RETUYT-INCO at MLSP 2024: Experiments on Language Simplification using Embeddings, Classifiers and Large Language Models

Ignacio Sastre **Leandro Alfonso** **Facundo Fleitas**

Federico Gil **Andrés Lucas** **Tomás Spoturno**

Santiago Góngora **Aiala Rosá** **Luis Chiruzzo**

Instituto de Computación, Facultad de Ingeniería, Universidad de la República
Montevideo, Uruguay

Abstract

In this paper we present the participation of the RETUYT-INCO team at the BEA-MLSP 2024 shared task. We followed different approaches, from Multilayer Perceptron models with word embeddings to Large Language Models fine-tuned on different datasets: already existing, crowd-annotated, and synthetic. Our best models are based on fine-tuning Mistral-7B, either with a manually annotated dataset or with synthetic data.

1 Introduction

History has shown that technology can mean a step forward for inclusion and social development. For instance, NLP can change how different social groups interact with texts, by automatically adapting texts to the reader's needs and hence improving digital accessibility. One of the many NLP tasks devoted to this objective is *lexical simplification*, where systems are built to replace complex words by simpler ones. This has an immediate impact on language learners and children, but also on people with different types of learning or reading difficulties (Paetzold and Specia, 2016).

The BEA-MLSP 2024 shared task (Shardlow et al., 2024a) proposes an excellent opportunity to explore two problems related to this path: to score how complex a word is in a given context (task 1), and to find simpler substitutes for that word (task 2). The dataset used both as trial and test sets covers 10 different languages: Catalan, English, Filipino, French, German, Italian, Japanese, Portuguese, Sinhala and Spanish (Shardlow et al., 2024b). This dataset was annotated using the MultiLS Framework (North et al., 2024).

In this paper we present the participation of the Uruguayan RETUYT-INCO team at this shared task, describing the approaches followed and the datasets used. The main challenge to solve these tasks is the scarcity of data: only 30 examples for each language were given as trial data, and no training data. We decided to use the trial data as a development set to compare our experiments against each other, and rely on other sources of data (already existing datasets, crowd-sourced, or synthetic).

2 Related Work

Lexical complexity prediction and lexical simplification tasks have been addressed in different challenges in the past. We discuss the most recent ones for each task.

In the SemEval-2021 Task 1: Lexical Complexity Prediction (Shardlow et al., 2021), participants developed systems that, given a word within a sentence, assign it a complexity value on a continuous scale. An extended version of the CompLex Corpus (Shardlow et al., 2020) was used, with 10,800 instances of words and multi-word expressions scored according to their complexity. Deep Learning based systems performed the best, followed closely by feature-based approaches.

The TSAR-2022 Shared Task on Lexical Simplification (Saggion et al., 2022) hosted a shared task on Multilingual Lexical Simplification for English, Portuguese, and Spanish. The participants had to propose simpler substitutes for a complex word in a given context. Some trial examples were provided in each language (10 for English, 10 for Portuguese, and 12 for Spanish). The best results were obtained by approaches based on masked language models.

3 Approaches

In this section we detail the five different approaches followed. We experimented with static word embeddings, contextual embeddings, fine-tuning Mistral 7B on synthetic data, crowd-sourced data and existing data, and also with the Groq platform. We describe each of them next.

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 618–626

3.1 Word Embeddings + Frequency Baselines

We created baseline approaches to the two tasks based on the use of word embeddings and word frequencies. In these baselines we prioritized using collections of embeddings and word frequency lists that were collected in the same way for all the languages in the task, so we used the Polyglot (Al-Rfou et al., 2013) word embedding collections, and the word frequency datasets collected from subtitles by Hermit Dave[1]. These resources are available for many languages, including all the languages in the shared task with the exception of the specific Filipino variety of the Tagalog language. In that case we used the corresponding resources for Tagalog, even if they could have some differences.

The approach for task 1 (Complex Word Prediction) is non-contextual, as no information from the context sentence is used: we take the 10 closest words to the target word in the embeddings collection, then use the frequency as a proxy to how complex a word is, assuming that more frequent words are simpler than less frequent ones. We sort the 10 closest words plus the target word by frequency and estimate the complexity of the target word as the relative position in this list, being 0 if it is the most frequent of the set and 1 if it is the least frequent.

