Learning to Paraphrase Sentences to Different Complexity Levels

Alison Chi, Li-Kuang Chen, Yi-Chen Chang*, Shu-Hui Lee*, Jason S. Chang
National Tsing Hua University, Hsinchu, Taiwan
achi@gapp.nthu.edu.tw, lkchen@gapp.nthu.edu.tw,
yichen@nlplab.cc, shlee@nlplab.cc, jason@nlplab.cc

Abstract

While sentence simplification is an active research topic in NLP, its adjacent tasks of sentence complexification and same-level paraphrasing are not. To train models on all three tasks, we present two new unsupervised datasets. We compare these datasets, one labeled by a weak classifier and the other by a rule-based approach, with a single supervised dataset. Using these three datasets for training, we perform extensive experiments on both multitasking and prompting strategies. Compared to other systems trained on unsupervised parallel data, models trained on our weak classifier labeled dataset achieve state-of-the-art performance on the ASSET simplification benchmark. Our models also outperform previous work on sentence-level targeting. Finally, we establish how a handful of Large Language Models perform on these tasks under a zero-shot setting.

1 Introduction

Paraphrasing a sentence to a targeted level of complexity is a natural language processing task that has not received much attention. Most work focuses solely on sentence simplification: decreasing the syntactic and lexical complexity of a sentence in order to make it easier to understand while preserving its original meaning (Siddharthan, 2002, 2006; Zhu et al., 2010; Woodsend and Lapata, 2011; Xu et al., 2015; Zhang and Lapata, 2017; Alva-Manchego et al., 2020b). This task has applications for second language (L2) learners and people with neural conditions that impede their reading comprehension abilities (Alva-Manchego et al., 2020b). There has been limited work on sentence complexification, which is the exact opposite of sentence simplification: increasing the syntactic and lexical complexity of a given sentence (Berov and Standvoss, 2018).

As far as we know, there has not been any work done on same-level paraphrasing, which we define as paraphrasing a given sentence without changing its complexity level. However, all three tasks have important potential applications in computer-assisted language learning.

Services like Grammarly[1] and LinggleWrite (Tsai et al., 2020) aim to correct grammatical and lexical writing errors, especially for L2 learners. Others aim to generate example usage sentences for new words (Huang et al., 2017), as well as suggest potential paraphrases of learners' sentences in order to improve the diversity of their writing (Chen et al., 2015). In addition to suggesting general paraphrase rewrites, the online writing assistant WordTune[2] allows users to control both the length (correlated to complexity) and formality level of its paraphrase suggestions (Zhao, 2022).

Despite the existence of these paraphrasing systems commercially, to the best of our knowledge, there has been no academic work on paraphrasing to different complexity levels. Writing assistants and general language learning systems could benefit from this. A learner might want to see more concise ways of expressing their ideas (simplifications), more advanced or idiomatic ways of expressing them (complexifications), or suggestions that match their writing level (same-level paraphrases). We present models for all three tasks. For these tasks, we construct two automatically labeled (unsupervised) datasets and compare them to one human-labeled (supervised) dataset.

Our first automatic labeling method is rule-based according to Flesch-Kincaid Grade Level (FKGL). FKGL can be calculated automatically

*Equal contribution.

[1] https://www.grammarly.com.
[2] https://www.wordtune.com.

Transactions of the Association for Computational Linguistics, vol. 11, pp. 1332–1354, 2023. https://doi.org/10.1162/tacl_a_00606
Action Editor: Sara Rosenthal. Submission batch: 3/2023; Revision batch: 7/2023; Published 11/2023.

as a weighted score consisting of sentence length and syllable information (Kincaid et al., 1975). A lower score means simpler output, and the lowest possible score is −3.40. Although this metric has been widely used for automatic evaluation of sentence simplification systems, it has been criticized for being easy to manipulate without increasing the simplification quality of the output (Tanprasert and Kauchak, 2021).

Our second automatic labeling method is weak classification according to the six Common European Framework of Reference for Languages (CEFR) levels. The CEFR is used in standardized testing around the world to describe the language ability of L2 learners.[3] It contains six levels in the order of increasing complexity: A1, A2, B1, B2, C1, and C2.[4] Unlike FKGL, the CEFR is based on a holistic combination of lexical, syntactic, and conceptual features and requires professionals to determine scoring (Council of Europe, 2001). We construct a new, weakly labeled CEFR-annotated sentence and phrase dataset from the English Profile and Cambridge Dictionary, which we call CEFR-CEP (CEFR-Cambridge-English-Profile). We train a classifier to classify sentences and phrases into any of the six levels.

From the ParaNMT dataset (Wieting and Gimpel, 2018), we create both CEFR-labeled and FKGL-labeled unsupervised sentence simplification, complexification, and same-level paraphrasing datasets. We also use a supervised dataset called Newsela-Auto (Jiang et al., 2020). On all three datsets, we fine-tune T5 models. We conduct ablation studies on multitasking configurations, comparing performance of single-task, two-task, and three-task models. We also compare two prompting strategies: absolute prompting, where we prepend target complexity level to the input sentence, and relative prompting, where we prepend level direction to the input sentence. Finally, we assess how Large Language Models (LLMs) perform on these tasks in a zero-shot setting. Our contributions are as follows:

- To our knowledge, we are the first to attempt the task of changing complexity level in any direction. From our in-depth fine-tuning experiments as well as a brief study on how well LLMs can change complexity, we establish new benchmarks.
- Our CEFR-labeled ParaNMT dataset produces state-of-the-art results on the ASSET simplification benchmark for models trained on unsupervised parallel data.
- Our absolute prompting models outperform previous level targeting work on the Newsela-Manual benchmark.
- We release our ParaNMT data, CEFR classifier, and best fine-tuned paraphrasing models to the public.[5] We also release the CEFR-CEP test data used for human evaluation. The source dataset is publicly available on EVP, EGP, and Cambridge websites and can be obtained via their data request process.[6]

2 Related Work

2.1 Sentence Complexity Classification

Much work has been done on complexity level classification as a component of Automatic Readability Assessment, but it has mostly focused on the document level (Xia et al., 2016; Lee et al., 2021) and not the sentence level due to a shortage of sentence-level datasets. In English, data from Newsela,[7] which contains articles that have been manually simplified to four different target levels, has been widely used (Xu et al., 2015; Lee et al., 2021; Lee and Vajjala, 2022). Newsela sentence levels can be automatically derived for sentence-level research. However, since Newsela levels (US grade ranges) are per document, not every sentence level corresponds to its document's level. The OneStopEnglish corpus (Vajjala and Lučić, 2018), which consists of sentences and documents labeled at three ESL levels, is also widely used. Since readability is highly subjective and dependent on a specific audience or set of standards, it is difficult to apply a single readability assessment scheme to a variety of domains. Lee and Vajjala's (2022) pairwise ranking model

[3] https://www.cambridgeenglish.org/exams-and-tests/cefr.

[4] Levels that fall within the same letter are closer together than those that belong to different letters. For example, A1 and A2 are more similar to each other than A1 and B1.

[5] https://github.com/alisonhc/change-complexity.

[6] https://languageresearch.cambridge.org/academic-research-request-form.

[7] https://newsela.com.

has made progress on this, demonstrating strong accuracy on out-of-domain (OOD) data.

As the CEFR is a widely used international standard, readability classification into CEFR levels has been attempted (Xia et al., 2016; Khallaf and Sharoff, 2021; Arase et al., 2022). But most of this work has focused on documents, collections of documents, and individual words (Settles et al., 2020; Kerz et al., 2021; Schmalz and Brutti, 2021; Gaillat et al., 2022). There is a very limited amount of work on sentence level classification (Volodina et al., 2013; Khallaf and Sharoff, 2021; Arase et al., 2022). Arase et al. (2022) present CEFR-SP, the first human-labeled CEFR English sentence-level dataset, sourcing sentences from Newsela-Auto and Wiki-Auto (Jiang et al., 2020) in addition to the Sentence Corpus of Remedial English (SCoRE).[8] A BERT classifier trained on CEFR-SP achieves 84.5% F1 on the in-domain test set (Arase et al., 2022).

2.2 Changing Sentence Complexity

Most work in changing sentence complexity focuses on lowering sentence level to specific grades. The Newsela corpus (Xu et al., 2015; Jiang et al., 2020) has been used to train controlled simplification models to target level (Scarton and Specia, 2018; Agrawal and Carpuat, 2019; Nishihara et al., 2019; Kew and Ebling, 2022; Tani et al., 2022). To our knowledge, there have been three previous attempts at sentence complexification, also known as text or discourse embellishment. Berov and Standvoss (2018) introduce the task and train a LSTM on a story corpus and the inverse of a simplification corpus, WikiLarge, which contains aligned sentence pairs from English and Simple English Wikipedia articles (Zhang and Lapata, 2017). Naskar et al. (2019) also use WikiLarge. And more recently, Sun et al. (2023) train BART (Lewis et al., 2020) on reversed simplification sentence pairs from Newsela. There has been no previous work on same-level paraphrasing.

