MultiLS: An End-to-End Lexical Simplification Framework

Kai North[1], **Tharindu Ranasinghe**[2], **Matthew Shardlow**[3], **Marcos Zampieri**[1]
[1]George Mason University, USA
[2]Lancaster University, UK
[3]Manchester Metropolitan University, UK
knorth8@gmu.edu

Abstract

Lexical Simplification (LS) automatically replaces difficult to read words for easier alternatives while preserving a sentence's original meaning. Several datasets exist for LS and each of them specialize in one or two sub-tasks within the LS pipeline. However, as of this moment, no single LS dataset has been developed that covers all LS sub-tasks. We present MultiLS, the first LS framework that allows for the creation of a multi-task LS dataset. We also present MultiLS-PT, the first dataset created using the MultiLS framework. We demonstrate the potential of MultiLS-PT by carrying out all LS sub-tasks of (1) lexical complexity prediction (LCP), (2) substitute generation, and (3) substitute ranking for Portuguese.

1 Introduction

Despite the importance and growing popularity of LS (Paetzold and Specia, 2016b; Yimam et al., 2018; Shardlow et al., 2021a; Saggion et al., 2022), all publicly available datasets, regardless of language, fail to cover all sub-tasks within the LS pipeline: lexical complexity prediction (LCP), substitute generation (SG), selection (SS), and ranking (SR) as depicted in Figure 1.

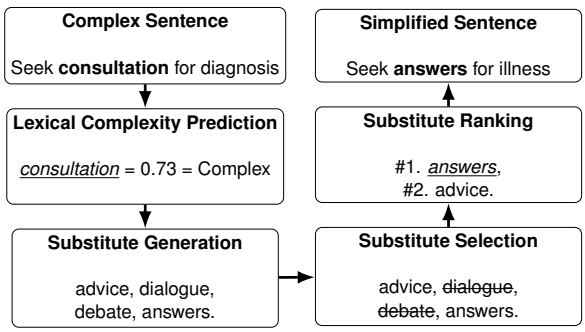

Figure 1: LS Pipeline. Example shows LS pipeline applied within the biomedical domain. Original figure adapted from (Paetzold and Specia, 2015)

End-to-end LS frameworks (McCarthy and Navigli, 2007; Specia et al., 2012; Horn et al., 2014; Hartmann and Aluísio, 2020; Saggion et al., 2022) have collected gold simplifications needed for SG, SS, and SR, but have excluded LCP. In contrast, lexical complexity datasets (Maddela and Xu, 2018; Shardlow et al., 2020) refrained from collecting gold simplifications. Each of these LS frameworks also annotated different target words meaning that their subsequent datasets cannot be combined to provide all the necessary information for LS.

In this paper, we introduce MultiLS, the first multi-purpose end-to-end framework for the creation of all-in-one LS datasets by providing target words with lexical complexity values required for LCP and gold candidate simplifications needed for SG, SS, and SR. MultiLS is an extensible framework allowing the creation of datasets in various languages. We use MultiLS to create MultiLS-PT, the first multi-task and multi-genre dataset for Portuguese LS. Portuguese is one of the ten most spoken languages in the world with over 250 million speakers (Eberhard et al., 2023). Many countries where Portuguese is spoken (e.g., Angola, Brazil, Mozambique) have low literacy rates. We chose to include texts from the Brazilian variety in MultiLS-PT as this is the most widely-spoken variety of Portuguese. While Brazil is one of the largest economies in the world, a large part of its population are either illiterate or functionally illiterate worsening existing socio-economic challenges (Ireland, 2008). As such, there is ample motivation for the development of assistive reading technologies for Portuguese.

The main contributions of this paper are:

1. **MultiLS**: the first multi-purpose framework for the full training and evaluation of all LS sub-tasks (Sections 2 to 3).
2. **MultiLS-PT**: the first Portuguese multi-genre dataset for LS to contain both continuous complexity values and ranked gold simplifications (Section 4).

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 1–11

3. **Evaluation**: the performance of multiple state-of-the-art models for LCP, substitute generation and ranking (Sections 5 to 7).

