# Enhancing Sentence Simplification in Portuguese: Leveraging Paraphrases, Context, and Linguistic Features

**Arthur Scalercio**[1] **, Maria José Finatto**[2]**, Aline Paes**[1]
[1]Institute of Computing, Universidade Federal Fluminense, Niterói, RJ, Brazil
[2]Universidade Federal do Rio Grande do Sul, Porto Alegre, RS, Brazil
{arthurscalercio@id, alinepaes@ic}.uff.br, mariafinatto@gmail.com

## Abstract

Automatic text simplification focuses on transforming texts into a more comprehensible version without sacrificing their precision. However, automatic methods usually require (paired) datasets that can be rather scarce in languages other than English. This paper presents a new approach to automatic sentence simplification that leverages paraphrases, context, and linguistic attributes to overcome the absence of paired texts in Portuguese. We frame the simplification problem as a textual style transfer task and learn a style representation using the sentences around the target sentence in the document and its linguistic attributes. Moreover, unlike most unsupervised approaches that require style-labeled training data, we fine-tune strong pre-trained models using sentence-level paraphrases instead of annotated data. Our experiments show that our model achieves remarkable results, surpassing the current state-of-the-art (BART+ACCESS) while competitively matching a Large Language Model.

## 1 Introduction

Text simplification consists of making a text easier to read and understand by wider audiences while preserving most of its original meaning (Al-Thanyyan and Azmi, 2021). Simplification has a variety of critical social applications, for example, increasing accessibility for individuals with reading difficulties (Aluísio and Gasperin, 2010), cognitive disabilities such as aphasia (Carroll et al., 1998), dyslexia (Rello et al., 2013), and autism (Evans et al., 2014), or for non-native speakers (Paetzold and Specia, 2016). Moreover, expert-written texts, such as those in science, medical, financial, and legal fields, can exhibit a high level of complexity, making them difficult for the public to read and understand. This complexity stems from specialized jargon and technical language, which, while precise and necessary within the field, can pose significant reading challenges to the layperson (Cao et al., 2020).

The Plain Language movement dates back to the 1940s, advocating for a straightforward writing style that avoids unnecessary jargon or complex vocabulary (Felsenfeld, 1981). However, legislation towards making legal and governmental texts more accessible to the public has become prominent only in this century[1]. In Brazil, for example, while a few local governments have specific laws for simple language since 2010 (Martins et al., 2023), only in 2023 a national law has emerged[2].

Given the amount of human knowledge still outside that movement, achieving the goals of plain language initiatives worldwide requires *automatic* approaches. Additionally, the unique characteristics of cultural writing styles necessitate developing customized solutions for each. For instance, Portuguese writers favor lengthy sentences, incorporate passive voice and complex verb conjugations, add implicit coreferences, and use extensively verbal phrases and other intricate structural elements. Those practices usually make sentences more complex than they could be. However, most text simplification work has targeted English, with a relative abundance of aligned pairs for supervised training of automatic models. Ryan et al. (2023) shows the amount of parallel simplification data available by language. Parallel pairs are very scarce in languages other than English, French, and Russian.

This paper focuses on Portuguese, although our model architecture is language-agnostic. The scarcity of both simplification models and datasets in Portuguese highlights the need to make available more resources for the community, mainly because it is the language spoken by about 260 million people [3]. Portuguese text simplification

[1] https://www.dni.gov/index.php/plain-language-act

[2] https://www.gov.br/gestao/pt-br/assuntos/inovacao-governamental/cinco/cinforme/edicao_1-2023/linguagem-simples

[3] https://www.ethnologue.com/



*Findings of the Association for Computational Linguistics ACL 2024*, pages 15076–15091
August 11-16, 2024 ©2024 Association for Computational Linguistics

also inherits the challenges of natural language generation, namely the lack of accurate evaluation metrics (Reiter and Belz, 2009; Gatt and Krahmer, 2018). Furthermore, we focus on models trained from non-aligned pairs to account for the lack of paired datasets.

Acknowledging that text simplification is highly audience-centric (Stajner, 2021), recent work has focused on developing techniques to control the degree of simplicity of the output. Controllable Text Simplification can be seen as a conditional language modeling task, where the source text $X$ is rephrased as an output $Y$ that presents attributes $V$ evaluated by a model $P(Y|X, V)$ (Prabhumoye et al., 2020). Agrawal and Carpuat (2023) provides an overview of the control token attributes introduced in prior work for text simplification. They range from high-level features, e.g., grade levels, to low-level syntactic, lexical, and semantic attributes.

This work proposes a new controllable sentence simplification model trained using pairs of paraphrase data ($s_{source}$, $s_{target}$) plus a sentence around the target paraphrase. In addition to conditioning the text generation on linguistic attributes of the target text, we also condition on its context. Similar to Riley et al. 2021, we take as context a sentence of the same document as the target sentence context, relying on the observation that style is a "slow-moving" feature, which tends to be uniform over spans of a document.

We learn a representation from the low-level features of the target sentence, and another one from its context. We combine and feed them into a simple neural network to obtain our final style representation, which will guide the decoder. PT-T5 (Carmo et al., 2020), a Portuguese version of T5 (Raffel et al., 2020), is our base neural sequence-to-sequence (Seq2Seq) architecture, given the successful results of this Transformer-based model on several NLP tasks.

The contributions of this paper are:

1. A novel few-shot approach to training simplification models with paraphrase data and the sentence adjacent to the target paraphrase.
2. To the best of our knowledge, this is the first study that uses both linguistic features and context to learn a representation to guide the simplification process.
3. In the experiments with three Portuguese simplification datasets, our method outperformed strong baseline approaches, including SOTA and large language models.
4. We release pre-trained models, paraphrase data, a new dataset, and code for training [4].

## 2 Related Work

### 2.1 Sentence Simplification

Most research in text simplification usually follow a generative or an edit-based supervised strategy. The first case includes sequence-to-sequence models (Nisioi et al., 2017) using transformer (Vaswani et al., 2017) architectures and reinforcement learning (Zhang and Lapata, 2017), leveraging external paraphrase datasets (Zhao et al., 2018), and integration of syntactic rules (Maddela et al., 2021). Conversely, edit-based supervised models have been crafted to use parallel complex-simple sentence pairs. Alva-Manchego et al. (2017) learns which operations should be performed to simplify a sentence, and Omelianchuk et al. (2021) predicts token-level operations in a non-autoregressive manner.