2.3 Sentence Simplification

Supervised Data Many sentence simplification systems adopt the architecture of machine translation, requiring complex-simple sentence pairs to train (Zhu et al., 2010; Wubben et al., 2012; Narayan and Gardent, 2014; Zhang and Lapata, 2017; Alva-Manchego et al., 2020b). WikiLarge (Zhang and Lapata, 2017), described in Section 2.2, has been widely used. Models trained on this dataset can be easily applied to test sets that source their data from Wikipedia such as ASSET (Alva-Manchego et al., 2020a) and the Turk Corpus (Xu et al., 2016). Newsela, also described in Section 2.2, has been a popular source for sentence simplification datasets (Xu et al., 2015; Zhang and Lapata, 2017). Jiang et al. (2020) present a sentence alignment model to generate the larger datasets of Wiki-Auto and Newsela-Auto. Their human annotators also developed the smaller Newsela-Manual dataset. Although most of the aforementioned corpora contain sentences that are automatically aligned, they are still considered supervised because the text was simplified by humans.

Unsupervised Data Since there are few supervised datasets, methods have been proposed to generate unsupervised datasets, which often consist of mined paraphrases. Backtranslation, or translating a sentence into a language and then back into the original language, has been used to generate paraphrases (Lu et al., 2021). Other work has used heuristics like embedding similarity to mine semantically similar sentence pairs (Martin et al., 2022). An effective way of training on unsupervised parallel data is the use of control tokens to allow models to hone in on features that correlate with sentence simplicity. For example, the ACCESS method prepends tokens that specify output length, similarity of output and input, output word rank, and output tree depth to the beginning of each input sentence (Martin et al., 2021). As these tokens are by default prepended in plain text **before tokenization**, they are functionally a form of prompt learning.

Multitask Learning Multitask learning has proven useful for overcoming lack of data and improving simplification quality. Entailment (Guo et al., 2018), paraphrase generation (Guo et al., 2018; Maddela et al., 2021), copy prediction (Maddela et al., 2021), translation (Agrawal and Carpuat, 2019; Mallinson et al., 2020), and summarization (Dmitrieva and Tiedemann, 2020) have all been used as auxiliary tasks for simplification models. It has been shown in the past that training a model on multiple very similar tasks can improve its performance on each individual task (Ratner et al., 2018; Liu et al., 2019). Although

[8] https://www.score-corpus.org.

simplification, complexification, and same-level paraphrasing belong to the same general task of changing sentence complexity, training a multi-task model with all three has not previously been attempted. The use of prompts for both training and inference has proven particularly useful for multitasking with pretrained models. Scialom et al. (2022) fine-tune a T5 model with eight new tasks, including sentence simplification, with prompts either prepended to the input text or embedded as part of a template depending on the task.

Inference with Large Language Models Research has been done on whether LLMs can simplify text without further training. Feng et al. (2023) show that GPT-3.5-Turbo produces a SARI score of 44.67 for zero-shot prompting and 47.06 for single-shot prompting, surpassing previous state-of-the-art scores. Ryan et al. (2023) find that BLOOM (Scao et al., 2023) achieves high meaning preservation and fluency but fails to simplify as well as smaller fine-tuned models. Aumiller and Gertz (2022) use an ensemble of prompts on GPT-3 (Brown et al., 2020), producing state-of-the-art results for lexical simplification specifically.

3 CEFR Level Classification

In order to automatically label paraphrase data with complexity levels, we first train a sentence-level classification model. In theory, any of the few English sentence-level readability datasets can be used for training. However, CEFR-SP (Arase et al., 2022) and Newsela (Xu et al., 2015) may contain data that we use for training and testing our later paraphrasing models, so we do not use either of those. The other option of OneStopEnglish (Vajjala and Lučić, 2018) has very few sentence pairs, and upon inspection, we find its simplest level to appear more complex than CEFR A1. Therefore, we create a new CEFR-labeled corpus for our needs, CEFR-CEP.

3.1 Data

We combine data from the English Profile and Cambridge Dictionary.[9] Our main source, English Profile (Capel, 2012), contains CEFR levels that map to word senses or grammar concepts. It contains two searchable databases, English Vocabulary Profile (EVP)[10] and English Grammar Profile (EGP).[11]

Source Distribution	EVP: 32079
	EGP: 3620
	Cambridge Dict: 3714
Level Distribution	A1: 1790
	A2: 3890
	B1: 7445
	B2: 10558
	C1: 5921
	C2: 9809
Sentence vs. Phrase	Sentence Count: 28638
	Phrase Count: 10775

Table 1: CEFR-CEP information.

Each entry in EVP corresponds to a word, and each of its possible definitions (word senses) is marked with its CEFR level along with one or more example usage sentences or phrases from either a real learner or a dictionary. EVP words, but not example sentences, have been used in the past to create lexical simplification datasets (Uchida et al., 2018; Fujinuma and Hagiwara, 2021). EGP and the Cambridge Dictionary are structured similarly to EVP, containing CEFR levels and examples for grammar concepts and word senses respectively. We automatically label these EVP, EGP, and dictionary examples with their entries' CEFR levels. We eliminate any duplicates from our combined dataset. Further details about CEFR-CEP are shown in Table 1.

This method assumes that for each word sense or grammar concept, its example sentences/phrases match its CEFR level. This is likely false some of the time. However, analysis on the CEFR-CEP sentences shows that our assumed CEFR levels correlate strongly with other metrics associated with sentence complexity: word count, tree depth, and FKGL, as shown in Figure 1.

3.2 Model

On CEFR-CEP, we train a BERT classifier (Devlin et al., 2019) in addition to SVM and LSTM

[9] https://dictionary.cambridge.org.

[10] https://www.englishprofile.org/wordlists/evp.

[11] https://www.englishprofile.org/english-grammar-profile/egp-online.

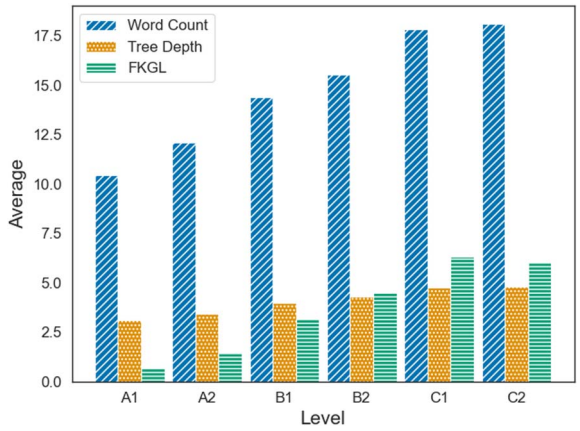

Figure 1: For texts in CEFR-CEP, the average word count, tree depth, and FKGL per CEFR level.

baselines, with an 80-10-10 train-validation-test split. The BERT-base-cased [CLS] token embedding serves as the sentence representation and the input to our classifier, which is made up of one linear layer and trained with cross-entropy loss as in previous work (Arase et al., 2022). Its outputs are softmax probabilities for each of the six CEFR levels, and we use an Adam optimizer (Kingma and Ba, 2015) with the best learning rate of 3e-5.[12]

In addition to the BERT model, we train two baselines on the same data. The first is a Support Vector Machine (SVM) classifier with Term Frequency-Inverse Document Frequency (TF-IDF) for its embeddings and a Radial Basis Function kernel (Scholkopf et al., 1997). We use the optimal cost and gamma hyperparameters of 10 and 1, respectively. We also train a LSTM classifier with a single dense layer and Word2Vec Google News vectors (Mikolov et al., 2013) as its embedding layer, Adam optimization with an optimal learning rate of 4e-3, softmax activation, and cross entropy loss.

3.3 Evaluation

We perform automatic evaluation on our held-out CEFR-CEP test data with four evaluation metrics. Our F1 scores are weighted to take label imbalance into account.

- **6-Level F1 (6-F1)**: The prediction F1 for the six CEFR levels.

[12]On a single NVIDIA GPU, we use the AllenNLP library (Gardner et al., 2018) to train for three epochs with a batch size of 32.

Model	6-F1↑	3-F1↑	Adj-Acc↑	MAE↓
SVM	57.40	71.29	80.54	0.68
LSTM	53.17	70.00	82.04	0.71
BERT	**59.78**	**76.80**	**90.64**	**0.52**

Table 2: CEFR classifier results on CEFR-CEP test set.