2 Related Work

Complexity Prediction The first-step within the LS pipeline is the identification of complex words (North et al., 2022d). There are two approaches to this task. Complex Word Identification (CWI), a binary classification task which assigns each target word with a non-complex (0) or complex (1) label (Paetzold and Specia, 2016b; Zampieri et al., 2017). LCP is a regression-based task that assigns a complexity value on a continuum often using a Likert-scale, including such labels as very simple (0), neutral (0.5), to very complex (1) (Shardlow et al., 2020). Words that have an assigned complexity value substantially greater than 0.5 are considered to be complex words, such as the word "*consultation*" within Figure 1. LCP datasets have employed the use of human annotators to assign gold complexity values (Horn et al., 2014; Paetzold and Specia, 2016b; Yimam et al., 2018).

Substitute Generation and Selection SG is the second-step within the LS pipeline and it aims to produce a pre-defined number: k candidate substitutions that are easier to understand than the original complex word while persevering its meaning (North et al., 2023b). SS filters these generated candidates to find the best possible simplification, commonly referred to as the top-k candidate substitution. For example, given the sentence: "*Seek consultation about your diagnosis*", and the target word: "*consultation*" within Figure 1, SG would produce k candidate substitutions, such as "*advice*", "*dialogue*", "*debate*", and "*answers*". SS then removes those generated candidates that are more complex, semantically dissimilar, or do not fit into the provided context resulting in the *top-k* candidate substitutions: "*advice*" and "*answers*". While SG and SS datasets provide gold candidate substitutions, these datasets are independent of CWI and LCP as they do not include annotated complexity values per target word. Examples of SG and SS datasets include the ALEXSIS datasets for English, Spanish, and Portuguese (Saggion et al., 2022; Ferres and Saggion, 2022; North et al., 2022b) and SIMPLEX-PB 3.0 for Portuguese (Hartmann and Aluísio, 2020). These datasets, however, do not include complexity values required for LCP.

Substitute Ranking SR is the final step within the LS pipeline and it sorts candidate substitutions from the most to the least appropriate simplifications. It arranges candidate substitutions based on their complexity and their semantic similarity to the target word and context (North et al., 2023b).The example shown in Figure 1 ranks "*answers*" as being a more appropriate simplification than "*advice*" for the target word "*consultation*". This may, in part, be due to "*answers*" having a higher frequency within a reference corpus or being more frequent within a training set. Alternatively, "*answers*" may have a lower age of acquisition, higher familiarity score, or even concreteness (abstractness) rating (North et al., 2022d).

End-to-End Frameworks The few previous end-to-end LS frameworks have focused on substitute generation, selection and ranking and not on LCP. In fact, traditional notions of LS consider the identification of complex words a precursor and a separate task to LS (Paetzold and Specia, 2017). BenchLS (Paetzold and Specia, 2016c) provided a suitable framework for the training and evaluation of substitute generation to ranking. BenchLS (Paetzold and Specia, 2016c) contains sentences, target words, and several candidate substitutions ranked per their simplicity, but does not supply the continuous complexity values needed for LCP. PLUMBErr (Paetzold and Specia, 2016a), an automatic error identification framework for LS, demonstrated its potential by assessing several LS systems that conducted CWI alongside all other LS sub-tasks. Nevertheless, its CWI component was trained on a dataset different from that used to evaluate its overall performance. FLELex (Tack et al., 2016) caters for LCP by aligning two datasets of authentic and simplified texts and providing continuous complexity ratings for each target word. However, only a portion of their target words were labeled with a maximum of one candidate substitution per target word limiting its usefulness.