The approach for task 2 (Lexical Simplification) was similar: finding the 10 most similar words to the target in the embeddings set, and sorting them by frequency. Besides the Polyglot embeddings and subtitle word frequency lists, for task 2 we also tried variants of this baseline approach using bigger and richer word embedding collections and frequency lists. For Spanish we used the `SBW-vectors-300-min5` embeddings[2] trained with the Spanish Billion Word Corpus[3]; for English the `googlenewsvectors` collection[4], and for Portuguese a word2vec collection trained from the ConLL17 corpus[5].

We also used other word frequency lists: for Spanish we used the Wiktionary Spanish frequency list[6], for English the Kaggle English Word Frequency dataset[7] compiled from the Google Web Trillion Word Corpus, and for Portuguese the frequency counts of the `wordfreq` library[8]. Another variant of this approach was sorting the replacement candidates by the distance with respect to the target word, without using word frequencies at all.

Besides these static word embedding approaches, we also tried with pre-trained contextual word embeddings such as BERT (Devlin et al., 2019) and ALBERT (Lan et al., 2019). We encode the context sentence and substitute the target word with the `[MASK]` token to obtain the 10 most probable replacements, that could be sorted either by probability or with word frequencies. In this case we used the BETO (Cañete et al., 2020) models `dccuchile/bert-base-spanish-wwm-cased` and `dccuchile/albert-xxlarge-spanish` for Spanish and HuggingFace models `google-bert/bert-large-cased` and `albert/albert-xlarge-v2` for English.

3.2 Fine-tuning Mistral 7B

This section presents two different approaches to fine-tuning an LLM to solve these tasks.

3.2.1 Fine-tuning on a Synthetic Dataset from Claude 3

It is well known that larger and more complex LLMs like the GPT family or Claude 3 Opus LLM from Anthropic[9] generally have good results in many NLP tasks. However, these are closed models, and we wanted to try if it was possible to at least distill some of their capabilities into a smaller model that is more resource-efficient, open and accessible to run in our available environment. To achieve this, and to alleviate the data scarcity problem, given that preliminary experiments with Claude 3 using the trial data showed promising results in a zero-shot scenario, we built a synthetic dataset using this LLM.

Generation of the synthetic data

Figure 1 shows a diagram of the synthetic dataset generation process. The complete prompts for each step can be found in Appendix C, while a comprehensive explanation of the entire process is provided in Appendix A. Below is a concise overview.

[1] https://github.com/hermitdave/FrequencyWords/

[2] https://github.com/dccuchile/spanish-word-embeddings

[3] https://crscardellino.github.io/SBWCE/

[4] https://www.kaggle.com/datasets/adarshsng/googlenewsvectors

[5] http://vectors.nlpl.eu/repository/

[6] https://en.wiktionary.org/wiki/User:\Matthias_Buchmeier#Spanish_frequency_list

[7] https://www.kaggle.com/datasets/rtatman/\english-word-frequency

[8] https://pypi.org/project/wordfreq/

[9] https://www.anthropic.com/news/claude-3-family

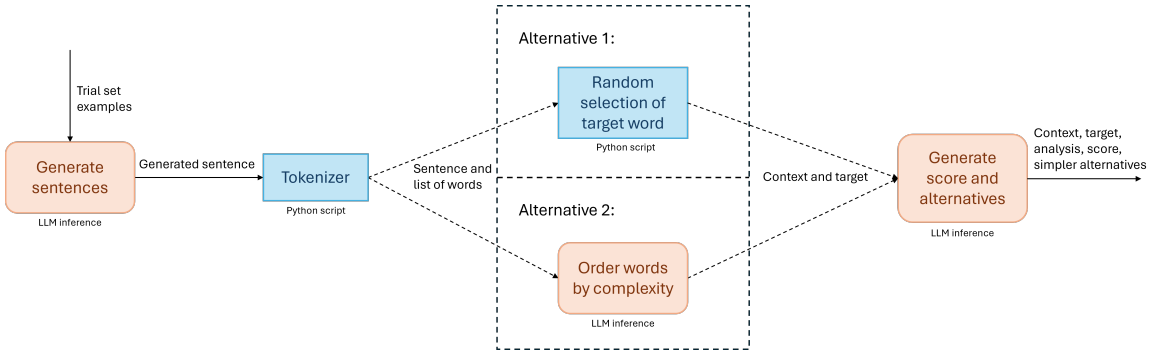

Figure 1: Diagram illustrating the process used to generate the synthetic dataset.