- **3-Level F1 (3-F1)**: The prediction F1 for the three CEFR levels A, B, and C.
- **Adjacent Accuracy (Adj-Acc)**: the percentage where the prediction's deviation from the test label is less than or equal to one.[13]
- **Mean Absolute Error (MAE)**: a number between 0 and 5. The average amount that the prediction deviates from the test label.[14]

Table 2 shows the results for each metric on the baseline and BERT models. For every metric, the BERT model performs better. But 6-F1 is only 59.78%, and we posit that it is so difficult to get an exact match with dataset CEFR level because of dataset flaws mentioned in Section 3.1: namely, that we label each example text according to the level of its corresponding word sense or grammar concept, which is not always correct. But Adj-Acc is a high value of 90.64%, showing that our model has very close estimation, and the low MAE of 0.52 is consistent with this. Our SVM baseline scores similarly to the LSTM despite having much more information-rich embeddings.

Since we will use our classifier to add CEFR labels to the OOD ParaNMT dataset, we conduct a study to see to what extent its labels match human labels on the ParaNMT data. On our preprocessed ParaNMT set (see Section 4.2 for details), we sample 60 sentence pairs: 20 where their classified levels are the same and 40 where their classified levels differ by at least two (e.g., A2-B2 but not A2-B1). We split the different-level pairs into two groups: simplification where the higher level sentence comes first and complexification where the lower level one does. We then ask four native English speakers to examine each sentence pair

[13]Under this metric, a prediction of A2 would be considered accurate if the test label was A1, A2, or B1, because the deviation from A2 is one or less.

[14]Prediction 0 (A1), test label 1 (A2) corresponds to MAE of 1. Prediction 1, test label 5 (C2) corresponds to MAE of 4.

Category	CEFR F1	FKGL F1
Simplification	**53.33**	46.15
Complexification	50.0	**64.52**
Same Level	12.50	**28.57**

Table 3: F1 of CEFR classifier vs. FKGL predictions on 39 human labels.

and label which sentence is simpler: the first, the second, or neither. **These three labels map to the categories of complexification, simplification, and same-level paraphrasing, respectively.**

Inter-rater agreement, or nominal Krippendorff's Alpha (Krippendorff, 2011), is a fairly low 0.27, where 0 means no agreement (chance) and 1 means perfect agreement. Because we want to evaluate on only reliable labels, we just consider the sentence pairs where three or more of the raters agree. These amount to 39 out of 60 pairs with agreement of 0.48. We test both our CEFR classifier and FKGL on these 39 gold labels.

We compare our CEFR classifier's predictions with those of FKGL. Table 3 shows the F1 of the CEFR versus FKGL methods on the gold labels for each of the three categories of simplification, complexification, and same-level paraphrasing. FKGL performs better for classifying complexification and same-level paraphrasing, while CEFR classification performs better for simplification. However, F1 is universally low, casting doubt on the reliability of our weak labeling approaches. Our gold human labels are also potentially problematic: Only six of the 60 sentence pairs that were rated as same-level paraphrasing met our criterion of three out of four raters agreeing, compared to 15 and 18 for simplification and complexification respectively. From these results, we tentatively hypothesize that sentence simplification models trained on data labeled by the CEFR classifier will perform better than those trained on FKGL-labeled data, while complexification and same-level paraphrasing models trained FKGL-labeled data will perform better than those trained on CEFR-labeled data.

4 Paraphrasing Data

Next, we construct datasets for simplification, complexification, and same-level paraphrasing. Details are included in Table 4.

Dataset	Tasks	Size
Newsela-Auto	Simplification	238,597
	Complexification	238,662
ParaNMT-CEFR	Simplification	1,287,794
	Complexification	1,287,795
	Same Level	1,287,795
ParaNMT-FKGL	Simplification	1,287,794
	Complexification	1,287,794
	Same Level	1,287,794

Table 4: Paraphrasing dataset details.

4.1 Supervised Data

Our supervised data source is Newsela-Auto,[15] a sentence simplification corpus derived from Newsela news articles targeted at five levels and written by education professionals, where level 0 is the complex original and 1–4 are simplifications of increasing degree (Xu et al., 2015). Their sentences must be aligned to create a sentence pair corpus from these original articles. Previous methods have aligned using metrics like Jaccard similarity (Zhang and Lapata, 2017). Newsela-Auto's pairs are aligned according to a neural CRF model (Jiang et al., 2020), and its pairs are more numerous (666k) and creatively rewritten than previous Newsela alignments.[16] Newsela-Auto does not contain level labels, so we use string matching with the original Newsela to find each sentence's level (Xu et al., 2015).[17]

A limitation of Newsela-Auto and other simplification datasets like WikiLarge (Zhang and Lapata, 2017) and Wiki-Auto (Jiang et al., 2020) is that they are only meant to contain different-level pairs. Therefore, we only conduct simplification and complexification experiments on this dataset. For the two-task dataset, we flip the order of exactly half of the sentence pairs. For the two single-task datasets, we extract all simplification and complexification pairs from the two-task dataset but perform an additional filtering step of removing all pairs that were labeled as the same level according to our retroactive labeling

[15]Request data at `https://newsela.com/data`.

[16]To stay consistent with previous work, we employ the same train-test-validation split.

[17]Due to the limitations of this retroactive approach, our resulting corpus is slightly smaller than the original: 394,108 instead of 394,300 for training and 43,305 instead of 43,317 for validation.

algorithm. These pairs only number into a few thousand and are not enough to train a comparable same-level paraphrasing model.

4.2 Unsupervised Data

To contrast with our supervised dataset and fill the gap of missing same-level paraphrase pairs, we create two unsupervised datasets. We use ParaNMT, one of the largest paraphrase pair datasets available to the public, with 50 million sentence pairs generated through backtranslation of the Czeng1.6 corpus (Wieting and Gimpel, 2018). It contains data sourced from movie and educational video subtitles, European legislation proceedings, and medical websites (Bojar et al., 2016). ParaNMT has been used for sentence simplification in the past (Martin et al., 2022).

To determine our filtering techniques, we inspect samples from the corpus and find pairs that are identical or almost identical, very different in meaning, or that contain incomplete sentences. To alleviate these problems, we remove pairs where one sentence is contained in the other or where any sentence has less than three words.

To encourage our models not to directly copy the input sentence—a problem that occurs in both sentence simplification (Dong et al., 2019) and paraphrase generation (Thompson and Post, 2020)—we only include aggressive paraphrases. We remove pairs where Sentence-BERT cosine similarity (Reimers and Gurevych, 2019) is below 60% or above 80%. From our observations, these thresholds exclude pairs that are different in meaning or too similarly phrased.

We want ParaNMT-CEFR and and ParaNMT-FKGL to be as similar as possible for the sake of comparison. From our filtered data, we use the CEFR classifier to label the level of each sentence. To maximize the likelihood that a level difference between the two sentences exists (see Table 2's Adj-Acc), we only select pairs where the level difference is two or greater.[18] For the same-level dataset, we select pairs where the sentences are classified as exactly the same level.[19]

We are left with 2,575,589 different-level pairs and 6,207,876 same-level pairs. For both the CEFR-based and FKGL-based labeling schemes, **we derive all of our simplification, complexification, and same-level paraphrasing data from these two sets**. For ParaNMT-CEFR, we halve the different-level dataset and re-order it to create one simplification and one complexification dataset. We then sample from the same-level pairs to get an equal-sized same-level set. To create ParaNMT-FKGL, we calculate the FKGL of each sentence (rounded to two decimal points). If the FKGL of the two sentences in a pair differs at all, we consider it a different-level pair. If it is **exactly the same**, we consider it a same-level pair. We are able to derive 65.16% of our different-level pairs from the ParaNMT-CEFR different-level set. The other 878,449 are taken from the ParaNMT-CEFR same-level pairs. We sample from the resulting data to match ParaNMT-CEFR's in size. The train-validation-test split is 80-10-10 for both ParaNMT datasets. We have made these data available to the public.[20]

[18]For example, we keep A1-B1 pairs but remove A2-B1 pairs.

[19]For example, A1-A1 but not A1-A2.

Task	Prompt(s)
Simplification	''level down: '', ''change to level X: ''
Complexification	''level up: '', ''change to level X: ''
Same-level	''same level: ''

Table 5: Prompt(s) for each task. For same-level paraphrasing single-task models, we only train REL prompt ablations. For simplification, complexification, and all two-task and three-task configurations, both REL and ABS prompt ablations are trained.

5 Paraphrasing Experiments

We train models on the three tasks of sentence simplification, sentence complexification, and same-level paraphrasing. We train ablations for training dataset (Newsela-Auto, ParaNMT-CEFR, ParaNMT-FKGL), multitasking configuration (1–3 tasks), and prompting strategy (relative/absolute). Including our baselines, we train 42 models in total. See Table 5 for details.

5.1 Models

For all models, we use a single NVIDIA GPU, a batch size of 32 after gradient accumulation, and maximum decoding length of 300 tokens. We fine-tune 34 ablations on **T5** (Raffel et al., 2020),

[20]https://github.com/alisonhc/change-complexity.

a pre-trained transformer.[21] We also perform limited experiments with Flan-T5-base (Chung et al., 2022), a more recent instruction-tuned version of T5. We train ParaNMT-CEFR single-task and 2-task simplification and complexification ablations (6 models). However, since we find in Section 6.1.4 that it does not perform as well as T5, we focus our main experiments on T5.