3 MultiLS Framework

As discussed in the last section, LS datasets often have a narrow specialization focusing on one or two tasks. They only include lexical complexity values, candidate substitutions, or candidate features, restricting their use to either LCP or substitute generation, selection, or ranking (Table 1). Unlike previous frameworks, the MulitLex framework supplies all the necessary data required for

		Original Datasets (English)				MultiLS Framework (New MultiLS-PT Dataset)			
						Step 1→	*Step 2→*	*Step 3→*	*Step 4*
T.	D.	Token	Context (Sentence)	Val.	Substitutions	Selection	New Context (Sentence)	New Val.	New Substitutions
Task 1: LCP	**CompLex**	colleagues	pointed out colleagues	0.26	–	colegas	controlado por colegas	0.13	amigos (friends),.
		uncertainties	uncertainties in the	0.37	–	incertezas	influenciada por incertezas	0.08	dúvidas (doubts),.
		gentiles	teacher of the gentiles	0.26	–	gentios	doutor dos gentios	0.46	multidão (crowd),.
		prophet	raise up a prophet	0.27	–	profeta	um profeta semelhante	0.21	mensageiro (messenger),.
		maximum	a maximum of two	0.14	–	máximo	máximo corrigido	0.32	extremo (extreme),.
Tasks 2-3: SG & SS	**ALEXSIS-EN**	observers	the number of observers	–	watchers, spectators,.	observadores	observadores que tiveram	0.19	examinadores (examiners),.
		authorities	assistance to authorities	–	officials, powers,.	autoridades	alegando que as autoridades	0.23	forças (forces),.
		condolences	sincere condolences to	–	sympathy, comfort,.	condolências	suas condolências pedindos	0.21	compaixão (compassion),.
		regime	between Assad's regime	–	government, rule,.	regime	aregime do presidente	0.11	governo (government),.
		monitoring	it was monitoring the	–	watching, observing,.	monitoramento	sistema de monitoramento	0.32	acompanhamento,.
	ALEXSIS+	criteria	meet the criteria	–	requirements, standards,.	critério	critério de visão pública	0.24	normas (standards),.
		pledges	the agreement pledges	–	promises, guarantees,.	promessas	faz promessas e	0.11	compromissos,.
		acquisition	the acquisition announced	–	transaction, purchase,.	aquisição	local de aquisição	0.33	obtenção (obtaining),.
		residence	the residence next door	–	house, apartment,.	residência	tenham residência habitual	0.17	casa (house),.
		inclusion	ensure the inclusion	–	participation, presence,.	inclusão	inclusão das opções	0.12	inserção (insertion),.
Task 4: SR	**CompLex-BC**	exchange	(exchange, brains)	1	–	intercâmbio	intercâmbio efetivo das artes	0.21	troca (replacement),.
		sight	(sight, implants)	0	–	vista	agradável à sua vista	0.12	visão (view),.
		wisdom	(wisdom, women)	1	–	sabedoria	na muita sabedoria há	0.13	conhecimento (knowledge),.
		sword	(sword, densities)	0	–	espada	ferimentos por espada	0.07	faca (knife),.
		spirit	(spirit, Mesopotamia),.	0	–	espírito	há um espírito	0.08	almas (souls),.

Table 1: Illustrates the creation of MultiLS-PT. **"–"** indicates missing data in previous datasets. **T.** stands for sub-tasks within the LS pipeline that the corresponding dataset could be used for prior to MultiLS expansion. **D.** is Dataset. **Val.** represents assigned complexity value. Only a snapshot of contexts and candidate substitutions are shown.

the training and evaluation of the entire LS pipeline, including LCP. We use the MultiLS framework to guide the creation of the first multi-purpose, and multi-genre LS dataset, named MultiLS-PT (Table 1). The MultiLS framework consists of the following summarized steps.

Selection We identified target words from four pre-existing English datasets: CompLex (Shardlow et al., 2020), ALEXSIS-EN (Saggion et al., 2022), ALEXSIS+ (North et al., 2023a), and CompLex-BC (North et al., 2022c). Only words with a similar use and meaning within both English and Portuguese were hand-selected to provide comparable data for future multilingual and cross-lingual experiments (Section 8.1). Selection was done by a trained linguist fluent in both languages.

Context Retrieval Once target words had been identified, we automatically scraped several genres (bible extracts, news articles, and biomedical papers) to obtain new and varied sentences, hereby referred to as contexts, for each target word ready for annotation. Bible instances were obtained from Portuguese translations of the King James Bible. News instances were scraped from the PorSimplesSent dataset (Leal et al., 2018) as well as from the CC-News (Common Crawl-News) corpus (North et al., 2023a). Biomedical instances were extracted from abstracts of biomedical literature supplied by WMT-2019 (Bawden et al., 2019).