The process starts prompting the Claude model multiple times, including four random sentences from the trial dataset (i.e. following a few-shot strategy (Brown et al., 2020)) to get 250-500 sentences for each language. Then, each generated sentence is processed to generate different (context,target) pairs, as needed for the task 2 of the shared-task. Finally we generate the complexity score (1 to 5) and simpler alternatives for a given (context, target) pair, as needed for the task 1. In this final step we prompt the Claude model with an example of a word with a score of 1 and another one with a score of 5. Additionally we also include a Chain-of-Thought analysis (CoT, Wei et al. (2022)) to improve the performance of the model.

Each row of the resulting dataset consists of the context sentence, the target word, the analysis (CoT), the complexity score and the simpler alternatives. We elaborated a dataset of 2211 examples: 961 in Spanish, 750 in English and 500 in Portuguese. Our decision to focus on these three languages was due to time constraints and also because these are languages that we are familiar with, so we were able to check the overall quality of the synthetic text.

Fine-tuning details In order to fine-tune a smaller model for both task 1 and 2, each example of the dataset is transformed into a string which is a concatenation of the context sentence, the target word, the complexity score, and the simpler alternatives. Each of the parts is separated using XML tags, as can be observed in appendix C.4.

We tried adding the analysis (CoT) before the score when using the Spanish dataset. Table 3 and 4 in appendix B show the results for all the combinations of these techniques (CoT and SC). As can be seen, using a variation of SC without CoT gave the best results. Because of this, we decided to use this method for the rest of the languages.

These formatted examples are utilized for fine-tuning `Mistral 7B Instruct v0.2`[10]. Due to resource constraints, the model is 4-bit quantized and is fine-tuned using the Low-Rank Adaptation (LoRA) method (Hu et al., 2022).

Three different models were trained this way: using only the Spanish portion of the dataset, using Spanish and English, and using the whole dataset (Spanish, English and Portuguese). As a consequence, these were also the languages we focused more on evaluating. We also tried the last model on the Catalan language, given the similarity with Spanish and for testing the generalization capabilities of the fine-tuned model.

When doing inference with these models, a variation of the Self-Consistency technique (SC, Wang et al. (2023)) was employed. For the Lexical Simplification task, inference is conducted 10 times per example with a temperature of 0.7, resulting in up to 30 simpler alternatives with repetitions. These are then counted and arranged in order or frequency, with the most frequent ones appearing first, and any repeated occurrences are eliminated.

For the Lexical Complexity Prediction task, the prompt is structured such that the immediate next token represents the score, so only one token is sampled.This process is performed concurrently 100 times (within one batch) with temperature set to 1. The average score is then computed and normalized to the 0-1 range.

3.2.2 Fine-tuning on an existing English Dataset

We also tried another approach to fine-tune Mistral by curating the LCP2021 (Shardlow et al., 2020) dataset for English[11], recommended by the organiz-

[10] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.2

[11] https://github.com/MMU-TDMLab/CompLex

Experiment	Lang.	Pearson	Spearman	MAE	MSE	RMSE	r2
LoRA Mistral-7B, LCP2021	en	0.8061	0.7596	0.1405	0.0252	0.1587	0.3154
MLP with RoBERTa embeddings	en	0.5502	0.4923	0.1561	0.0328	0.1811	0.1062
LoRA Mistral7B, SC en-es dataset	en	0.7599	0.7406	0.1867	0.0433	0.2081	-0.1796
MLP with BETO embeddings	es	0.3126	0.2369	0.1433	0.0349	0.1868	0.0131
LoRA Mistral7B, SC en-es-pt dataset	es	0.6641	0.6547	0.1311	0.0254	0.1594	0.2808
LoRA Mistral7B, SC en-es-pt dataset	pt	0.6772	0.7121	0.2067	0.0557	0.2360	-1.5487
LoRA Mistral7B, SC en-es-pt dataset	ca	0.3948	0.3862	0.199	0.0569	0.2385	-1.3972
LoRA Mistral7B, SC en-es-pt dataset	all	0.4858	0.4892	0.2089	0.0623	0.2496	-0.6746

Table 1: Results for Task 1 over the test data.