### 2.2 Controllable Sentence Simplification

Recently, researchers have introduced explicit parameters to guide and control the simplified output (Nishihara et al., 2019; Martin et al., 2020; Agrawal et al., 2021). Martin et al. (2020) introduced four hyperparameters in the AudienCe-CEntric Sentence Simplification (ACCESS): the number of characters, Levenshtein similarity (Ristad and Yianilos, 1996), word rank and dependency tree depth, to control the length, similarity, lexical complexity, and syntactic complexity, respectively. These features are added to the input sequence, and subsequently, the model undergoes training to produce the desired target sequence. Agrawal et al., 2021 replaced those parameters with more straightforward simplification grades, overcoming the need for specific linguistic knowledge. These approaches are supervised, relying on parallel complex-simple pairs available in English.

The Multilingual Unsupervised Sentence Simplification (MUSS) (Martin et al., 2022) technique involves gathering paraphrase datasets in various languages. Then, instead of using complex-simple parallel corpora, they trained their simplification models using these paraphrases, incorporating ACCESS control tokens. This method has surpassed other unsupervised text simplification (TS) models, establishing SOTA results. Conversely, Agrawal

[4] https://github.com/scalercio/portuguese-simplification

and Carpuat (2023) points out some drawbacks of applying control tokens at the corpus level rather than at the sentence level. To address this, they suggest a control token predictor that leverages surface-form features from the source text and considers both the source and target grade levels.

Our system utilizes the MUSS mining procedure to extract Portuguese paraphrases, but we took it a step further by also mining the context surrounding the target paraphrase. Our training framework also diverges considerably from the MUSS approach. In addition to incorporating context, we have used linguistic features in a novel way, as detailed in Section 3.

### 2.3 Simplification in Portuguese

Previous simplification works in Portuguese that rely on machine learning extensively use parallel corpora. Specia (2010) formulated a Statistical Machine Translation (SMT) framework to learn how to translate from complex to simplified sentences, given a parallel corpus of original and simplified texts. Hartmann and Aluísio (2020) introduced a pipeline designed explicitly for the lexical simplification of informational texts in Brazilian Portuguese, targeting elementary school children. Considering the limited resources available, zero-shot, few-shot, and unsupervised approaches emerge as promising avenues for Portuguese text simplification.

In this context, Martin et al. (2022) contributed a neural model[5] that is trained on a substantial corpus of mined Portuguese paraphrases. Furthermore, Feng et al. (2023) analyzed the zero-/few-shot learning ability of LLMs to simplify sentences in several languages by evaluating them on benchmark test sets in several languages, Portuguese included. Our work also follows the tendency to design simplification models that operate even without parallel annotated corpora.

## 3 Method

Figure 1 illustrates our proposed architecture. At a high level, our approach follows (Riley et al., 2021), training a denoising autoencoder conditioned on a fixed-width style vector. However, our approach differs in two fundamental ways: firstly, it leverages paraphrases instead of reconstructing a corrupted input; secondly, it introduces additional parameters to the neural style extractor, aiming to incorporate linguistic features alongside the context.

### 3.1 Mining Paraphrases in Portuguese

Our model is trained in a purely unsupervised way, solely using paraphrases. We adopted the mining procedure described in Martin et al. (2022) to extract Portuguese texts from CCNet (Wenzek et al., 2020), an extraction of Common Crawl. In line with that methodology, we break down documents into multiple sequences. We compute n-dimensional embeddings for each sequence using the LASER tool (Artetxe and Schwenk, 2019). These embeddings are then indexed using Faiss (Johnson et al., 2019). At last, we query each sequence against the Faiss index to retrieve its most similar counterpart. In addition to discarding poor alignments, we also eliminate paraphrases where at least one paraphrase is not recognized as Portuguese by a language classifier (Joulin et al., 2016).

To capture the query paraphrase context, we randomly select a different sequence from the same document to which the query belongs. This selection is subject to three conditions: the sequence must not be identical to the query, it must not encompass the query, nor should the query encompass it. As a result, we generated a dataset comprising $472,530$ triplets, of which $470,530$ were used for training purposes.

### 3.2 Architecture and Style Learning

Similar to the approach designed in Riley et al. (2021), we employ a T5 (Raffel et al., 2020) Seq2Seq model enhanced with an extra T5 encoder. The extra encoder is dedicated to learning a style representation that aids the decoder during generation. The model is fed with the source and target sentences and the target sentence context ($s_{source}$,$s_{target}$, $context_{target}$), respectively. It is trained to generate the target sentence $s_{target}$. In our setup, the source and target sentences are paraphrases of each other, and the context is a sentence from the same document as the target sentence.

We augmented the architecture by enhancing the style extractor to incorporate linguistic attributes of the target paraphrase. Concretely, we integrated two straightforward feed-forward networks into the style extractor, as illustrated in Figure 2. Both networks have only one intermediary layer. The first network is tasked with processing the linguistic attributes, producing a linguistic representation

[5]https://github.com/facebookresearch/muss.git

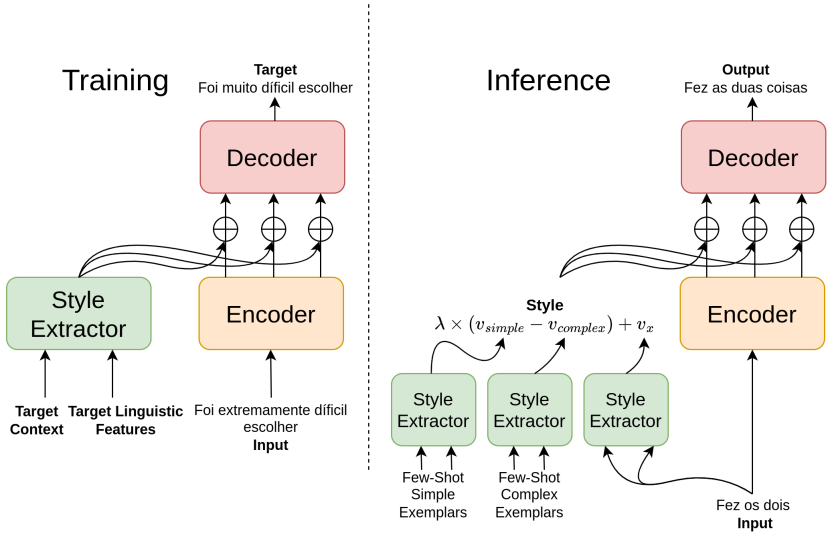


Figure 1: Our architecture for few-shot Portuguese Simplification. All the transformer stacks are initialized from pre-trained Portuguese T5. During training, the model learns to generate a paraphrase conditioned on a fixed-width "style vector" extracted from the context sentence and the target linguistic features. At inference time, a new style vector is formed via "targeted restyling": adding a directional delta to the extracted style of the input text.