New Complexities (Val.) We presented target words in bold within the scraped contexts to annotators and asked annotators to rate their perceived difficulty using a 5-point Likert-scale: very easy (1), easy (2), neutral (3), difficult (4), to very difficult (5) (Shardlow et al., 2020, 2022). Each target word was annotated by 25 crowd-sourced Amazon Mechanical Turk (MTurk) workers located in Brazil. Table 2 shows an example Human Intelligence Task (HIT) presented to each of the 25 annotators. We selected a high number of annotators in order to get a representative gold complexity value for each target word by averaging the returned labels. Annotators were paid 2 cents of US Dollar per annotation allowing them to surpass the minimum hourly wage in Brazil.

New Substitutions Additionally, we asked annotators to suggest a valid simplification to the target word that fits within its surrounding context. Generated candidate substitutions were ranked per their suggestion frequency providing a list of gold simplifications.

4 MultiLS-PT Dataset

The uniqueness of the MutliLex framework is the collection of both continuous complexity values and gold candidate substitutions. This is what gives MultiLS-PT and future datasets that follow the MultiLS framework their distinctive multi-task functionality. The resulting MultiLS-PT dataset is unlike any other prior Portuguese dataset for LS. As referenced in Section 2, only two datasets exist made specifically for Portuguese LS: SIMPLEX-PB (Hartmann and Aluísio, 2020), and ALEXSIS-

Example MTurk HIT for Annotation	Difficulty
Identify the word ***authorities*** in the sentence below: "One of the greatest ***authorities*** on the subject, says that the destruction of the biome is irreversible."	1. Very Easy 2. Easy 3. Neutral 4. Difficult 5. Very Difficult

Tasks

(1). In your opinion, how difficult is the word in bold in this sentence? Select from 1 to 5.

(2). Write a simpler alternative to the word in bold (if any). Your suggestion must maintain the meaning of the sentence above and be easier to understand than the word in bold.

Table 2: An example HIT provided to the annotators. The HIT asks for both a continuous complexity rating and a suggested simplification. Each HIT was provided in Portuguese. Example has been translated for illustrative purposes.

PT (North et al., 2022b). However, these datasets only contain candidate substitutions without complexity values for target words. Moreover, both datasets are restricted to a specific genre. MultiLS-PT, on the other hand, contains 5,165 Portuguese target words annotated with complexity values in context taken from the Bible (2,321), news articles (1,817), and biomedical texts (1,237) with each target word also having an average of two gold candidate substitutions. Table 3 shows a direct comparison between MultiLS-PT and existing Portuguese datasets for LS.

	SIMPLEX-PB	ALEXSIS-PT	MultiLS-PT
Genre	children's books	newspapers	multi-genre
# Annotators	5	25	25
# Target Words	730	387	5,165
# Complexity Vals.	**-**	**-**	**5,165**
# Substitutions	**3,650**	**9,605**	**9,932**

Table 3: Comparison of Portuguese datasets for LS. MultiLS-PT is the first LS dataset to contain both gold complexity values (vals.) and candidate substitutions.

5 Tasks

We showcase three applications of the MultiLS-PT dataset for LS. We believed substitute selection to be conducted simultaneously during substitute generation and ranking, and therefore have only focused on LCP, substitute generation, and substitute ranking in the form of binary comparative LCP (North et al., 2022c). Each task was defined as follows: *LCP:* a regression-based task. Models were trained to automatically identify complex words by predicting their complexity value, between 0 (very easy) and 1 (very hard), of a target word in context. *SG:* a text generation task. Models were set to generate top-10 (k) candidate substitutions. *Binary Comparative LCP (BC-LCP):* a binary classification task used for substitute ranking (North et al., 2022c). Models were trained to rank candidate substitutions by assigning either 0 or 1 labels; 0 indicated that candidate 1 has a greater complexity than candidate 2 and 1 denoted the opposite.

Data for each task was formatted differently for model training. Example instances with gold labels are provided below (Table 4). Gold labels for the three tasks were averaged complexity values, most frequently suggested simplifications, and a binary label showing which of two candidate words was more complex, respectively.