5.2 Prompting Strategies

At inference time, we prepend the corresponding prompt to the beginning of each input sentence, as this strategy was used for T5 (Raffel et al., 2020).

Relative Simplification, complexification, and same-level paraphrasing correspond exactly to the prompts *''level down: ''*, *''level up: ''*, and *''same level: ''*. We train on the data of one, two, or all three tasks, adding the corresponding task prompt to the front of each input sentence. We call this relative (REL) prompting because the prompt denotes the relative difference between the levels of the input and output sentence: down, up, or same. This scheme has 7 possible task combinations.

Absolute For each task combination besides single-task same-level paraphrasing, we use prompts that specify absolute (ABS) output level. For training, we insert *''change to level X: ''*, where X is the level of the output.[22] ABS prompting theoretically has an advantage over REL prompting because we can change the prompt to match the level of a test dataset's output sentence. To compare the two prompting strategies on equal footing, we remove this advantage. With the exception of Section 6.1.6, for ABS prompting inference, **we use the same prompt for every test input no matter the output level**. Therefore, we can only evaluate these models on simplification and complexification and not on same-level paraphrasing.

5.3 Baselines

We train paraphrasing baselines, the first trained on the entire ParaNMT-CEFR dataset and the other trained on ParaNMT-FKGL. Each dataset consists of one third simplification data, one third complexification data, and one third same-level paraphrasing data, but at train time, we use the prompt *''paraphrase: ''* for each input.[23]

6 Paraphrasing Evaluation

We perform both automatic and human evaluation. To compare all 40 experiment models, we only report automatic evaluation results. We perform human evaluation on just one model per task.

6.1 Automatic Evaluation

We first discuss each individual task. Then, we discuss our ablation results more generally.

Metrics We report SARI and FKGL.[24]

- **SARI** (System output Against References and against the Input sentence) is the most important automatic metric for text simplification. Ranging from zero to 100, it represents the F1 for a model's added, kept, and deleted n-grams when comparing the input and reference sentences (Xu et al., 2016).
- **FKGL** (Flesch–Kincaid Grade Level) is a weighted score with sentence length and syllable information (Kincaid et al., 1975). It was introduced in Section 1. We consider the best FKGL score to be that closest to the gold reference FKGL in a given test set.

Data For simplification and complexification, we use ASSET and Newsela-Manual. These simplification benchmarks can be easily reversed for the complexification task. There are no existing benchmarks that can be applied to same-level paraphrasing. Therefore, we use sentence pairs from the ParaNMT corpus. In all tables and figures, we denote task type to u/d/s for up (complexification), down (simplification), and same.

- **ASSET** has 359 test sentences, each with 10 human-written reference sentences (Alva-Manchego et al., 2020a). For simplification,

[21] We fine-tune T5-base with the transformers library (Wolf et al., 2020). After 3 epochs, we automatically select the model checkpoint with the lowest validation loss.

[22] For ParaNMT-CEFR, X is the CEFR level A/B/C. For ParaNMT-FKGL, X is FKGL rounded to two decimal points. And for Newsela-Auto, X is one of the Newsela levels 0–4.

[23] We also train an LSTM baseline per task per ParaNMT dataset using REL prompting, but we do not report the results because they do not add to the analysis.

[24] We use the EASSE Python library to compare with previous sentence simplification research (Alva-Manchego et al., 2019).

we use this dataset as-is. For complexification, we consider each reference sentence to be an input and the corresponding test sentence to be an output, resulting in 3590 one-to-one pairings.

- **Newsela-Manual** contains Newsela sentence pairs where each pair is annotated as *aligned*, *partially aligned*, or *not aligned* (Jiang et al., 2020).[25] We collect all *aligned* and *partially aligned* pairs and follow Kew and Ebling's (2022) method to automatically fix the alignments between *partially aligned* pairs. We include pairs from all input levels to all output levels and remove pairs where the output is an exact copy of the input, resulting in 2,748 pairs.[26]
- **Newsela-Manual by Level** contains sentences where the complex level 0 maps to each of the simple levels 1–4. To evaluate our models' level targeting ability, we use the same configuration as Kew and Ebling (2022).[27] We also create a complexification version with the simple input of level 4 and the possible output levels 3-0.
- **ParaNMT-s** Since there is no publicly available same-level paraphrasing dataset, we sample from both the FKGL and CEFR versions of the ParaNMT-same set to collect 128,779 pairs.[28] This corpus is inherently noisy due to its unsupervised nature. We hope that in future work, a cleaner same-level paraphrasing dataset with human labels will be available.

6.1.1 Simplification Results

We report results in Table 6 on both the ASSET and Newsela-Manual test sets. Besides baselines, we divide the table into two sections, one for models trained on unsupervised data and the other for supervised data. We only report our two best performing ablations per training dataset. For ABS prompting CEFR and Newsela-Auto (News) models, we try all possible prompts. For FKGL models, we try a range of prompts (0.0-7.0) and pick the best ones. For MUSS models, which are open source (Martin et al., 2022), we report their best scores on ASSET and do our own parameter search on the Newsela-Manual validation set to derive optimal prompts. On both benchmarks, all models outperform baselines in SARI score. We achieve a new state-of-the-art for unsupervised parallel data, with the highest SARI score of 43.65 on ASSET going to T5-CEFR-u-d-ABS (prompt B).[29] Our supervised model T5-News-u-d-REL has the highest SARI score on the Newsela-Manual benchmark, outperforming baselines and MUSS.

	ASSET		Newsela-Manual	
Model	**SARI↑**	**FKGL**	**SARI↑**	**FKGL**
Baselines				
Reference	44.89	6.49	–	5.80
T5-CEFR-Para	39.58	9.88	36.13	9.73
T5-FKGL-Para	39.45	9.90	36.0	9.69
Unsupervised Data				
MUSS-mined	42.65	8.23	38.80	7.26
Lu et al. 2021	42.69	7.94	–	–
T5-CEFR-u-d-ABS *(B)*	**43.65**	7.91	39.13	8.09
T5-CEFR-d-s-ABS *(B)*	43.45	8.51	**39.67**	8.24
T5-FKGL-d-ABS *(see caption)*	42.38	**7.03**	37.81	2.47
T5-FKGL-d-s-ABS *(3.0)*	42.31	6.81	39.29	**5.90**
Supervised Data				
MUSS-wiki-mined	**44.15**	**6.05**	41.38	6.67
Clive et al. 2022	43.58	5.97	–	–
T5-News-d-ABS *(4)*	40.87	5.96	41.54	**5.76**
T5-News-u-d-REL	39.97	5.92	**42.44**	5.91

Table 6: **Simplification** on ASSET and Newsela-Manual. Models abbreviated to [Model]-[Data]-[Tasks]-[ABS or REL prompting]. MUSS-mined and MUSS-wiki-mined come from Martin et al. (2022). MUSS and Clive et al. (2022) use ACCESS prompting (Martin et al., 2021). Lu et al. (2021) create their own corpus via backtranslation. For ABS models, we enclose in parentheses the target level we used for prompting at inference time. T5-FKGL-d-ABS uses 3.0 for ASSET and 0.0 for Newsela-Manual.

6.1.2 Complexification Results

We report SARI and FKGL on reversed ASSET and reversed Newsela-Manual. Table 7 contains results arranged in the same way as for simplification. For FKGL prompts, we try a range of

[25]There is no overlap between Newsela-Auto training or validation data and Newsela-Manual test data.

[26]For simplification, we use the dataset as-is, and for complexification, we reverse it.

[27]This configuration does not filter out input-output copies.

[28]There is no overlap between our resulting test set and either of the training or validation sets.

[29]We say unsupervised **parallel** data because GPT-3.5-Turbo, mentioned in Section 2.3, has a higher score (Feng et al., 2023).

	ASSET		Newsela-Manual	
Model	**SARI↑**	**FKGL**	**SARI↑**	**FKGL**
Baselines				
Reference	–	10.46	–	10.14
T5-CEFR-Para	42.09	7.46	39.41	6.92
T5-FKGL-Para	42.28	7.40	39.83	6.93
Unsupervised Data				
MUSS-mined	44.06	7.92	38.46	7.85
T5-CEFR-u-ABS *(C)*	43.87	7.79	**40.98**	7.61
T5-CEFR-u-s-ABS *(C)*	43.44	7.70	39.60	7.50
T5-FKGL-u-ABS *(12.0)*	43.86	13.76	40.21	12.36
T5-FKGL-u-s-ABS *(11.0)*	**44.07**	**11.87**	40.32	**11.10**
Supervised Data				
MUSS-wiki-mined	**42.51**	7.89	37.97	7.40
Sun et al. (2023)	40.0	8.30	–	–
T5-News-u-ABS *(0)*	38.96	**9.82**	**42.21**	**9.46**
T5-News-u-REL	36.90	8.10	42.07	7.64

Table 7: **Complexification** on ASSET and Newsela Manual. See Table 6's caption for naming details. We obtained model weights and data for Sun et al.'s (2023) ComplexBART model and ran inference ourselves. However, since their Newsela training data overlaps with the Newsela-Manual test set, we only report ASSET scores for ComplexBART.