that matches the size of the encoder's hidden state. In line with the TextSETTR model, the context is processed through the T5 encoder, yielding a context representation. These two vectors – the linguistic and context representations – are then concatenated and fed into the second feed-forward network. This process culminates in our final style representation, which encapsulates low-level linguistic features and higher-level attributes such as the target simplicity style, among others, like humor and formality.

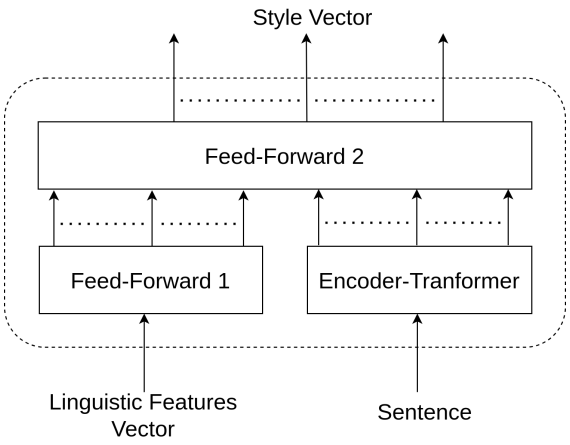


Figure 2: Our Enhanced Style Extractor

Building upon the work of Sheang and Saggion (2021), we equip our style extractor with four linguistic features to guide various facets of text simplification: character length, number of words, word rank, and dependency tree depth. Additionally, we experimented with sentence syllable count as a feature, but the outcomes were not satisfactory. To integrate the style vector into the rest of the model, we incorporate it into each of the encoder's final hidden states. Our model initializes its weights with a pre-trained Portuguese T5 model (Carmo et al., 2020). Both the context representation extractor and the encoder are initialized from this pre-trained encoder; however, their weights are not tied throughout the training process.

### 3.3 Inference Procedure

Regarding the style representation approach, most previous work on style transfer (Dai et al., 2019; Scalercio and Paes, 2023) predominantly employs a method where a fixed set of discrete styles is predefined. Each style's unique representation is learned and integrated into the network. In contrast, our approach diverges significantly. We do not impose predefined style constraints, aiming instead to obtain rich and expressive style representations not specified in advance. This allows for a more flexible and nuanced understanding of style. For instance, a specific style vector in our model could encode that a sentence is formal, simple, and exhibiting a regional accent, among other characteristics.

At inference time, we assume to have access to a limited number of demonstration sentences for both complex and simple styles (varying from 1 to 100). We employ a style extractor to derive style

vectors for each demonstration sentence. Next, we calculate the average vector for each style category. This process forms two distinct averaged vectors $v_{complex}$ and $v_{simple}$.

To transform an input sentence $x$, we implement restyling in the relevant direction, as in Riley et al. (2021). This process begins with extracting the original style vector $v_x$ from the input. We then determine the target output style by shifting in the direction of the difference between the simple and complex style attributes, in line with Equation 1. This calculated shift generates a new style vector, which is then used for decoding. We have observed that the scale $\lambda$ is a crucial hyperparameter to adjust. Typically, values within the range of [1, 14] yield good results, with the optimal values varying based on the specific exemplars involved.

$$v_x + \lambda \times (v_{simple} - v_{complex}) \quad (1)$$

## 4 Experimental Setting

### 4.1 Datasets

We used the mined data described in subsection 3.1 as training data. For validation, testing, and as the source for our few-shot demonstration, we use PorSimplesSent (Leal et al., 2018), which was built from the parallel corpus PorSimples (Aluísio and Gasperin, 2010). PorSimplesSent features multiple versions, distinguishing whether the complex texts were split during simplification. Our primary experiments use the version where the complex sentences remain unsplit. To enhance diversity, we exclude any simplifications that originated from an already simplified pair. This results in a total of 1515 sentences. We divided these into distinct sets: 200 sentences are allocated for few-shot demonstrations, 709 for the validation set, and 606 for the test set.

Additionally, we evaluated our model on two other datasets. The first is ASSET-PT, the Portuguese version of the ASSET dataset, also assessed by Martin et al. (2022) and Feng et al. (2023). It comprises 359 complex Portuguese sentences, each accompanied by ten simplifications. ASSET-PT is a translation of the original English version (Alva-Manchego et al., 2020) using the Google Translate API and made available by Martin et al., 2022 at their repository.

The other, Museum-PT, is a document simplification dataset proposed in Finatto and Tcacenco (2021) and curated explicitly to this work. The Museum-PT dataset originated from simplifications carried out by linguists, aiming to reduce or eliminate complexity by applying Plain Language techniques and adhering to principles of Textual and Terminological Accessibility. The set comprises written texts accompanying experiments and objects from science and technology museums, aimed at a general audience. Our curated version of the dataset includes both sentence and paragraph alignments and can be a valuable validation/testing resource for Portuguese document simplification Cripwell et al. (2023). The dataset comprises 42 documents, 80 document simplifications, 168 paragraphs, and 460 sentences. Each sentence is also annotated with the operation performed during the simplification, which can be *copy, rephrase, split* or *delete*. For testing our model, we selected all the sentences annotated with *rephrase* or *split*, totaling 476 complex-simple pairs.

### 4.2 Baselines

Our evaluation comprises two robust baselines to assess the performance of our models.

**MUSS-Unsupervised** (Martin et al., 2022) This is an unsupervised multilingual simplification method that fine-tunes BART (Lewis et al., 2020), leveraging paraphrases and control tokens from ACCESS (Martin et al., 2020) during training. It is the only transformer-based unsupervised open-source implementation available for Portuguese.