Task	Example Instance with Gold Label(s)
LCP	"Procure **consulta** para diagnóstico" <\t> 0.73 (Gold) (Translation: Seek **consultation** for diagnosis) "Múltiplas feridas de **espada**" <\t> 0.08 (Gold) (Translation: Multiple **sword** wounds)
SG	"**consulta**" <\t> **respostas**, **conselho**, ... (Gold) (Translation: **consult** <\t> **answers**, **advice**) "**espada**" <\t> **faca**, **lâmina** ... (Gold) (Translation: **sword** <\t> **knife**, **blade**)
BC-LCP	"**respostas**" <\t> "**conselho**" <\t> 1 (Gold) (Translation: **answers** <\t> **advice**) "**lâmina**" <\t> "**faca**" <\t> 0 (Gold) (Translation: **blade** <\t> **knife**)

Table 4: Example instances with gold labels used for training each task. Only a snapshot of gold simplifications for SG are shown. For BC-LCP, a gold label of 1 shows candidate word 1 as being less complex than candidate word 2; i.e. "answers" is less complex than "advice", whereas 0 shows the opposite.

#	Task	Train	Dev	Test	Total
1	LCP	3,615	516	1,034	5,165
2	SG	-	-	462	462
3	BC-LCP	20,113	2,873	1,029	24,015

Table 5: MultiLS-PT's train, dev, and test splits per task. No training was conducted for the SG task.

MultiLS-PT was divided to have a 70/10/20 corresponding train, dev, and test split for the LCP and binary comparative LCP tasks, whereas the SG task had no train, dev, and test split since it was conducted in a zero-shot setting (Table 5). The test set of the binary comparative LCP task was also reduced by removing candidate substitution pairs that contained unrelated words and therefore were unsuitable for candidate ranking. Each task used a different number of total instances. The LCP task leveraged all 5,165 instances. The SG and BC-LCP tasks, on the other hand, utilized smaller subsets of the MulitLex-PT dataset. The SG task used a total of 462 instances that had a minimum of 5 gold simplifications in order to conduct meaningful eval-

Sub-Task	Num.	Name	Prompt
LCP	1	ZeroShot-5-Likert	On a scale from 1 to 5 with 5 being the most difficult, how difficult is the "target word"? Answer:
	2	Context-5-Likert	On a scale from 1 to 5 with 5 being the most difficult, how difficult is the "target word" in the above sentence? Answer:
	3	ZeroShot-10-Likert	On a scale from 1 to 10 with 10 being the most difficult, how difficult is the "target word"? Answer:
	4	Context-10-Likert	On a scale from 1 to 10 with 10 being the most difficult, how difficult is the "target word" in the above sentence? Answer:
	5	Ensemble-5-Likert	Average returned complexity from prompts 1 to 2.
	6	Ensemble-10-Likert	Average returned complexity from prompts 3 to 4.
SG	1	ZeroShot	Find ten easier words in Portuguese for "target word". Answer:
	2	Context	Find ten easier words in Portuguese for "target word" in the above sentence. Answer:
BC-LCP	1	Difficulty	Which word is more difficult "target word1" or "target word2"? Answer:
	2	Frequency	Which word is less common: "target word1" or "target word2"? Answer:
	3	Context	Which sentence is more difficult: (a). "sentence1" or (b). "sentence2"? Answer:
	4	Ensemble	All of the above.

Table 6: Prompts used per task.

uation. The BC-LCP task used a total of 24,015 instances comparing words of similar meaning and usage per a substitute ranking scenario.

6 Models

Multiple approaches using state-of-the-art models were applied to all three tasks. These approaches ranged from prompt-learning, regression, masked-language modeling (MLM) to binary classification depending on the task. Several LLMs were chosen to perform various prompt learning experiments given their high performance on a variety of NLP-related tasks. These LLMs, all of varying sizes, included GPT-3.5 (text-davinci-003) from OpenAI's API, alongside Mistral (Jiang et al., 2023), Llama-2 (Touvron et al., 2023), Falcon, and MPT available on Hugging Face. The prompts fed into these LLMs for LCP, substitute generation, and binary comparative LCP are shown in Table 6. These prompts were designed to artificially replicate answers provided by human annotators by copying the instruction supplied via MTurk.