Experiment	Lang.	MAP@1/POT@1	MAP@3	MAP@5	MAP@10	Pot@3	Pot@5	Pot@10	Acc@1@tg1	Acc@2@tg1	Acc@3@tg1
LoRA Mistral7B, SC es dataset	es	0.6138	0.4124	0.2980	0.1595	0.7875	0.8246	0.8532	0.3288	0.4435	0.4839
Groq Prompting + CREA Freq	es	0.3136	0.2412	0.1650	0.089	0.5233	0.556	0.5893	0.1045	0.1905	0.2698
Baseline ALBERT + Distance	es	0.2563	0.154	0.1193	0.0731	0.4097	0.4907	0.5986	0.1079	0.1551	0.1854
LoRA Mistral7B, SC en-es dataset	en	0.5789	0.3718	0.2542	0.1355	0.7666	0.8087	0.8578	0.3438	0.4701	0.5526
LoRA Mistral7B, SC en-es-pt dataset	en	0.5947	0.3832	0.2634	0.1394	0.7824	0.828	0.8543	0.3789	0.5105	0.5701
Baseline ALBERT + Distance	en	0.1596	0.0920	0.0629	0.0379	0.2771	0.3438	0.4649	0.0824	0.1263	0.1561
LoRA Mistral7B, SC en-es dataset	pt	0.4021	0.2094	0.1360	0.0712	0.5784	0.6137	0.6631	0.2768	0.3844	0.4514
LoRA Mistral7B, SC en-es-pt dataset	pt	0.3756	0.2062	0.1336	0.0695	0.5414	0.5855	0.6172	0.2592	0.3562	0.4197
Baseline Static Embeddings + Word Freq	pt	0.0670	0.0380	0.0251	0.0136	0.1604	0.1922	0.2204	0.0582	0.0934	0.1358
LoRA Mistral7B, SC dataset en-es	all	0.3925	0.5233	0.5560	0.5893	0.2412	0.1650	0.0890	0.2156	0.2912	0.3324
LoRA Mistral7B, SC dataset en-es-pt	all	0.3818	0.2351	0.1608	0.0862	0.5091	0.5436	0.5772	0.2074	0.2851	0.3216

Table 2: Results for Task 2 over the test data.

ers, formatting this dataset to align with the task requirements. We fine-tuned the `Mistral-7B-v0.1` model, using a customized LoRA (Hu et al., 2022), choosing specific configurations to disable cache usage during training and to adapt the tokenizer for the corresponding task.

3.3 BERT and MLP models

As a totally different approach for task 1, we tried to use BERT embeddings as a text-representation input for Multilayer Perceptron models. For English we used the original BERT (Devlin et al., 2019), while for Spanish we used BETO (Cañete et al., 2020).

3.3.1 English (BERT)

To fine-tune the English BERT model we used the previously mentioned LCP2021 dataset. We trained for over 10 epochs with validation splits to monitor overfitting and batch processing for efficiency.

3.3.2 Spanish (BETO)

We had an additional problem when trying to fine-tune the BETO model, because there was not a Spanish dataset that was similar to LCP2021. The most similar set we found was the EASIER_CORPUS (Alarcon et al., 2023) dataset, but it only categorizes words in a binary way between easy and complex, and in this case we needed a more fine-grained distinction.

We first tried to generate synthetic text in Spanish using `gpt-3.5-turbo-0125`. In order to get data as balanced as possible, the prompts for the API were designed to produce sentences of two complexity levels, with a 50% probability each.