**Open AI's GPT-3.5-Turbo** Given the impressive results of LLMs in a wide range of NLP tasks, we also benchmark our model against the GPT-3.5-Turbo. This particular LLM stands out with the best results in a study that benchmarks several LLMs on English Sentence Simplification (Kew et al., 2023). Following their settings, we use Nucleus Sampling with a probability of 0.9, a temperature of 1.0, and a maximum output length of 100 tokens. We perform each inference run three times to account for the probabilities. We first investigated the performance of zero, one, and few-shot in the PorSimplesSent dataset. Confirming the findings of Feng et al. (2023), the one-shot approach proved the most successful. Therefore, this is the setting displayed throughout the results section. The other results and more details about the prompts and demonstration selection are in Appendix A.

### 4.3 Evaluation Metrics

Our evaluation comprises automatic metrics widely used in text simplification task (Sheang and Saggion, 2021; Martin et al., 2022). We measure simplicity using SARI (Xu et al., 2016), meaning preservation using BERTScore (Zhang* et al., 2020), and BLEU (Papineni et al., 2002). These metrics are computed using the EASSE package (Alva-Manchego et al., 2019)[6]. We also report the percentage (%) of unchanged outputs (i.e., exact copies), following Agrawal and Carpuat (2023).

### 4.4 Training and Inference Details

All models undergo training using identical hyperparameters, including a batch size of 80 for T5-base and 8 for T5-large, a maximum token limit of 85, a learning rate of 1e-4, Adam epsilon of 1e-8, and 10 epochs. The remaining parameters are left at default values from the Transformers library. Additionally, the seed is set to 123 to ensure reproducibility. Our models are trained using a server with a single RTX4090 GPU with 24GB of memory and 64GB of RAM. Training the T5-large model for ten epochs typically requires approximately 20 hours, while the base takes 5 hours. Our base version has 334M parameters and the large 1.1B. We used the Spacy package with its default configuration for Portuguese to calculate our syntactic attribute.

To perform inference, we follow the procedure from Section 3.3. For our default setup, we sampled 100 complex and 100 simple exemplars from our few-shot exemplars resource and fixed them for all experiments. Unless otherwise specified, we use greedy decoding and a delta scale of $\lambda = 12$. The model with the highest SARI score on the validation set was selected to be run on the test set.

## 5 Results

### 5.1 Automatic Evaluation

We evaluate our models automatically on three different datasets. Table 1 reports the results of the automatic evaluation of our models compared with the baselines. In PorSimplesSent, our model obtained a +0.46 SARI improvement over the LLM baseline and +1.34 SARI improvement over the open-source state-of-the-art unsupervised method (MUSS). We also have the highest content preservation metric (BERTScore).

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

In the Museum-PT dataset, our model achieved the highest score in content preservation. However, it did not surpass GPT-3.5 Turbo in terms of the simplicity metric, despite attaining a considerably high value. However, the reader must remember this is a closed-source model with a paid API.

When tested on the ASSET-PT dataset, our model was outperformed by both baseline models. We attribute this outcome to the nature of the ASSET-PT dataset, a translation from English. This translation process often leads to significant content deviations in the reference texts compared to the original input texts. Given that our model is very conservative about meaning preservation, the substantial differences in the reference texts adversely affect its performance, as it tends to be penalized for maintaining closer adherence to the original content.

| Model | Metrics | | | |
|---|---|---|---|---|
| | SARI | BScore | BLEU | (%) U |
| **PorSimplesSent** | | | | |
| MUSS | 38.30 | .8976 | 51.38 | 3.46 |
| GPT3.5T | 39.18 | .8805 | 38.01 | 0.26 |
| Ours | **39.64** | .9024 | 48.2 | 3.79 |
| **ASSET-PT** | | | | |
| MUSS | 40.04 | .9467 | 81.05 | 3.35 |
| GPT3.5T | **45.66** | .9271 | 66.50 | 0.46 |
| Ours | 38.28 | .9408 | 69.59 | 3.90 |
| **Museum-PT** | | | | |
| MUSS | 39.31 | .8534 | 32.12 | 3.99 |
| GPT3.5T | **47.23** | .8468 | 26.27 | 0.63 |
| Ours | 41.62 | .8550 | 32.36 | 5.46 |

Table 1: Automatic Evaluation Results on Portuguese text simplification for three datasets

**Statistical significance analysis** Given the small-sized test sets, we conducted statistical tests to compare the SARI metric achieved by our model and the other two baselines. We omitted the ASSET-PT dataset since it is generated through translation from another dataset in English. SARI is a corpus-based metric, so we applied the Paired Bootstrap Resampling test (Koehn, 2004) to evaluate the significance of our results against the baselines. We repeatedly (1000 times) create new virtual test sets with the same original size from each test set by drawing sentences with replacement from the test set. The statistical significance in Table 2 is the number of times one system outperforms the other system.

| Dataset | Hypothesis Test | Significance |
|---|---|---|
| Porsimpl. | Ours > Muss | 97.7% |
| Porsimpl. | Ours > GPT3.5T | 99.0% |
| Museum-PT | Ours > Muss | 99.8% |
| Museum-PT | GPT3.5T > Ours | $\approx 100\%$ |

Table 2: Paired Bootstrap Resampling Statistical Tests on SARI metric

| Model | Simplicity | Content | Fluency |
|---|---|---|---|
| MUSS | 3.1 | 3.4 | 4.1 |
| Gpt3.5Turbo | **3.8** | 3.8 | **4.6** |
| Ours | 3.1 | 4.0 | 4.2 |
| Reference | 3.6 | **4.3** | 4.5 |

Table 3: Results from human evaluation in PorSimplesSent dataset.

These tests endorse our main results. In the Porsimplessent dataset, our model is better than MUSS and GPT3.5T baselines with 97.7% and 99.0% of statistical significance, respectively. In the Museum-PT, our model is better than Muss with 99.8% of significance, while GPT3.5T is better than ours with almost 100% of significance.