We also experimented with several pre-trained transformers and feature engineering models such as support vector machine (SVM) and random forest (RF). Transformers and feature engineered models are currently state-of-the-art for LCP and binary comparative LCP, respectively (Shardlow et al., 2021a; North et al., 2022c). Transformers trained with a MLM objective were also state-of-the-art for substitute generation and selection prior to the arrival of recently proposed LLMs (Saggion et al., 2022; North et al., 2022a). MLM models replace the target word with a "[MASK]" special token and then attempt to provide a suitable simplification based on the masked target word and its surrounding context (Qiang et al., 2020).

We selected several transformers pre-trained on English and/or Portuguese data. These included BERT, mBERT (Devlin et al., 2019), RoBERTa (Zhuang et al., 2021), XLM-R (Conneau et al., 2020), BR-BERTo[1], Albertina PT-BR[2], ALbertina PT-PT[3] (Rodrigues et al., 2023), RoBERTa-PT-BR[4], and BERTimbau[5] (Souza et al., 2020) and were also obtained from Hugging Face. Each transformer was fine-tuned on the LCP and binary comparative LCP data supplied by MultiLS-PT as shown in Table 4. Fine-tuning was conducted over 5 epochs with a learning rate of 2e-5, a batch size of 8 and a max sequence length of 256 using a NVIDIA GeForce RTX 3060 GPU. No fine-tuning was conducted for substitute generation given that it is a zero-shot text generation task. Feature engineered approaches were trained on features previously shown to be indicative of lexical complexity (Desai et al., 2021; Shardlow et al., 2021b). Training was conducted over 5 epochs on features ranging from word length, syllable count, frequency, prevalence, and age-of-acquisition (AoA). Our SVM was set to have a sigmoid activation function and our RF was set to have 100 trees. Frequencies were calculated using the Exquisite Corpus[6] for Portuguese. English prevalence and AoA values were taken from Brysbaert et al. (2019) and Brysbaert and Biemiller (2017), respectively. These values were mapped to Portuguese due to the limited availability of Portuguese psycholinguistic datasets.

Evaluation Metrics Tasks were evaluated using their respective evaluation metrics found through-

[1] huggingface.co/rdenadai/BR_BERTo
[2] huggingface.co/PORTULAN/albertina-ptbr
[3] huggingface.co/PORTULAN/albertina-ptpt
[4] huggingface.co/josu/roberta-pt-br
[5] huggingface.co/neuralmind/bert-base-portuguese-cased
[6] https://github.com/LuminosoInsight/exquisite-corpus

out LS literature (Štajner et al., 2022). Mean squared error (MSE), Pearson Correlation (R) and Spearman Correlation (ρ) were used to evaluate LCP, with lower MSE values correlated with greater performance (Shardlow et al., 2021a). Weighted average recall, precision, and F1-score were used to assess binary comparative LCP (North et al., 2022c). However, substitute generation was evaluated using a alternative set of evaluation metrics introduced in the TSAR-2022 shared-task (Štajner et al., 2022; Saggion et al., 2022), including potential and accuracy at top-k = 1. Potential is the ratio of the predicted candidate substitutions that match the most frequently suggested gold label. Accuracy at top-k = 1 (A@1@Top1) is the ratio of best predicted candidate substitutions at rank #1 that are equal to the most appropriate gold simplification also at rank #1. It is important to note, A@1@Top1 is different from ACC@1 that is reported alongside A@1@Top1 at TSAR-2022 (Saggion et al., 2022). ACC@1 takes into consideration multiple generated candidates, whereas A@1@Top1 only considers the top-k = 1 candidate generated. We decided to use A@1@Top1 as it is a more competitive evaluation metric.

7 Results

In this section we present the results for each task using the MultiLS-PT dataset. We report model performances on LCP (Table 9 in the Appendix) before moving to substitution generation (Table 8 in the Appendix), and finally substitute ranking via binary comparative LCP (Table 7). For each task, we look into LLM versus transformer performance, impact of genre and context, and compare model performances on MultiLS-PT to prior datasets.

