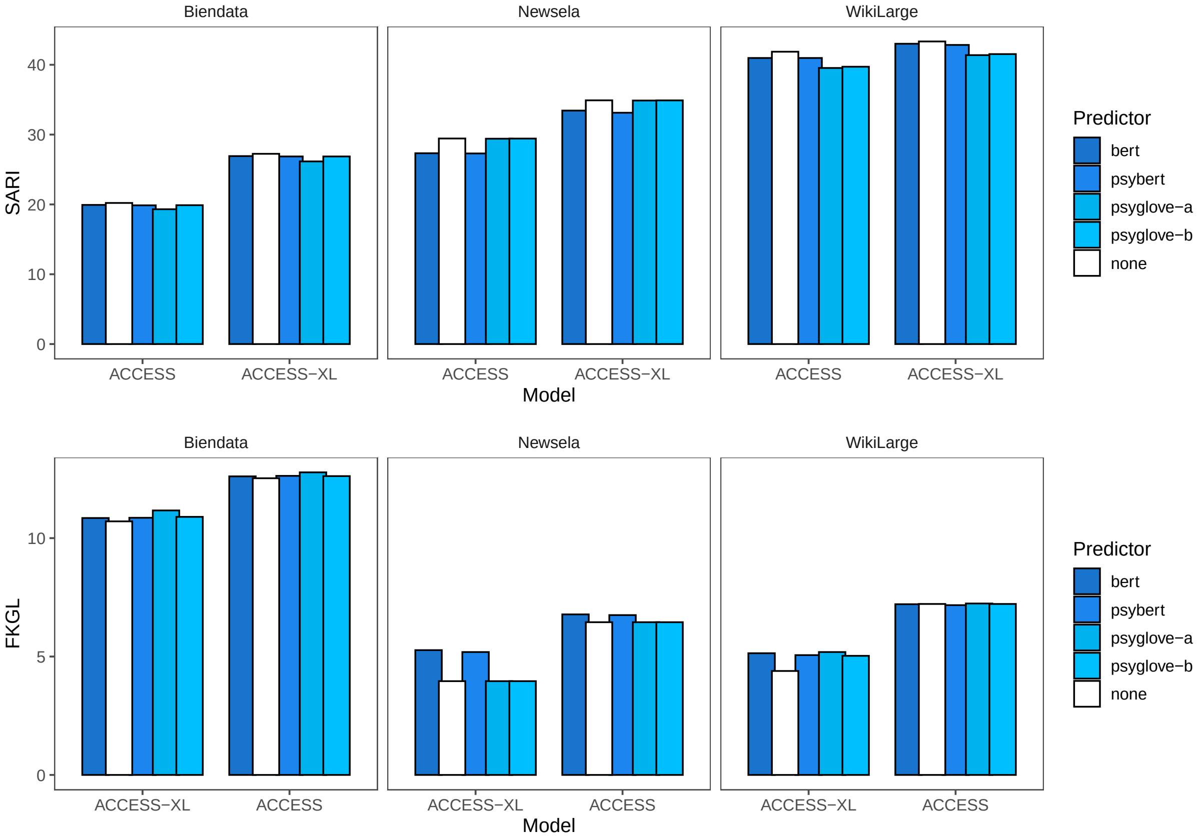

Figure 6: **Simplification Generation:** Performance of text simplification models as measured by SARI (top, higher is better) and Flesch-Kinkaid Grade Level (FKGL, bottom; lower is better) across datasets and use of complexity prediction methods.

Table 7: Average ED between simple sentences and original ACCESS output predictions with and without complexity prediction. ED are calculated using the tseval library, which EASSE relies on.

Dataset	Filter	ED
WiKiLarge	none	15.641
	BERT	15.566
	BERT_PSYLING	15.590
	GloVe-PSYLING_a	15.717
	GloVe-PSYLING_b	15.771
	LSTM	15.705
biendata	none	13.298
	BERT	13.269
	BERT_PSYLING	13.267
	GloVe-PSYLING_a	13.240
	GloVe-PSYLING_b	13.281
	LSTM	13.220
newsela	none	16.378
	BERT	15.958
	BERT_PSYLING	15.957
	GloVe-PSYLING_a	16.377
	GloVe-PSYLING_b	16.376
	LSTM	16.008

Table 8: Avg ED between complex sentences and original ACCESS outputs with/without complexity prediction

Dataset	Filter	ED
WiKiLarge	none	6.684
	BERT	5.916
	BERT_PSYLING	5.979
	GloVe-PSYLING_a	5.639
	GloVe-PSYLING_b	5.765
	LSTM	4.516
biendata	none	2.823
	bert	2.719
	BERT_PSYLING	2.699
	GloVe-PSYLING_a	2.529
	GloVe-PSYLING_b	2.723
	LSTM	1.585
newsela	none	5.368
	BERT	3.918
	BERT_PSYLING	3.960
	GloVe-PSYLING_a	5.358
	GloVe-PSYLING_b	5.363
	LSTM	3.022

Table 9: **Simplification Generation:** Proportion of sentences retained after complexity prediction after complexity prediction (step 1) across prediction model and dataset

	Complexity prediction model			
Dataset	BERT	PsyBERT	PsyGloVe$_a$	PsyGloVe$_b$
Biendata	0.947	0.945	0.881	0.961
Newsela	0.572	0.578	0.959	0.991
WiKiLarge	0.807	0.805	0.675	0.764

Evaluation metrics for simplification generation FKGL is computed as a linear combination of the number of words per simple sentence and the number of syllables per word:

$$FKGL = 0.39\frac{N\ word}{N\ sent} + 11.8\frac{N\ syl}{N\ word} - 15.59$$

SARI compares the predicted simplification with both the source and the target reference. It is an average of F1 scores for three n-gram operations: additions (add), keeps ($keep$) and deletions (del). For each operation, these scores are then averaged for all n-gram orders (from 1 to 4) to get the overall F1 score.

$$f_{ope}(n) = \frac{2 \times p_{ope}(n) \times r_{ope}(n)}{p_{ope}(n) + r_{ope}(n)}$$

$$F_{ope} = \frac{1}{k} \sum_{n=[1,\ldots,k]} f_{ope}(n)$$

$$SARI = \frac{F_{add} + F_{keep} + F_{del}}{3}$$

SARI thus rewards models for adding n-grams that occur in the reference but not in the input, for keeping n-grams both in the output and in the reference, and for not over-deleting n-grams. Xu et al. (2016) show that SARI correlates with human judgments of simplicity gain.