### 5.2 Human Evaluation

Furthermore, to thoroughly assess the effectiveness of our method, we conducted a human evaluation focusing on three key aspects: adequacy, fluency, and simplicity. The results are detailed in Table 3. Simplifications are evaluated on a five-point Likert scale (1-5). It assesses three critical dimensions: is the meaning preserved? (adequacy), is the simplification fluent? (fluency), and has the simplification indeed made the text simpler to understand (simplicity). We recruited two volunteer native Portuguese speakers with a background in linguistics and asked them to assess sentences based on the above dimensions. More detailed instructions can be found in Appendix B. We randomly select 80 complex sentences from our PorSimplesSent test set for this evaluation. We presented the corresponding simplified reference for each sentence and three additional simplifications generated by our model, MUSS, and GPT-3.5-Turbo. Each simplification was rated once.

The results in Table 3 highlight the great capability of LLM in text simplification, with GPT-3.5 achieving the highest scores in the simplification metric. Interestingly, GPT-3.5's performance was so effective that it tricked the automatic metric, and even surpassed the score reached by the references, which are sentences created by human experts.

The other results align with our automatic evaluation and add confidence to the efficacy of our proposed model and experimental techniques. Notably, our approach attained the highest score in content preservation compared to all baselines. Furthermore, all models evaluated demonstrated a high level of fluency, indicating these models' overall effectiveness and linguistic competence in producing coherent and natural-sounding text. Some poorly and well-evaluated sentences are in the appendix D. We reached a moderate inter-annotator agreement with a Cohen's Kappa (Cohen, 1960) of 0.5.

### 5.3 Ablation Analyses

We performed ablation tests on Museum-PT and PorSimplesSent datasets, leaving out ASSET-PT because it comprises translations instead of human experts generating simplifications.

**Architecture of the Network** We ablate over the size of the neural network and its architecture. We compared the models trained with the large and base versions of the Portuguese T5. Furthermore, we also analyze the impact of using context and features to obtain our style representation.

Table 4 points out that increasing the model size leads to enhanced performance. Regarding the components selected for learning style representation, incorporating both context and linguistic features significantly benefits the Museum-PT performance across all evaluated metrics. In the PorsimplesSent dataset, adopting context and linguistic features and adopting only the context achieved similar results regarding simplicity and meaning preservation. Nevertheless, we can see by the metric % of outputs unchanged that adopting both context and linguistic features brings more lexical diversity to the outputs without losing semantics, avoiding the copy-source flaw. Besides, adopting the linguistic feature extractor network is almost parameter-free compared to the context extractor, which requires a whole transformer encoder network.

**Few-Shot Demonstration** Although our method learns unsupervised, it requires few-shot demonstration examples during the inference phase. To assess their influence on the model's performance, we varied the samples available at inference time

| Size | Type | PorSimplesSent | | | | Museum-PT | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | SARI | BScore | BLEU | (%) U | SARI | BScore | BLEU | (%) U |
| **PT-T5 BASE** | FEAT | 38.12 | .9091 | 52.49 | 9.24 | 37.89 | .8581 | 33.14 | 9.87 |
| | CTX | 39.28 | .9129 | 54.31 | 9.90 | 39.18 | .8589 | 33.70 | 10.29 |
| | FEAT+CTX | 39.10 | .9071 | 49.36 | 6.43 | 39.38 | .8555 | 31.86 | 9.03 |
| **PT-T5 LARGE** | FEAT | 38.25 | .9101 | 53.09 | 11.88 | 38.61 | .8570 | 32.86 | 8.40 |
| | CTX | **39.65** | .9112 | 53.12 | 10.72 | 39.86 | .8578 | 33.26 | 7.77 |
| | FEAT+CTX | **39.64** | .9024 | 48.2 | 3.79 | **41.62** | .8550 | 32.36 | 5.46 |

Table 4: We display metrics for various architectural choices

qualitatively and quantitatively. Due to computational resource constraints, only the smallest model was used in this ablation.

To evaluate the limits of our model's generalization capabilities, we constrained the set of exemplars to 50 randomly chosen and four manually selected types of simplification: syntactic simplifications, changes in word order, anaphora, and eliminating redundant information. These manually selected exemplars are in Appendix C. Conversely, we also increased the number of demonstration examples to 200. All exemplars were picked from our PorSimplesSent few-shot demonstrations set.

To determine the impact of sample quality on performance, we experimented with two sets, ensuring that each set consistently comprised 100 samples. We tested the following variations: (1) SPLIT: the demonstration examples are 100 examples that the complex sentence was split; and (2) MIXED: 50 simplifications used in our default configuration and 50 simplifications that the complex sentence was split. We also extended the validation set to include simplification pairs with split operations in the SPLIT and MIXED configurations. These simplifications that cause the complex sentence to suffer a split are available in the PorSimplesSent Project, but they are not part of the default configuration of our experiments. We left out the BLEU metric since it correlates poorly with human judgments and often penalizes more straightforward sentences (Sulem et al., 2018), as we could further evidence from our human evaluation. In these ablations, we reduced the delta scale to $\lambda = 8$.

The results presented in Table 5 show that none of the datasets benefited from increasing the number of available instances to 200 at inference time. On the other hand, using only four few-shot exemplars, there was an improvement in performance for both datasets when compared to using only 100 instances. In the PorSimplesSent dataset, the result with just four samples matched our best result with the large version of PT-T5.

The results when we change the type of exemplars are even more interesting. In the PorsSimplesSent dataset, the performance deteriorated, which aligns logically with the nature of the dataset. In this setup, the validation set and available few-shot pairs contain simplifications that have undergone splitting. Since this type of simplification is absent in the test set, employing strategies that rely on split simplifications is not helpful.