7.1 Lexical Complexity Prediction

Pre-trained transformers outperformed our LLMs for LCP, regardless of genre or prompt (Table 9). Transformers fine-tuned on all of the instances from MultiLS-PT, depicted lower MSE values alongside higher R and ρ values compared with our prompt learning approaches. The highest performing models were BERTimbau (#1) and XLM-R-L (#3) having achieved R values of 0.8423 and 0.8295, ρ values of 0.8081 and 0.8054, and MSE values of 0.0664 and 0.0698, respectively. In comparison, our best performing LLMs achieved noticeably worst performances when asked to rate the complexity of the target word in a zero-shot setting (ZeroShot-5-Likert, Table 6). Mistral-8X7B (#8) achieved a R value of 0.1810, a ρ value o 0.4816, and a MSE value of 0.1810. Llama-2-13B (#13) produced a R value of 0.2249, a ρ value o 0.3441, and a MSE value of 0.2249. All other prompts that took into consideration context or had their answers averaged within an ensemble resulted in worst performances. Without prior exposure to gold complexity ratings, our prompts were ineffective at modeling the complexity assignments of Portuguese speakers.

Differences in LCP performance per genre were observed by both transformers and LLMs. Transformers fine-tuned and evaluated on biomed instances returned the best results followed by Bible and news extracts. BERTimbau (#1) produced R values of 0.8959, 0.8260, and 0.7244 on biomed, Bible, and news instances, respectively. Likewise, XLM-R-L (#3) achieved R values of 0.8907, 0.8055, and 0.0.7212 on biomed, Bible, and news instances, respectively. Interestingly, Mistral-8x7B performed best on Bible instances having achieved a R value of 0.5608, followed by news instances attaining a R value of 0.4663, and lastly biomedical instances scoring a R value of 0.3762. Varying performances between genre can be seen throughout the remaining tasks.

7.2 Substitute Generation

Simplifications generated by LLMs were of a greater quality compared to those generated by the majority of MLM approaches for all instances (Table 8). The best LLM, being Falcon-40B (#1), achieved an A@1@Top1 of 0.01708 and a potential of 0.5291, closely followed by Mistral-8x7B (#3) having obtained an A@1@Top1 of 0.1375, and a potential of 0.4083. (Table 8). The majority of MLM approaches, including transformers such as XLM-R (#12), RoBERTa-PT-BR (#13), mBERT (#14), and so on, produced less suitable candidate substitutions with A@1@Top1 scores of 0.0458, 0.0333, 0.0229, respectively. However, the best performing MLM model, being BERTimbau (#9), achieved an A@1@Top1 of 0.0916, that surpassed the performance of smaller LLMs, such as Mistral-7B, and Llama-2-7B. A direct correlation was therefore observed between LLM size and overall performance.

Context influenced prompt performance. The three best performing LLMs, Falcon-40B (#1), Mistral-8x7B (#3), and Llama-2-13B (#5), produced their best simplifications across all genres

when fed prompts referring to the target word's context. Their Zero-shot counterparts, on the other hand, performed noticeably worst. Falcon-40B scored a A@1@Top1 of 0.1708 with context dropping to 0.1375 without context. Mistral-8x7B achieved a A@1@Top1 of 0.1375 with context falling to 0.1125 without context. Llama-2-13B showed the greatest decrease in performance having fell from a A@1@Top1 of 0.1104 with context to a much lower A@1@Top1 of 0.0520 without context. This signifies the vital role context plays in substitute generation.

Substitute generation performance also varied between genres. Falcon-40B (#1), Mistral-8x7B (#3), and Llama-2-13B (#5) achieved greater A@1@Top1 and potential scores for Bible instances when compared to biomed and news instances. Falcon-40B Mistral-8x7B, and Llama-2-13B produced candidate substitutions with A@1@Top1 scores of 0.2086, 0.1695, and 0.1260 for Bible instances, respectively. However, the same LLMs produced inferior candidate substitutions for news extracts with A@1@Top1 scores of 0.1329 by Falcon-40B, 0.1040 by Mistral-8x7B, and 0.0867 by Llama-2-13B. We observed little variation between these LLMs performance on the biomedical extracts with Mistral-8x7B and Llama-2-13B achieving the same A@1@Top1 of 0.1168.