Then we gathered crowd-sourced data using a public website developed by us. This website allowed users to rate the complexity of words within sentences on a scale from 1 to 5. First we included only the synthetic sentences, and later on we also added the EASIER_CORPUS sentences, trying to include a wider range of linguistic contexts. We got approximately 2300 entries over a seven-day period[12].

After normalizing the scores of the whole dataset to match the expected score ranges, we fed BETO with all this Spanish text.

3.4 Use of pretrained models in Groq

As a final experiment for task 2 in Spanish, we used the Groq platform[13] to leverage the prompting capabilities of several pretrained LLMs: LLAMA (`llama2-70b-4096`), GEMMA (`gemma-7B-it`), and MIXTRAL (`mixtral-8x7B-32768`). We created a pipeline that prompts each of these models into giving simpler alternatives to a word in the context of a sentence, following a one-shot mechanism to illustrate the expected response. Using the Groq API, we collected the responses of the three models, combined them and used the word frequencies of the CREA corpus[14] to sort the possible answers.

[12]This manually annotated dataset will be published.

[13]https://groq.com/

[14]https://www.rae.es/banco-de-datos/crea

4 Results

In the appendix B we include tables 3 and 4, which show the results of our methods over the trial data. We used those preliminary results to choose which submissions to send to the competition, trying to keep the most promising systems but also a mix of different approaches. The experiments selected for submission are underlined in the tables. Tables 1 and 2 show the results of the submitted systems over the test set.

5 Conclusions

We presented a series of experiments for solving the Complex Word Prediction and Lexical Simplification tasks, ranging from simpler non-contextual static embeddings baselines, to more advanced fine-tuning of LLMs. The most important challenge in these tasks was the data scarcity, and because of this we had to use different resources like synthetic datasets, adapting existing datasets, or crowd-annotating new data. Our best approaches for both tasks where achieved by fine-tuning Mistral 7B, either with synthetic data or with already existing resources.

6 Limitations

Due to time constraints there were many experiments and combinations that we did not try, being the most salient one the fine-tuning of Mistral 7B with the manually annotated data collected through crowd-sourcing. We look forward to complete this experiment in the future.

References

Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual nlp. *arXiv preprint arXiv:1307.1662*.

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2023. Easier corpus: A lexical simplification resource for people with cognitive impairments. *Plos one*, 18(4):e0283622.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared 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 M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher 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. *Preprint*, arXiv:2005.14165.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

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.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. Multils: A multi-task lexical simplification framework. *arXiv preprint arXiv:2402.14972*.

Gustavo Paetzold and Lucia Specia. 2016. Benchmarking lexical simplification systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3074–3080, Portorož, Slovenia. European Language Resources Association (ELRA).

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2022. Findings of the TSAR-2022 shared task on multilingual lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 271–283, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024a. The BEA 2024 Shared Task on the

Multilingual Lexical Simplification Pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Huelsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024b. An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI)*.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of the 1st Workshop on Tools and Resources to Empower People with REAding DIfficulties (READI)*, pages 57–62, Marseille, France. European Language Resources Association.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. SemEval-2021 task 1: Lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 1–16, Online. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

A Synthetic data generation using Claude 3

In section 3.2.1 we briefly described the strategy to generate synthetic data, summarized by figure 1. In this appendix we will describe this pipeline in detail.

(1) Generation of synthetic sentences

A prompt was designed using four different sentence examples from the trial dataset as few-shot examples (Brown et al., 2020). To increase the diversity of the generated sentences and avoid overfitting, the examples are selected at random, so the prompt is not always the same. In some cases, to enhance variety, an additional phrase is added to the prompt asking the model to generate sentences containing at least one complex word.

The inference of the model is done multiple times with temperature or top-p set to 1 for maximizing diversity, and 250-500 sentences are created for each language, half of them with the complex word restriction added to the prompt.

(2) Selection of the target word

Given a generated sentence, we need to select a target word to generate (context, target) pairs, so we first tokenize the sentence to obtain a list of candidate words. Our simple tokenization means lower-casing, separating by spaces and removing punctuation and stopwords from NLTK (Bird and Loper, 2004)).