Conversely, we observed a significant improvement in the results of the Museum-PT dataset. Notably, the ablation model which uses as exemplars only simplifications that suffered split exceeded the performance of the best model (as initially identified) by 0.82 points in SARI, despite the model being substantially smaller. This finding suggests that our original selection of the best model might not have been optimal. It implies that using a more diverse and extensive range of validation and few-shot sets could result in more general and superior models.

| Model | Metrics | | |
|---|---|---|---|
| | SARI | BScore | (%) U |
| PorSimplesSent | | | |
| 4 exemplars | **39.59** | .9065 | 9.08 |
| 50 exemplars | 38.77 | .8952 | 2.64 |
| 200 exemplars | 38.54 | .9067 | 7.76 |
| SPLIT | 38.82 | .8966 | 3.3 |
| MIXED | 38.54 | .8963 | 3.46 |
| Museum-PT | | | |
| 4 exemplars | 39.77 | .8537 | 9.87 |
| 50 exemplars | 40.61 | .8498 | 2.31 |
| 200 exemplars | 39.14 | .8564 | 6.51 |
| SPLIT | **42.44** | .8536 | 1.47 |
| MIXED | 41.20 | .8535 | 3.78 |

Table 5: Evaluation results for the Few-Shot ablation study using the PT-T5 base. $\lambda = 8$

### 5.4 Embedding Visualization

To analyze our learned style representations, we projected the vectors of simple and complex sentences from the museum test set into a three-dimensional space using t-SNE and analyzed the vector distributions. Figure 3 shows a good visual separation between the two classes in the museum dataset. We also do not expect a clean linear separation within each attribute, since we aim for the learned vectors to encode many style attributes simultaneously. That is why our targeted restyling inference procedure is essential, as it washes out most untargeted attributes.

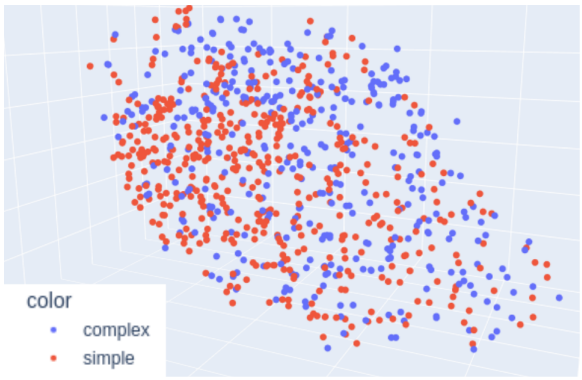


Figure 3: 3D t-SNE embeddings of the style vectors of the sentences from the museum test set extracted by our model

## 6 Conclusion

This paper proposed a new method that leverages a pre-trained text-to-text model (T5), paraphrases, context, and linguistic features to address the Controllable Sentence Simplification task. Integrating linguistic features enhances the control and interpretability of the generated output. We experimented with three datasets, two meticulously curated by linguistics experts and one translated from English. The results demonstrated notable performance improvements on the SARI metric, surpassing the current open-source state-of-the-art model and showing competitive results against an LLM. Our strategy is particularly advantageous due to its minimal requirement for exemplars during inference. Future work could explore obtaining richer representations from the parsed dependency tree with graph neural networks and applying our models in other languages besides Portuguese using multilingual models.

## 7 Limitations

One issue with our approach is that the context style extractor requires a 50% increase in the parameters of a seq2seq model. This increase translates to 110M parameters for the T5-base model and 385M for the T5-large model.

As indicated in our ablation studies, using a smaller and less diverse validation set may have impacted the selection of the model. Although the validation set is not used in training, the best model is invariably selected based on its performance on this set. Employing a smaller set can result in a less versatile model that may struggle with generalization.

Our experiments were exclusively focused on sentence simplification in Portuguese. To replicate our proposed method in other languages, paraphrase datasets would be necessary for training, and annotated simplification datasets would be necessary for validation and testing. Furthermore, considering that the nature of simplification varies across languages, this would demand the involvement of human experts with specific language expertise for conducting the human evaluation.

Another point worth mentioning is the background of the individuals who conducted our evaluations. Although we selected linguistic experts knowledgeable in Portuguese simplification theory and techniques, they are not the intended end-users of text simplification. This suggests the need for evaluations involving individuals who require simplified texts, potentially involving tailored questionnaires to address their specific needs.

## 8 Ethics Statement

This research aligns with the ACL Ethics Policy. Our contributions enhance expertise in text simplification (Section 2.6). We have ensured that all models, datasets, and computing resources are utilized with proper authorization, respecting access rights and licenses (Section 2.8). This work fosters the professional development of the research team (Section 3.5) and intends to benefit the research community and society broadly (Section 3.1) by expanding the understanding of machine learning models capabilities in the specific task of text simplification for Portuguese. We checked the datasets to ensure they did not have offensive content (to the best of our understanding and cultural background). Some sentences might contain names of public figures, such as politicians and celebrities, but only

stating facts and not subjective opinions. To keep the meaning of sentences as initially intended by the experts, we decided not to take any action to anonymize them.

While our study adheres to the ethical code, it is crucial to address some aspects highlighted by Gooding (2022) regarding ethical considerations in text simplification. For instance, our motivation for text simplification in the introduction and subsequent experiments does not specifically target any audience. The techniques used for text simplification should be tailored to the needs and requirements of diverse groups, such as individuals with disabilities, low-literacy readers, young children, and non-experts.

Regarding potential risks, if the model is integrated into educational tools, it is advisable to refrain from using it with students who should be presented with complex text skills in reading and writing. Simplification may diminish exposure to complex vocabulary, advanced grammar, and discourse features. Moreover, it might distort the original texts' meaning, nuance, or tone and foster dependency or lack of challenge for students.

This way, if our method is to be implemented in assistive or educational technologies in the future, further research is needed to determine the most suitable audience and include constraints to make it fully aligned with those critical ethical aspects.

## References


Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Conference on Empirical Methods in Natural Language Processing*.

Sweta Agrawal, Weijia Xu, and Marine Carpuat. 2021. A non-autoregressive edit-based approach to controllable text simplification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3757–3769, Online. Association for Computational Linguistics.

Suha S. Al-Thanyyan and Aqil M. Azmi. 2021. Automated text simplification: A survey. *ACM Comput. Surv.*, 54(2).

Sandra Aluísio and Caroline Gasperin. 2010. Fostering digital inclusion and accessibility: the porsimples project for simplification of portuguese texts. In *Proceedings of the NAACL HLT 2010 Young Investigators Workshop on Computational Approaches to Languages of the Americas*, pages 46–53.

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

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

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.

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1061–1071, Online. Association for Computational Linguistics.

Diedre Carmo, Marcos Piau, Israel Campiotti, Rodrigo Nogueira, and Roberto Lotufo. 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. *arXiv preprint arXiv:2008.09144*.

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

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37 – 46.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Context-aware document simplification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13190–13206, Toronto, Canada. Association for Computational Linguistics.