10.0–17.0. We do a grid search to find MUSS parameters. MUSS-mined almost matches our best performing model's SARI on ASSET, even beating its supervised data counterpart and Sun et al.'s (2023) ComplexBART. But it falls much shorter of our best models on Newsela-Manual.

Between ParaNMT-CEFR and ParaNMT-FKGL models, the latter produce the highest SARI on ASSET and highest FKGL on both test sets. However, after inspecting model outputs, we find that for every FKGL model whose SARI surpasses our highest ParaNMT-CEFR SARI score, the outputs contain many degenerate repetitions. For example, consider the ASSET input simple sentence **The state capital is Aracaju.** T5-CEFR-u-s-ABS with prompt C produces the slightly longer sentence **the capital of the state is Aracaju.** But T5-FKGL-u-s-ABS with prompt 11.0 produces a 295-word output starting with **the capital of the state is Aracaju, the capital of the state is the capital of the state of the state of**. MUSS SARI also surpasses ParaNMT-CEFR on ASSET. However, their outputs contain fewer degenerate repetitions according to an inspection

Model	**SARI↑**	**FKGL**
Baselines		
Reference	–	2.82
T5-CEFR-Para	**49.40**	2.76
T5-FKGL-Para	48.21	**2.82**
Experiment Models		
T5-CEFR-u-d-s-REL	48.26	2.86
T5-FKGL-u-d-s-REL	45.75	2.90

Table 8: **Same-level paraphrasing** on ParaNMT-s. See Table 6's caption for naming details.

of the outputs. We believe this quality difference is due to problems with the ParaNMT dataset that are exacerbated by organizing it by FKGL score, a length-based metric. The MUSS-mined training data contains human-written sentences that were mined according to similarity metrics (Martin et al., 2022). ParaNMT, on the other hand, is the result of machine translation (Wieting and Gimpel, 2018), which can sometimes enter repetitive loops during decoding (Holtzman et al., 2019; Welleck et al., 2019). Future work on backtranslation datasets could attempt to filter out sentences that contain these repetitions.

We also find that degenerate repetitions are not adequately captured by SARI, which only counts **unique *n*-grams** that are added, kept, and deleted compared to the gold references (Xu et al., 2016; Alva-Manchego et al., 2019). This means that **as long as a model's repetitions have added no or very few unique new words to the sentence, they will not be reflected in SARI**. Therefore, we suggest that for sentence complexification, a modified SARI should be used that takes word counts into consideration. We leave this to future work.

6.1.3 Same-level Paraphrasing Results

In Table 8, we report results for all of our baselines along with our best performing CEFR and FKGL models. Notably, both of our CEFR and FKGL paraphrasing *baselines* outperform their corresponding experiment models, which were trained on the **exact same data**, the only difference being prompting strategy. When we compare T5-CEFR-Para's outputs with those of

	Simplification (d)		Complexification (u)	
Model	**ASSET**	**News**	**ASSET**	**News**
Best T5-CEFR	**43.65**	**39.67**	**43.87**	**40.98**
CEFR-ABS-d *(B)*	**42.91**	**39.28**	–	–
CEFR-ABS-u *(see caption)*	–	–	**42.84**	**40.57**
CEFR-ABS-u-d *(d-B, u-C)*	42.45	38.81	42.33	39.46
CEFR-REL-d	42.46	38.75	–	–
CEFR-REL-u	–	–	42.73	40.55
CEFR-REL-u-d	42.64	38.79	42.12	39.66

Table 9: Flan-T5 SARI for all trained ablations. For ABS models, the best prompt(s) are shown in parenthesis. CEFR-ABS-u uses B for ASSET and C for News.

T5-CEFR-u-d-s-REL, we find that after tokenization, the former copies the input 4.40% of the time, while the latter does so 10.42% of the time. The ParaNMT-s test set copies input 0.31% of the time after tokenization. Since we are unable to perform a quantitative human evaluation comparing these outputs, we are left with two possible theories.

The first is that our T5 paraphrasing baselines are actually learning to same-level paraphrase. When presented with data where a third increases level, a third decreases level, and a third keeps level the same, the model picks the average option, which is same-level paraphrasing. The second theory is that the sentences in our same-level paraphrasing data are not actually the same level. After all, both our CEFR and FKGL methods in Section 3.3 have extremely low F1 on human labels for same-level paraphrasing: 12.5% for CEFR and 28.57% for FKGL. However, we doubt this theory because of our positive human evaluation results (see Section 6.2).

6.1.4 Flan-T5 Results

Table 9 shows SARI for the six Flan-T5 ablations we trained along with the best SARI scores from our T5 experiments on the same dataset of ParaNMT-CEFR. Interestingly, the best Flan-T5 scores never surpass the best from T5. And when directly comparing scores for each ablation, **T5 outperforms Flan-T5 for 12 out of the 16 cases**.

This may be surprising, as Flan-T5 performs better on a variety of tasks and benchmarks for zero- and few-shot inference (Chung et al., 2022). But Flan-T5 has **not** been shown to be better than T5 for fine-tuning on new datasets. We suspect that the reason for its degraded performance compared to T5 is that fine-tuning incurs catastrophic forgetting, diminishing the benefits gained from its previous instruction-tuning. While Scialom et al. (2022) report that T5 models can continually learn new tasks without catastrophic forgetting, rehearsal (Shin et al., 2017) is still required for the models to retain their previously learned skills.

6.1.5 Ablation Study Results

Figure 2 shows results for all T5 experiment models on all test sets, the x-axis being number of tasks per model and the y-axis being SARI score. Each data point is annotated with task combination.

Multitasking There is no clear winner among multitasking configurations. Single- and two-task models often perform better than three-task ones, with the exception of same-level models, where SARI increases with the number of tasks. Many high-scoring two-task models were trained on tasks that are not opposite (i.e., u-s and d-s but not u-d). However, for simplification, the highest scoring models for ASSET and Newsela-Manual were both trained on the u-d ablation. For T5-News-u-d-REL, this is not noteworthy because REL prompts are distinct for each task (see Table 5). But strikingly, T5-CEFR-u-d-ABS scores best on ASSET with prompt **B** even though in theory, upon seeing the middle prompt B (as opposed to A or C), the model should not know whether to increase or decrease a sentence's complexity. Upon further investigation, we find that the reason for this is likely that the training dataset contains approximately double the amount of $C \rightarrow B$ simplifications as $A \rightarrow B$ complexifications.

Prompt Type For FKGL models, ABS prompting always performs better than REL prompting. For News models, ABS prompting performs better in all but one case. For CEFR models, results are mixed, but ABS prompting performs slightly better on average. Compared to CEFR and Newsela levels, FKGL is very fine-grained, with up to two decimal point precision. The fact that FKGL models always perform better for ABS prompting than for REL, while CEFR and News models do not, suggests that using prompts that contain very fine-grained output information might improve performance. Additionally, among

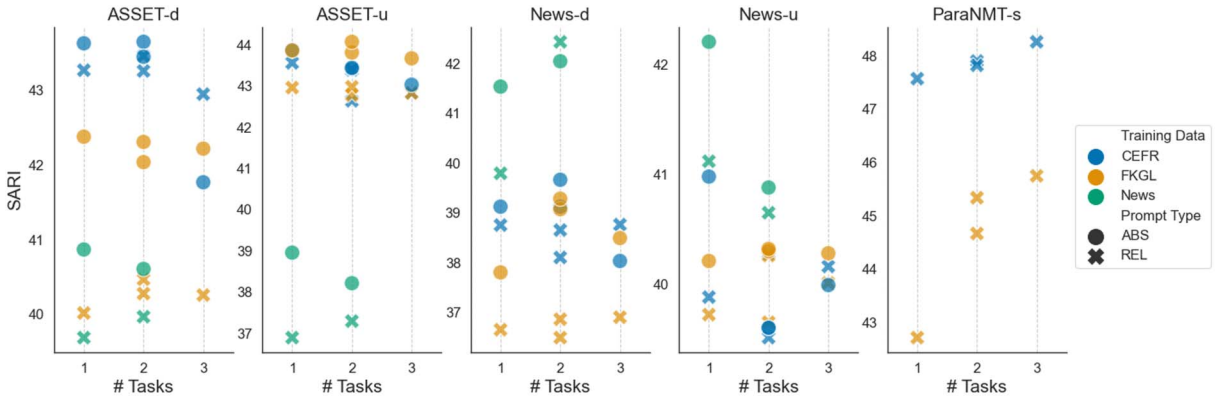

Figure 2: All ablation results. Tasks abbreviated as u (up, complexification), d (down, simplification), and s (same, same-level paraphrasing). ASSET-d and News-d correspond to the original ASSET and Newsela-Manual sets. The -u indicates that they were reversed for complexification.

just single-task models, ABS prompting always performs best, but this strategy is favored less and less as the number of tasks increases. This indicates that using a more complex prompting strategy incurs a greater performance cost as the number of tasks increases.