We explored two methods for selecting the words from the list of candadates. One approach was two select two or three words at random, not taking into account the complexity of the words. The other was to order the candidate words by decreasing complexity by prompting the LLM for this task, and then selecting the most complex and least complex words as target words.

(3) Generation of the complexity score and simpler alternatives

The prompt used to generate the complexity score and simpler alternatives for a given (context, target) pair consists of instructions for the model to generate the following three parts: a Chain-of-Thought analysis (CoT, Wei et al. (2022)) of the complexity of the target word in the given context sentence; a 1 to 5 complexity score for the target word, following the annotation guidelines used for the trial dataset; a list of at most three simpler alternatives for the target word. If no simpler alternatives exist, the model should return the same target word. Two handcrafted score examples are added to the prompt: one with a score of 1 and the other with a score of 5.

Each row of the resulting dataset consists of the context sentence, the target word, the analysis (CoT), the complexity score and the simpler alternatives. We elaborated a dataset of 2212 examples: 961 in Spanish, 750 in English and 500 in Portuguese.

Experiment	Lang.	Pearson	Spearman	MAE	MSE	RMSE	r2
BERT emeddings into MLP	en	0.3813	0.4331	0.2084	0.0543	0.2330	-0.3981
LoRA Mistral7B, LCP2021	en	0.8640	0.8574	0.1678	0.0330	0.1816	0.1514
MLP with RoBERTa embeddings	en	0.3957	0.2948	0.1607	0.0375	0.1936	0.0333
LoRA Mistral7B, SC en-es dataset	en	0.7363	0.7126	0.2243	0.0591	0.2431	-0.5199
MLP with BETO embeddings	es	0.4528	0.3925	0.2079	0.0622	0.2493	-0.1815
LoRA Mistral7B, es dataset	es	0.3892	0.3592	0.1942	0.0557	0.2360	-0.0570
LoRA Mistral7B, CoT es dataset	es	0.6355	0.6282	0.1458	0.0349	0.1867	0.3385
LoRA Mistral7B, SC es dataset	es	0.6461	0.6260	0.1483	0.0307	0.1754	0.4164
LoRA Mistral7B, SC-CoT es dataset	es	0.6102	0.6708	0.1575	0.0357	0.1890	0.3219
LoRA Mistral7B, SC en-es dataset	es	0.7283	0.7522	0.1337	0.0262	0.1618	0.5030
LoRA Mistral7B, SC en-es-pt dataset	es	0.7369	0.7180	0.1351	0.0259	0.1608	0.5090
LoRA Mistral7B, SC en-es-pt dataset	pt	0.7410	0.7754	0.1541	0.0415	0.2036	-0.5839
LoRA Mistral7B, SC en-es-pt dataset	ca	0.5460	0.5624	0.1299	0.0276	0.1662	-0.8219
LoRA Mistral7B, SC en-es-pt dataset	all	0.5301	0.5427	0.2060	0.0618	0.2486	-0.3930
Baseline Polyglot Embeddings + Word Freq	all	0.2106	0.2014	0.3711	0.2130	0.4615	-3.8008

Table 3: Results for Task 1 over the trial data. The underlined experiments are the ones we chose to send as submissions for the shared task.