Ning Dai, Jianze Liang, Xipeng Qiu, and Xuanjing Huang. 2019. Style transformer: Unpaired text style

transfer without disentangled latent representation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5997–6007, Florence, Italy. Association for Computational Linguistics.

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

Carl Felsenfeld. 1981. The plain english movement. *Can. Bus. LJ*, 6:408.

Yutao Feng, Jipeng Qiang, Yun Li, Yunhao Yuan, and Yi Zhu. 2023. Sentence simplification via large language models. *ArXiv*, abs/2302.11957.

Maria José Bocorny Finatto and Lucas Meireles Tcacenco. 2021. Tradução intralinguística, estratégias de equivalência e acessibilidade textual e terminológica. *Tradterm*, 37(1):30–63.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *J. Artif. Intell. Res.*, 61:65–170.

Sian Gooding. 2022. On the ethical considerations of text simplification. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies, SLPAT@ACL 2022, Dublin, Ireland, May 27, 2022*, pages 50–57. Association for Computational Linguistics.

Nathan Siegle Hartmann and Sandra Maria Aluísio. 2020. Adaptação lexical automática em textos informativos do português brasileiro para o ensino fundamental. *Linguamática*, 12(2):3–27.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. BLESS: Benchmarking large language models on sentence simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13291–13309, Singapore. Association for Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Sidney Evaldo Leal, Magali Sanches Duran, and Sandra Maria Aluísio. 2018. A nontrivial sentence corpus for the task of sentence readability assessment in Portuguese. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 401–413, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

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.

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

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 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.

Heloísa Tavares Martins, Adriano Rosa da Silva, and Márcia Teixeira Cavalcanti. 2023. Linguagem simples: um movimento social por transparência, cidadania e acessibilidade. *Cadernos do Desenvolvimento Fluminense*, (25).

Daiki Nishihara, Tomoyuki Kajiwara, and Yuki Arase. 2019. Controllable text simplification with lexical constraint loss. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 260–266, Florence, Italy. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Kostiantyn Omelianchuk, Vipul Raheja, and Oleksandr Skurzhanskyi. 2021. Text Simplification by Tagging. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 11–25, Online. Association for Computational Linguistics.

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

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Shrimai Prabhumoye, Alan W Black, and Ruslan Salakhutdinov. 2020. Exploring controllable text generation techniques. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1–14, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, 35(4):529–558.

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

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. 2021. TextSETTR: Few-shot text style extraction and tunable targeted restyling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3786–3800, Online. Association for Computational Linguistics.

Eric Sven Ristad and Peter N. Yianilos. 1996. Learning string-edit distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20:522–532.

Michael Joseph Ryan, Tarek Naous, and Wei Xu. 2023. Revisiting non-english text simplification: A unified multilingual benchmark. In *Annual Meeting of the Association for Computational Linguistics*.

Arthur Scalercio and Aline Paes. 2023. Masked transformer through knowledge distillation for unsupervised text style transfer. *Natural Language Engineering*, page 1–36.

Kim Cheng Sheang and Horacio Saggion. 2021. Controllable sentence simplification with a unified text-to-text transfer transformer. In *Proceedings of the 14th International Conference on Natural Language Generation*, pages 341–352, Aberdeen, Scotland, UK. Association for Computational Linguistics.

Lucia Specia. 2010. Translating from complex to simplified sentences. In *Computational Processing of the Portuguese Language: 9th International Conference, PROPOR 2010, Porto Alegre, RS, Brazil, April 27-30, 2010. Proceedings 9*, pages 30–39. Springer.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. BLEU is not suitable for the evaluation of text simplification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 738–744, Brussels, Belgium. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting high quality monolingual datasets from web crawl data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

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.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

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.

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

## A GPT Inference Details

We conducted the LLM evaluation with GPT-3.5-TURBO-INSTRUCT engine following recent work

on text simplification (Kew et al., 2023; Feng et al., 2023) and to be in line with our budget. We performed three runs for zero-shot, one-shot, and few-shot calls, always with Nucleus Sampling with a probability of 0.9, a temperature of 1.0, and a maximum output length of 100 tokens. The one-shot setup includes one example each of syntactic simplification, changes in word order, anaphora, and elimination of redundant information. The selected exemplars are in Table 6. The few-shot includes those four exemplars together.

The instruction follows Feng et al. (2023):

> *"Substitua a frase complexa por uma frase simples. Mantenha o mesmo significado, mas torne-a mais simples.*
> *Frase complexa: {original}*
> *Frase Simples: "*[7].

We also conducted preliminary results with the prompt of Kew et al. (2023) but their results were worse than the previous one. In this case, the prompt was:

> *"Reescreva a frase complexa a seguir para que fique mais simples. Você pode trocar as palavras complicadas por sinônimos mais simples, pode tirar informação que não considere útil ou encurtar uma frase, fazendo outras menores. A frase simplificada deve ficar gramaticalmente correta, ter sentido e deve manter as ideias principais, sem mudar o significado."*[8]

## B Human Evaluation Instructions

Figure 4 shows the instructions the volunteers received before answering the questions. The evaluators, linguists in the academic field, provided their evaluations without charging solely to contribute to science.

## C 4-Shot Experiment

Table 6 shows the four manually selected pairs used during the ablation experiments.

## D Assessment of Simplifications

Table 7 and Table 8 contain the simplifications performed by our main model that were worst and best evaluated by the judges, respectively.

[7] In English: *"Replace the complex sentence with a simple sentence. Keep the same meaning but make it simpler.*
*Complex sentence: {original}*
*Simple Sentence: "*