Data Labeling Scheme As expected, models trained on Newsela-Auto perform better on Newsela-Manual than models trained on ParaNMT data. However, they mostly fail to achieve as high of SARI on non-Newsela data as ParaNMT models achieve on Newsela data, and they are some of the worst performing models on ASSET. For ABS prompting, FKGL models often outperform CEFR models on complexification, but for REL prompting, FKGL models almost universally do worse. For same-level paraphrasing, it is notable that ParaNMT-CEFR models have much higher SARI than ParaNMT-FKGL ones despite the fact that the ParaNMT-s test dataset is half ParaNMT-CEFR and half ParaNMT-FKGL. This, and the fact that complexification FKGL model outputs contain degenerate repetitions that SARI does not reflect, shows that the CEFR method is the most robust automatic labeling method. Future work could experiment with finer-grained CEFR labels (6, not 3) and fewer fine-grained FKGL labels (intervals instead of two decimal precision).

6.1.6 Level Targeting Results

Table 10 shows our Newsela-Auto models' abilities to target specific levels for simplification and complexification. For brevity, we show results from only one of our models per table along with the best previous work baseline, supervised MUSS (Martin et al., 2022), for which we derive optimal parameters via grid search. For every level, our models achieve higher SARI than previous work, with the exception of $0 \rightarrow 2$ simplification, where MUSS wins. However, it appears that our models are better at targeting aggressive simplifications and complexifications than slight ones: SARI generally increases as target level deviates further from input level. The results from Section 6.1.5 show that even when we are not using ABS prompting to its full strength, it

Simplification			Complexification		
Target Level	**SARI↑**	**FKGL**	**Target Level**	**SARI↑**	**FKGL**
$0 \rightarrow 1$	–	9.05	$4 \rightarrow 3$	–	5.46
MUSS	38.71	7.34	MUSS	35.14	6.24
Ours	**39.81**	**10.30**	Ours	**41.82**	**4.90**
$0 \rightarrow 2$	–	7.13	$4 \rightarrow 2$	–	7.05
MUSS	**42.37**	**7.06**	MUSS	37.25	**6.22**
Ours	41.81	7.82	Ours	**40.97**	5.90
$0 \rightarrow 3$	–	5.51	$4 \rightarrow 1$	–	9.06
MUSS	40.21	**4.88**	MUSS	37.19	6.10
Ours	**44.81**	6.31	Ours	**41.52**	**6.85**
$0 \rightarrow 4$	–	3.89	$4 \rightarrow 0$	–	11.46
MUSS	40.08	**4.64**	MUSS	34.53	5.80
Ours	**46.77**	4.83	Ours	**42.44**	**8.33**

Table 10: Level targeting for **simplification** and **complexification** on Newsela-Manual. We compare our scores to supervised MUSS (Martin et al., 2022). Our simplification and complexification models are T5-News-u-d-ABS and T5-News-u-ABS respectively. For each level, we display reference FKGL. See Table 6 for naming conventions.

often surpasses REL prompting in performance. These level-targeting results confirm that ABS prompting at its full strength does better.

6.2 Human Evaluation

We carry out a human evaluation on all three tasks. We use a 1-5 Likert scale across three separate categories: task performance, meaning preservation, and fluency. Due to limited resources, we choose just one model per task. We choose ParaNMT models for our evaluation. For simplification, T5-CEFR-u-d-ABS with prompt B scores best on ASSET, but due to the prompt B task ambiguity discussed in Section 6.1.5, we choose T5-CEFR-d-ABS with prompt B, which scores second best with a SARI of **43.63**. For complexification, we use the highest scoring CEFR model, T5-CEFR-u-ABS with prompt C, even though some of the FKGL models have higher SARI scores on ASSET. This is because, as mentioned in Section 6.1, FKGL models produce numerous degenerate repetitions that do not hurt SARI score. Finally, for same-level paraphrasing, we choose T5-CEFR-u-d-s-REL because of its highest SARI score on ParaNMT-s.

Due to limited human evaluation resources, out of the three tasks, we only compare our simplification model to a baseline. We choose supervised MUSS (Martin et al., 2022), a publicly available state-of-the-art model that we also used in Section 6.1. We use its best-performing ASSET prompts. So as to directly compare the three tasks of simplification, complexification, and same-level paraphrasing on the exact same dataset, something not done in Section 6.1, we do not use a benchmark simplification dataset. We instead source data from the CEFR-CEP test set, which our paraphrasing models have not seen and our CEFR classifier has not been trained or validated on. However, because of this choice, there are no reference paraphrases to compare model outputs to, preventing us from using a reference baseline. We do not use any baseline because in the absence of a single one that fits all three tasks, it would require dramatically more labeling work.

From CEFR-CEP, we sample 13 sentences from each level A2-C1, amounting to 52 sentences that we release to the public.[30] We exclude A1 and C2 because simplifying or complexifying those sentences may not have an effect. We then run each of the four models on these sentences, producing 208 outputs. Three native English speakers each rate all outputs.[31] For each output, we average the ratings of the three evaluators. We then take the 95% confidence interval across each model's rating category along with inter-rater agreement using ordinal Krippendorff's Alpha (Krippendorff, 2011), a number between zero (random agreement) and one (perfect agreement).

	Task	Meaning	Fluency
Simplification			
MUSS	$2.96_{\pm 0.23}$	$3.63_{\pm 0.34}$	$4.71_{\pm 0.15}$
Agreement	0.33	0.63	**0.28**
Ours	$\mathbf{3.04}_{\pm 0.26}$	$\mathbf{4.24}_{\pm 0.27}$	$\mathbf{4.74}_{\pm 0.14}$
Agreement	**0.44**	0.60	0.26
Complexification	$2.35_{\pm 0.23}$	$4.12_{\pm 0.33}$	$4.64_{\pm 0.14}$
Agreement	0.28	**0.77**	0.18
Same Level	$\mathbf{3.85}_{\pm 0.18}$	$\mathbf{4.72}_{\pm 0.15}$	$\mathbf{4.77}_{\pm 0.11}$
Agreement	0.01	0.52	0.16

Table 11: Human evaluation results. Each row contains a mean rating from 1 to 5 with a confidence interval, plus inter-rater agreement below it.

Table 11 shows our results. For simplification, our model performs better than MUSS across all categories, especially meaning preservation. Across tasks, fluency is universally very high. This is a testament to the quality of these fine-tuned language models. Agreement is highest for meaning preservation, perhaps the most objective metric. We find that task performance is lowest for complexification, which is consistent with our intuition that this is the most difficult task, demanding the most additions and leaving the most room for error. Finally, same-level paraphrasing has the highest scores out of 5 compared to the other tasks, likely because it requires the least amount of modification. This is particularly interesting because of the fact that our paraphrasing baseline T5-CEFR-Para outperformed this model according to SARI on ParaNMT-s, calling into question whether the task models were effective at all. We told our raters to dock task performance points when a model exactly copied its input, but upon

[30] https://github.com/alisonhc/change-complexity.

[31] The raters are not told which outputs are from our models and which are not.

inspection of their ratings, we find that this is very inconsistent. So, this may be why inter-rater agreement is extremely low for task performance.

7 Can LLMs Change Complexity Level?

In this section, we perform an exploratory investigation into the simplification, complexification, and same-level paraphrasing abilities of LLMs.

7.1 Experiments

7.1.1 Data

For simplification and complexification, we use ASSET like in Section 6.1.[32] For same-level paraphrasing, we randomly sample 400 sentence pairs from ParaNMT-s.[33]

7.1.2 Models

For all models, we set temperature to 1.0 and limit output length to 50 tokens. We run inference in a zero-shot setting and leave an investigation into more sophisticated inference settings to future work. Due to hardware limitations, we are unable to run inference for models with more than 20 billion parameters. We mostly select instruction-tuned models because we expect them to do better with new tasks and prompts. We select five: GPT-3.5-Turbo,[34] GPT-NeoX-20B (Black et al., 2022), Flan-UL2 (Tay et al., 2023), Flan-T5-xxl (Chung et al., 2022), and OPT-IML-MAX-1.3B (Iyer et al., 2023).

7.1.3 Prompts

As in our fine-tuning experiments, we attempt both ABS and REL prompting. However, in this case, we construct prompts with more descriptive wording to better fit the zero-shot setting. Table 12 shows the prompts for each task. To determine them, we try different wording with GPT-3.5-Turbo to check for obvious differences in behavior. We find that for complexification, explicitly telling the model to ‘‘increase the complexity’’ of a piece of text produces undesirably long outputs, but the wording ‘‘advanced English level’’ does not. We keep terminology consistent across prompts.