Experiment	Lang.	MAP@1/POT@1	MAP@3	MAP@5	MAP@10	Pot@3	Pot@5	Pot@10	Acc@1@tg1	Acc@2@tg1	Acc@3@tg1
LoRA Mistral7B, es dataset	es	0.7666	0.5240	0.3144	0.1572	0.8666	0.8666	0.8666	0.5333	0.6000	0.6000
LoRA Mistral7B, CoT es dataset	es	0.6666	0.4722	0.2833	0.1416	0.8333	0.8333	0.8333	0.4000	0.4666	0.5333
LoRA Mistral7B, SC es dataset	es	0.9333	0.6000	0.4173	0.2298	0.9333	0.9666	1.000	0.5666	0.6666	0.7000
LoRA Mistral7B, SC-CoT es dataset	es	0.8000	0.5944	0.4510	0.2549	0.9000	0.9000	0.9333	0.4333	0.6333	0.6666
LoRA Mistral7B, SC en-es dataset	es	0.8666	0.5925	0.4405	0.2305	0.9333	0.9666	0.9666	0.5666	0.6333	0.7333
LoRA Mistral7B, SC en-es-pt dataset	es	0.8666	0.6333	0.4736	0.2658	0.9333	0.9333	0.9666	0.5000	0.6666	0.7333
Groq Prompting + CREA Freq	es	0.4666	0.2888	0.2213	0.1386	0.7000	0.7666	0.9333	0.2000	0.3666	0.4333
Baseline Static Embeddings + Distance	es	0.3333	0.1759	0.1355	0.0842	0.5000	0.6000	0.6666	0.1666	0.2333	0.2666
Baseline ALBERT + Distance	es	0.3333	0.2296	0.1714	0.1047	0.5333	0.5666	0.6333	0.1666	0.3000	0.3666
Baseline Static Embeddings + Word Freq	es	0.3000	0.2129	0.1511	0.0929	0.6000	0.6666	0.6666	0.1333	0.2333	0.3000
Baseline BERT + Distance	es	0.2666	0.2222	0.1810	0.1022	0.5000	0.5333	0.5666	0.1000	0.2333	0.2666
Baseline ALBERT + Word Freq	es	0.2000	0.1055	0.0730	0.0547	0.3666	0.4666	0.5666	0.1000	0.1666	0.2000
Baseline BERT + Word Freq	es	0.1333	0.0962	0.0677	0.0479	0.2333	0.3333	0.5333	0.0666	0.1000	0.1333
LoRA Mistral7B, SC en-es dataset	en	0.5666	0.3462	0.2371	0.1267	0.8333	0.8333	0.8333	0.4000	0.5666	0.7333
LoRA Mistral7B, SC en-es-pt dataset	en	0.5000	0.3166	0.2326	0.1201	0.7333	0.8333	0.8333	0.3666	0.5666	0.6333
Baseline Static Embeddings + Word Freq	en	0.1666	0.1074	0.0831	0.0480	0.3666	0.4666	0.5000	0.1000	0.2333	0.3000
Baseline BERT + Distance	en	0.1333	0.0981	0.0832	0.0462	0.3666	0.4666	0.5333	0.1000	0.2000	0.3000
Baseline Albert + Distance	en	0.1333	0.0814	0.0675	0.0387	0.3666	0.4666	0.5000	0.0666	0.1333	0.3000
Baseline Statitc Embeddings + Distance	en	0.0666	0.0574	0.0451	0.026	0.2333	0.3000	0.4000	0.0333	0.0666	0.2000
Baseline ALBERT + Word Freq	en	0.0666	0.0314	0.0245	0.0167	0.1333	0.2666	0.3666	0.0333	0.0666	0.1000
Baseline BERT + Word Freq	en	0.0333	0.0222	0.0183	0.0125	0.1333	0.1666	0.3666	0.000	0.000	0.0333
LoRA Mistral7B, SC en-es dataset	pt	0.3000	0.1925	0.1278	0.0695	0.6000	0.6666	0.7000	0.2333	0.4333	0.5000
LoRA Mistral7B, SC en-es-pt dataset	pt	0.3000	0.1759	0.1258	0.0646	0.6333	0.6333	0.6666	0.2333	0.3666	0.4333
Baseline Static Embeddings + Word Freq	pt	0.1333	0.0555	0.0333	0.0175	0.2000	0.2000	0.2666	0.1000	0.1333	0.1333
Baseline Static Embeddings + Distance	pt	0.0333	0.0166	0.0130	0.0078	0.0666	0.1333	0.2333	0.000	0.0333	0.0333
LoRA Mistral7B, SC dataset en-es	all	0.4066	0.2199	0.1319	0.0659	0.5300	0.5300	0.5300	0.2666	0.3333	0.3600
LoRA Mistral7B, SC dataset en-es-pt	all	0.4066	0.2257	0.1354	0.0677	0.4866	0.4866	0.4866	0.2466	0.3166	0.3566
Baseline Polyglot Embeddings + Word Freq	all	0.1133	0.0562	0.0384	0.022	0.1766	0.2133	0.2833	0.0500	0.0800	0.0866

Table 4: Results for Task 2 over the trial data. The underlined experiments are the ones we chose to send as submissions for the shared task.