[8] In English: *"Rewrite the complex sentence below to make it simpler. You can swap the complicated words using simpler synonyms, you can remove information that you do not consider useful or shorten a sentence, making others shorter. The simplified sentence must be grammatically correct, make sense and must maintain the main ideas, without changing the meaning."*

| Syle | Simplification |
|---|---|
| Complex | Conforme moradores do bairro, a expressão identificaria um grupo de pichadores. |
| Simple | Os moradores do bairro dizem que a frase identificaria um grupo de pichadores. |
| Complex | According to neighborhood residents, the expression would identify a group of graffiti taggers. |
| Simple | The neighborhood residents say that the phrase would identify a group of graffiti taggers. |
| Complex | Entre os motivos da liderança gaúcha, estão a tradição no cultivo da soja, que hoje representa a maior parte da matéria-prima do biodiesel, e a predominância da agricultura familiar, condição para concessão do selo social. |
| Simple | A tradição na cultura da soja, que hoje representa a maior parte da matéria-prima do biodiesel, e o predomínio da agricultura familiar, condição para conceder o selo social, estão entre os motivos da posição gaúcha de líder. |
| Complex | Among the reasons for the leadership of Rio Grande do Sul are the tradition in soybean cultivation, which today represents the majority of the raw material for biodiesel, and the predominance of family agriculture, a condition for obtaining the social seal. |
| Simple | The tradition in soybean cultivation, which today represents the majority of the raw material for biodiesel, and the predominance of family agriculture, a condition for granting the social seal, are among the reasons for Rio Grande do Sul's leadership position. |
| Complex | E com eles amarrados a coleiras, do alto de uma duna a cerca de 50 metros do mar, tomava chimarrão às 19h de ontem. |
| Simple | Pandolfo tomava chimarrão às 19h de ontem, no alto de um monte de areia, com os poodles amarrados a coleiras. |
| Complex | And with them tied to leashes, from the top of a dune about 50 meters from the sea, he drank mate at 7 p.m. yesterday. |
| Simple | Pandolfo was drinking mate at 7 p.m. yesterday, atop a sand dune, with the poodles tied to leashes. |
| Complex | Numa entrevista coletiva conduzida ontem à noite, os gerentes da Nasa deram o veredicto. |
| Simple | Numa entrevista coletiva ontem à noite, os gerentes da Nasa decidiram. |
| Complex | In a press conference conducted last night, NASA managers delivered the verdict. |
| Simple | In a press conference last night, NASA managers made a decision. |

Table 6: The four manually selected simplifications used as few-shot exemplars, together with their English versions. Note that the translations might not fully express the simplification if they were done in the original translated sentence.

| Complex | Simplification |
|---|---|
| Os aposentos privativos onde Bento XVI estará acomodado de hoje a sexta-feira. | Os aposentos privados onde Bento XVI ficará de hoje à sexta-feira. |
| E, de fato, o azar é todo das formigas, porque essas aranhas, ao chegarem ao formigueiro, se alimentam dos ovos e do trabalho das trabalhadoras desavisadas. | E o azar é todo das formigas, porque, ao chegarem ao formigueiro, elas se alimentam dos ovos e do trabalho das mulheres desprotegidas. |
| O Discovery precisa de reparos antes de voltar para casa. | O Discovery precisa de um recall antes de voltar para casa. |
| O trabalho de Someya e seus colegas está publicado na última edição da revista da Academia de Ciências dos EUA, a "PNAS" (www.pnas.org). | Someya e seus colegas de pesquisa publicaram o trabalho deles na edição mais recente da revista da Academia de Ciências dos Estados Unidos, "PNAS" (www.pnas.org), revista da Academia de Ciências dos Estados Unidos. |

Table 7: Simplifications in which both evaluators gave the minimum score (1)

| Complex | Simplification |
|---|---|
| Foi conduzida então uma caminhada espacial para retirada das duas tiras. | Foi então realizada uma caminhada espacial para a retirada das duas tiras. |
| Na área anexa, uma esplanada para circulação de torcedores, com rampas que levam ao interior do estádio e três torres que abrigarão centro de convenções, hotel e prédio comercial. | Na área anexa, haverá uma esplanada para a circulação de torcedores, com rampas que levam ao interior do estádio e três torres que abrigarão hotel, centro de convenções e prédio comercial. |
| Eles é que emitem luz, que aparentemente pode vir tanto da parte inferior do chapéu quanto do "cabo" ou do cogumelo inteiro. | Eles são os responsáveis por emitir a luz, que pode vir tanto do topo do chapéu como do "cabo" ou cogumelo inteiro. |

Table 8: Simplifications in which both evaluators gave the maximum score (5)

# Avaliação de Textos Simplificados

**Instruções**

Neste formulário há 80 sentenças e, para cada uma, há 4 simplificações dela geradas por sistemas de IA. O objetivo é julgar cada sentença simplificada em uma escala de 1 até 5. É necessário que você leia a sentença de entrada e suas 4 versões simplificadas e depois dê a sua opinião levando em consideração três aspectos:

- Fluência - O texto é bem formado e correto gramaticalmente?
- Simplicidade - O texto é mais simples que o texto de entrada?
- Adequação (preservação de conteúdo) - O texto preserva o conteúdo/sentido do texto de entrada?

1 = Discordo Fortemente, 5 = Concordo Fortemente

**Esclarecimentos:**

- É Válido que a versão simplificada seja composta por mais de uma sentença. Dividir uma sentença grande e complexa em menores pode melhorar a leiturabilidade em algumas situções. Cabe a você julgar se a divisão tornou o texto mais simples e compreensível.
- Diferentes sistemas podem gerar a mesma simplificação
- Fluência deve ser julgada olhando somente o texto simplificado. Em sua avaliação, considere erros gramaticais e/ou de escrita, mas também considere quão natural é o texto.
- Simplicidade e Adequação devem ser julgadas observando a sentença de entrada e a versão simplificada. Julgue se as modificações realizadas mudaram ou não o sentido da sentença original, e se tornaram o texto mais fácil de ser compreendido.
- É provável que a sentença simplificada não contenha todos os detalhes da sentença de entrada. No julgamento da adequação, cabe a você julgar o impacto dessas mudanças no significado do texto
- Julgar a qualidade de simplificações é subjetivo. Cada pessoa tem sua própria opinião sobre o que é fluente, simples e adequado. Por isso, estamos coletando um grande número de respostas para que seja possível estudar a concordância/discordância das avaliações. Por esse motivo, não apresentamos quaisquer exemplos: trata-se de uma forma de evitar que nosso viés de julgamento não afete seus julgamentos pessoais.
- As 80 sentenças foram divididas em 8 grupos para que a avaliação possa ser feita pausadamente. Pede-se que as sentenças dos 8 grupos sejam avaliadas. Apesar de não ser obrigatório, sugere-se começar a partir do grupo 1 e seguir a ordem.

Escolha o grupo abaixo para iniciar a avaliação dos textos desse grupo.
MUITO OBRIGADO!!

Figure 4: Instructions provided to the volunteers preceding human evaluation.