Task	REL Prompt	ABS Prompt
Simplification or Complexification	‘‘Please rewrite the following text to a [less/more] advanced English level: ’’	‘‘Please rewrite the following text so that its [CEFR/FKGL] level is X: ’’
Same-level	‘‘Please rewrite the following text to the same English level: ’’	–

Table 12: Prompt(s) for each task. For CEFR ABS prompting, we use A for simplification and C for complexification. For FKGL ABS prompting, in two-point intervals, we try levels 0–6 for simplification and 8–14 for complexification.

	d		u		s	
Model	**SARI↑**	**FKGL**	**SARI↑**	**FKGL**	**SARI↑**	**FKGL**
Best Fine-tuned	43.65	**7.03**	**44.07**	9.82	**48.26**	**2.86**
GPT-3.5-Turbo *(d-A, u-8)*	**45.76**	8.28	**42.84**	**10.72**	**41.73**	4.98
GPT-NeoX-20B *(d-2, u-REL)*	35.85	5.77	34.78	3.89	34.52	2.43
Flan-UL2 *(d-4, u-10)*	32.50	4.91	34.58	5.51	21.85	**2.73**
Flan-T5-xxl *(d-A, u-10)*	28.99	1.47	30.25	6.79	20.79	2.63
OPT-IML-MAX-1.3B *(d-0, u-8)*	36.26	**6.01**	33.52	3.98	31.07	0.0

Table 13: LLM results based on best SARI per model, tested on ASSET. Tasks are simplification (d), complexification (u), and same-level paraphrasing (s). For u and d, best prompts are included in the Model column. Reference FKGL is 6.49 for d, 10.46 for u, and 2.82 for s.

7.2 Results and Discussion

Table 13 shows results for each LLM and task, and Figure 3 shows SARI for each LLM per task and prompt type. On all tasks, GPT-3.5-Turbo outperforms the rest of the models by a large margin. None of the other models produce SARI scores that come close to the paraphrasing baselines from Tables 6, 7, and 8, much less the fine-tuned T5 scores. We confirm this by inspecting model outputs: all besides GPT-3.5-Turbo contain hallucinations. For example, in response to CEFR prompting (and FKGL to a lesser degree), Flan-T5-xxl and Flan-UL2 often return a single letter instead of a sentence as the output, while OPT-IML-MAX-1.3B and GPT-NeoX-20B attach discussions of the CEFR to their outputs.

[32] We do not use Newsela-Manual because we were not able to obtain clarity on whether sending data through OpenAI's API violates Newsela's licensing agreement.

[33] Cutting down on the original 128,779 pairs reduces both API costs and inference time.

[34] https://platform.openai.com/docs/model-index-for-researchers.

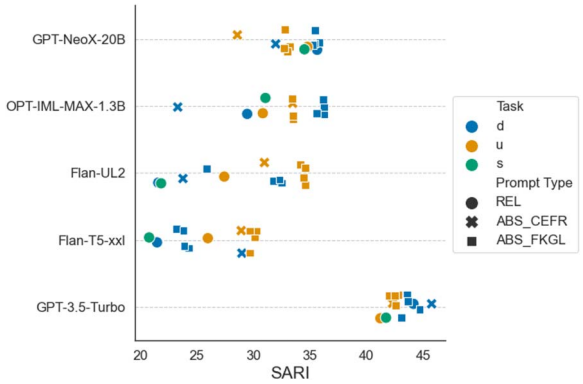

Figure 3: All SARI scores per model, task, and prompt.

Despite the fact that the ABS prompting outputs contain more hallucinations than those from REL prompting, Figure 3 shows that ABS prompting generally produces higher SARI, echoing our findings from the fine-tuning experiments. For GPT-3.5-Turbo in particular, the ABS-CEFR prompt produces outputs with **higher** SARI for simplification than Feng et al.'s (2023) REL prompting score of **44.67** in the zero-shot setting.

Notably, although GPT-3.5-Turbo outperforms our fine-tuned models on simplification, it does not on complexification, demonstrating the difficulty of the task. Models perform the worst at same-level paraphrasing, but this may be due to the unsupervised same-level dataset being worse in quality than supervised ASSET.

The huge gap in performance between GPT-3.5-Turbo and the other models may be in part due to its size of 176B parameters being much larger than the next largest size of 20B. However, there is no obvious pattern regarding model size for the other four: For example, the smallest model of OPT-IML-MAX-1.3B performs competitively with the two 20B-parameter models.

8 Conclusion

In this paper, we provide a general investigation of the task of changing sentence complexity, with thorough fine-tuning experiments and brief experiments with LLMs. For sentence simplification, our models surpass or are comparable to state-of-the-art systems. For sentence complexification and same-level paraphrasing, we set new benchmarks. We show that weak classification is an effective way to create strong unsupervised datasets and that target level absolute prompting is more effective than level direction relative prompting.

This research leaves opportunities for future work. For example, using a stronger level classifier to label paraphrase data might improve performance for the paraphrasing tasks. In the same vein, different filtering of ParaNMT or another paraphrasing dataset (Hu et al., 2019) could potentially be used. A human-labeled same-level paraphrasing test dataset does not yet exist, and a modified SARI metric that adequately penalizes repetitions is needed for sentence complexification. Our methods focus on English data, but they can be easily applied to other languages if a different classifier is trained (Khallaf and Sharoff, 2021; Vásquez-Rodríguez et al., 2022) and a non-English paraphrasing dataset is used (Scherrer, 2020; Lu et al., 2021; Martin et al., 2022). Finally, a thorough investigation on how well LLMs can change sentence complexity is necessary.

Acknowledgments

We thank the reviewers and editor Dr. Sara Rosenthal for providing valuable feedback that made this paper much better. We would also like to thank Dr. Laura Vásquez-Rodríguez and Jhih-Jie Chen for their helpful advice, as well as Andrew Cavicchi for lending us compute power. Finally, we thank those who provided assistance with our human evaluation.

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1166

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020a. 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. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.424

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. https://doi.org/10.18653/v1/D19-3009

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020b. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187. https://doi.org/10.1162/coli_a_00370

Yuki Arase, Satoru Uchida, and Tomoyuki Kajiwara. 2022. CEFR-based sentence difficulty annotation and assessment. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6206–6219, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.416

Dennis Aumiller and Michael Gertz. 2022. UniHD at TSAR-2022 shared task: Is compute all we need for lexical simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 251–258. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Virtual). https://doi.org/10.18653/v1/2022.tsar-1.28

Leonid Berov and Kai Standvoss. 2018. Discourse embellishment using a deep encoder-decoder network. In *Proceedings of the 3rd Workshop on Computational Creativity in Natural Language Generation (CC-NLG 2018)*, pages 11–16, Tilburg, the Netherlands. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6603

Sidney Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, Michael Pieler, Usvsn Sai Prashanth, Shivanshu Purohit, Laria Reynolds, Jonathan Tow, Ben Wang, and Samuel Weinbach. 2022. GPT-NeoX-20B: An open-source autoregressive language model. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 95–136, virtual+ Dublin. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.bigscience-1.9

Ondřej Bojar, Ondřej Dušek, Tom Kocmi, Jindřich Libovickỳ, Michal Novák, Martin Popel, Roman Sudarikov, and Dušan Variš. 2016. CzEng 1.6: Enlarged Czech-English parallel corpus with processing tools dockered. In *Text, Speech, and Dialogue: 19th International Conference, TSD 2016, Brno, Czech Republic, September 12–16, 2016, Proceedings 19*, pages 231–238. Springer. https://doi.org/10.1007/978-3-319-45510-5_27

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Annette Capel. 2012. Completing the English vocabulary profile: C1 and C2 vocabulary. *English Profile Journal*, 3:e1. https://doi.org/10.1017/S2041536212000013

Mei-Hua Chen, Shih-Ting Huang, Jason S. Chang, and Hsien-Chin Liou. 2015. Developing a corpus-based paraphrase tool to improve EFL learners' writing skills. *Computer*

Assisted Language Learning, 28(1):22–40. https://doi.org/10.1080/09588221.2013.783873

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416v5*.

Jordan Clive, Kris Cao, and Marek Rei. 2022. Control prefixes for parameter-efficient text generation. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 363–382, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.gem-1.31

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Anna Dmitrieva and Jörg Tiedemann. 2020. A multi-task learning approach to text simplification. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 78–89. Springer. https://doi.org/10.1007/978-3-030-71214-3_7

Yue Dong, Zichao Li, Mehdi Rezagholizadeh, and Jackie Chi Kit Cheung. 2019. EditNTS: An neural programmer-interpreter model for sentence simplification through explicit editing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3393–3402, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1331

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *arXiv preprint arXiv:2302.11957v1*.