B Results over the trial data

Tables 3 and 4 show the results of our methods over the trial data.

C Prompts

C.1 Generation of synthetic sentences prompt

System prompt: not used.

Message with role user:

`Your task is to create new sentences in` **`{language}`**.

`Here are some examples of the type of sentences we expect:` **`{few_shot}`**

`Try to write similar sentences to the examples provided.` **`{complex_sentence}`** `You should write` **`{n}`** `different and diverse sentences, each in a new line. No other text should be written.`

Where:

1. **language** is the expected language of the sentences. For example: Spanish.

2. **few_shot** is a list of four examples of sentences from the trial dataset, separated by new-

lines.

3. **complex_sentence** is either an empty string or the sentence: `It is essential for the new sentences to use some extremely complex words.`

4. **n** is the amount of sentences the model should create in one run.

C.2 Order candidate words by complexity prompt

System prompt:

```
You are an annotator for a dataset of lexical simplification.

<task_description>
Given a context sentence and an a list of words from that context, your task is to order these words by decreasing complexity. The most complex word should go first, and the least complex word should go last.
</task_description>

<answer_format>
Your answer must follow the following format: Each word should be written in a new line. Nothing else should be written.
</answer_format>
```

Message with role user:

```
<context>
{context}
</context>

<words>
{words}
</words>
```

Where:

1. **context** is the sentence where the words appear.

2. **words** is the candidate words list separated by newlines.

C.3 Complexity score and simpler alternatives prompt

System prompt:

```
You are an annotator for a dataset of lexical simplification.

<task>
Given a context sentence and an a identified (whole-word) target to be evaluated, your task is to annotate the following information:
1) An step-by-step analysis of the target in the context to justify you following decisions.
2) A complexity score for the target in its context on a scale of 1 (easy) to 5 (difficult). This number should come as a consequence of the analyisis.
3) A list of no more than 3 simpler alternatives for the target, or the target itself if no simpler alternative can be found. The words should appear in increasing order of complexity. Do not add the target if simpler alternatives exist.
</task>

<considerations>
- The analysis should always have language learners in mind, not just native speakers.
- It is important to make decisions based on how other words could be evaluated, to make a grounded decision.
- If there are no simpler alternatives, the alternatives should only be the word itself.
</considerations>

<expected_answer>
Your answer must follow the following format:
- Inside XML tags <analysis></analisis> you must write (1) as free form text in english (regardless of source language). Remember to write in english.
- Inside XML tags <score></score> you must write (2) as one of the following numbers: 1, 2, 3, 4 or 5. Write only the number, without periods or text.
- Inside XML tags <simpler_alternatives> </simpler_alternatives> you must write (3) as a list of words separated by commas. No newlines between words should be used.
```

```
</expected_answer>

<score_examples>
Example of a score of 5:
 - Context: {example_context_1}
 - Target: {example_target_5}
Example of a score of 1:
 - Context: {example_context_1}
 - Target: {example_target_1}
</score_examples>
```

Message with role user:

```
<context>
{context}
</context>

<target>
{target}
</target>
```

Where:

1. **example_context_5** and **example_target_5** correspond to a hand-crafted score 5 example of a sentence and a target word respectively. Varies depending on the language.
2. **example_context_1** and **example_target_1** correspond to a hand-crafted score 1 example of a sentence and a target word respectively. Varies depending on the language.
3. **context** is the context sentence where the target word occurs.
4. **target** is the target word to evaluate.

C.4 Fine-tuning prompt format

The following is the prompt format used for the fine-tuning examples:

```
<context>
{context}
</context>
<target>
{target}
</target>
<score>
{score}
</score>
<simpler_alternatives>
{simpler_alternatives}
</simpler_alternatives>
```

Where:

1. **context** is the context sentence where the target word occurs.
2. **target** is the target word to evaluate.
3. **score** is a number between 1 and 5 corresponding to the complexity score.
4. **simple_alternatives** is a list of simpler alternatives for the target word, separated by commas.