Yoshinari Fujinuma and Masato Hagiwara. 2021. Semi-supervised joint estimation of word and document readability. In *Proceedings of the Fifteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-15)*, pages 150–155, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.textgraphs-1.16

Thomas Gaillat, Andrew Simpkin, Nicolas Ballier, Bernardo Stearns, Annanda Sousa, Manon Bouyé, and Manel Zarrouk. 2022. Predicting CEFR levels in learners of English: The use of microsystem criterial features in a machine learning approach. *ReCALL*, 34(2):130–146. https://doi.org/10.1017/S095834402100029X

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-2501

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, USA. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2019. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751v2*. https://doi.org/10.48550/arXiv.1904.09751

Edward Hu, Rachel Rudinger, Matt Post, and Benjamin Van Durme. 2019. ParaBank: Monolingual bitext generation and sentential paraphrasing via lexically-constrained neural machine translation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6521–6528. https://doi.org/10.1609/aaai.v33i01.33016521

Chieh-Yang Huang, Mei-Hua Chen, and Lun-Wei Ku. 2017. Towards a better learning of near-synonyms: Automatically suggesting example sentences via fill in the blank. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, pages 293–302, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee. https://doi.org/10.1145/3041021.3054163

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. OPT-IML: Scaling language modelinstruction meta learning through the lens of generalization. https://doi.org/10.48550/arXiv.2212.12017

Chao Jiang, Mounica Maddela, Wuwei Lan, Yang Zhong, and Wei Xu. 2020. Neural CRF Model for Sentence Alignment in Text Simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7943–7960, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.709

Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. 2021. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, Online. Association for Computational Linguistics.

Tannon Kew and Sarah Ebling. 2022. Target-level sentence simplification as controlled paraphrasing. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 28–42, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.tsar-1.4

Nouran Khallaf and Serge Sharoff. 2021. Automatic difficulty classification of arabic sentences. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 105–114, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

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, Naval Technical Training Command Millington TN Research Branch. https://doi.org/10.21236/ADA006655

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015, Conference Track Proceedings*. https://doi.org/10.48550/arXiv.1412.6980

Klaus Krippendorff. 2011. Computing Krippendorff's Alpha-Reliability. https://api.semanticscholar.org/CorpusID:59901023

Bruce W. Lee, Yoo Sung Jang, and Jason Lee. 2021. Pushing on text readability assessment: A transformer meets handcrafted linguistic features. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10669–10686, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.emnlp-main.834

Justin Lee and Sowmya Vajjala. 2022. A neural pairwise ranking model for readability assessment. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3802–3813, Dublin, Ireland. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.findings-acl.300

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed,

Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.703

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Multi-task deep neural networks for natural language understanding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4487–4496, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-1441

Xinyu Lu, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2021. An unsupervised method for building sentence simplification corpora in multiple languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 227–237, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.22

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.277

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.415

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2022. MUSS: Multilingual Unsupervised Sentence Simplification by mining paraphrases. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1651–1664, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Eric Villemonte de La Clergerie, Antoine Bordes, and Benoît Sagot. 2021. Multilingual unsupervised sentence simplification. Working paper or preprint.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics. https://doi.org/10.3115/v1/P14-1041

Subhajit Naskar, Soumya Saha, and Sreeparna Mukherjee. 2019. Text embellishment using attention based encoder-decoder model. In *Proceedings of the 4th Workshop on Computational Creativity in Language Generation*, pages 28–38, Tokyo, Japan. Association for Computational Linguistics.

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable Text Simplification with Lexical Constraint Loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/P19-2036

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Alex Ratner, Braden Hancock, Jared Dunnmon, Roger Goldman, and Christopher Ré. 2018. Snorkel MeTaL: Weak supervision for multi-task learning. In *Proceedings of the Second Workshop on Data Management for End-To-End Machine Learning*, pages 1–4.

https://doi.org/10.1145/3209889.3209898, PubMed: 30931438

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Michael Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-English text simplification: A unified multilingual benchmark. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4898–4927, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.269

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier De la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, Mario Šaško, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Davut Emre Taşar, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M. Saiful Bari, Maged S. Al-shaibani, Matteo Manica, Nihal Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre François Lavallée, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey

Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Daniel Mc-Duff, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A. Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S. Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2023. BLOOM: A 176B-parameter open-access multilingual language model. https://doi.org/10.48550/arXiv.2211.05100

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, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-2113

Yves Scherrer. 2020. TaPaCo: A corpus of sentential paraphrases for 73 languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association (ELRA).

Veronica Juliana Schmalz and Alessio Brutti. 2021. Automatic assessment of English CEFR levels using BERT embeddings. *http://ceur-ws.org/Vol-3033/*, 3033.

Bernhard Scholkopf, Kah-Kay Sung, Christopher J. C. Burges, Federico Girosi, Partha Niyogi, Tomaso Poggio, and Vladimir Vapnik. 1997. Comparing support vector machines with Gaussian kernels to radial basis function classifiers. *IEEE Transactions on Signal Processing*, 45(11):2758–2765. https://doi.org/10.1109/78.650102

Thomas Scialom, Tuhin Chakrabarty, and Smaranda Muresan. 2022. Fine-tuned language models are continual learners. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6107–6122, Abu Dhabi, United

Arab Emirates. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.emnlp-main.410

Burr Settles, Geoffrey T. LaFlair, and Masato Hagiwara. 2020. Machine learning–driven language assessment. *Transactions of the Association for Computational Linguistics*, 8:247–263. https://doi.org/10.1162/tacl_a_00310

Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. 2017. Continual learning with deep generative replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 2994–3003, Red Hook, NY, USA. Curran Associates Inc.

Advaith Siddharthan. 2002. An architecture for a text simplification system. In *Language Engineering Conference, 2002. Proceedings*, pages 64–71. IEEE.

Advaith Siddharthan. 2006. Syntactic simplification and text cohesion. *Research on Language and Computation*, 4:77–109. https://doi.org/10.1007/s11168-006-9011-1

Renliang Sun, Wei Xu, and Xiaojun Wan. 2023. Teaching the pre-trained model to generate simple texts for text simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9345–9355, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.595

Kazuki Tani, Ryoya Yuasa, Kazuki Takikawa, Akihiro Tamura, Tomoyuki Kajiwara, Takashi Ninomiya, and Tsuneo Kato. 2022. A benchmark dataset for multi-level complexity-controllable machine translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6744–6752, Marseille, France. European Language Resources Association.

Teerapaun Tanprasert and David Kauchak. 2021. Flesch-Kincaid is not a text simplification evaluation metric. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 1–14, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.gem-1.1

Yi Tay, Mostafa Dehghani, Vinh Q. Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Dara Bahri, Tal Schuster, Steven Zheng, Denny Zhou, Neil Houlsby, and Donald Metzler. 2023. UL2: Unifying language learning paradigms. In *The Eleventh International Conference on Learning Representations*.

Brian Thompson and Matt Post. 2020. Paraphrase generation as zero-shot multilingual translation: Disentangling semantic similarity from lexical and syntactic diversity. In *Proceedings of the Fifth Conference on Machine Translation*, pages 561–570, Online. Association for Computational Linguistics.

Chung-Ting Tsai, Jhih-Jie Chen, Ching-Yu Yang, and Jason S. Chang. 2020. LinggleWrite: A coaching system for essay writing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 127–133, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.17

Satoru Uchida, Shohei Takada, and Yuki Arase. 2018. CEFR-based lexical simplification dataset. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sowmya Vajjala and Ivana Lučić. 2018. OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 297–304, New Orleans, Louisiana. Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-0535

Laura Vásquez-Rodríguez, Pedro-Manuel Cuenca-Jiménez, Sergio Morales-Esquivel, and Fernando Alva-Manchego. 2022. A benchmark for neural readability assessment of texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 188–198, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.tsar-1.18

Elena Volodina, Dijana Pijetlovic, Ildiko Pilán, and Sofie Johansson Kokkinakis. 2013. Towards a gold standard for swedish CEFR-based ICALL. In *Proceedings of the Second Workshop on NLP for Computer-Assisted Language Learning. NEALT Proceedings Series*, volume 17.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2019. Neural text generation with unlikelihood training. *arXiv preprint arXiv:1908.04319v2*.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics. https://doi.org/10.18653/v1/P18-1042

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-demos.6

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. Association for Computational Linguistics.

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

Menglin Xia, Ekaterina Kochmar, and Ted Briscoe. 2016. Text readability assessment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 12–22, San Diego, CA. Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-0502

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. https://doi.org/10.1162/tacl_a_00139

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. https://doi.org/10.1162/tacl_a_00107

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. Association for Computational Linguistics. https://doi.org/10.18653/v1/D17-1062

Xin Zhao. 2022. Leveraging artificial intelligence (AI) technology for English writing: Introducing Wordtune as a digital writing assistant for EFL writers. *RELC Journal*, page 00336882221094089. https://doi.org/10.1177/00336882221094089

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.