Performances on the news genre were lower than those achieved at the TSAR-2022 shared-task (Saggion et al., 2022). The wining system of TSAR-2022's Portuguese track was an BERTimbau-based system that achieved an A@1@Top1 of 0.2540 on the shared-task's news extracts (North et al., 2022a). Our best performing model, being Falcon-40B (#1), achieved an A@1@Top1 of 0.1329 for news instances. We attribute this performance to how MultiLS-PT's news instances were collected. Target words within MultiLS-PT's news genre were taken from CompLex's European Parliamentary proceedings (Parl) genre (Shardlow et al., 2020). This was done to maintain a level of similarity between the two datasets as described in Section 4. However, as a consequence, this resulted in more nuanced and complex sentences being present among MultiLS-PT's news instances in comparison to TSAR-2022's news extracts making substitute generation a more challenging task.

7.3 Binary Comparative LCP

GPT 3.5 achieved the best performance for binary comparative LCP. For the majority of instances, GPT 3.5 (#1) and Mistral-8x7B (#2) were able to predict which of two target words were more or less complex having achieved F1-scores of 0.7064 and 0.6992 for all instances, respectively. Unlike for LCP, no clear distinction was observed between the performances of several LLMs and transformers. For example, mBERT achieved an F1-score of 0.5223, whereas other larger LLMs, such as Llama-2-13B (#8), Falcon-7B (#6) and Falcon-40B (#11) attained F1-scores of 0.5043, 0.5097, 0.4737, respectively. This was likely due to the difficult nature of the task.

		F1-Score			
#	Model-Prompt/Features	**All**	**Bible**	**News**	**Biomed**
1	GPT 3.5-Frequency	0.7064	0.6555	0.7474	0.6063
2	Mistral-8x7B-Frequency	0.6992	0.5907	0.6986	0.6087
3	Mistral-7B-Difficulty	0.6276	0.6556	0.5989	0.6516
4	Llama-2-7B-Ensemble	0.6015	0.6168	0.5826	0.5984
5	mBERT	0.5223	1.0000	0.5932	0.3213
6	Falcon-7B-Frequency	0.5097	0.4771	0.2536	0.4781
7	RF-all	0.5044	0.5472	0.4938	0.3999
8	Llama-2-13B-Difficulty	0.5043	0.5225	0.6493	0.5986
9	SVM-all	0.4995	0.5030	0.4721	0.4875
10	MPT-7B-Difficulty	0.4789	0.5212	0.5633	0.5085
11	Falcon-40B-Sentence	0.4737	0.4692	0.4684	0.6355
13	XLM-R	0.4434	0.3995	0.4111	0.4579

Table 7: Shows weighted average binary comparative LCP F1-scores on instances separated by genre and language. Performances are shown as weighted averages. Models are ranked (#) from best to worst F1-Score for all instances. LLMs are separated by inputted prompt.

Features known to correlate with complexity were embedded within several prompts to better understand the thought process of our LLMs (Table 6). It was discovered that LLMs performed differently when taking into consideration different prompts. GPT 3.5 (#1) and Mistral-8x7B (#2) achieved their greatest F1-scores of 0.7064 and 0.6992 respectively when being asked to determine which target word was more or less complex based on its frequency. In contrast, these same models achieved noticeably worst F1-scores when being fed prompts that explicitly referred to word difficulty (Table 6). When inputted difficulty-based prompts, GPT 3.5 produced a F1-score of 0.6273 and Mistral-8x7B achieved a F1-score of 0.6154 amounting to a -0.0791 and -0.0838 decrease in performance respectively. Therefore, it would appear that our best performing LLMs considered frequency as being a highly influential factor in determining a word's overall complexity.

On several occasions, prompt performance varied between genres for binary comparative LCP. GPT 3.5 (#1) and Mistral-8x7B (#2) were able

to use frequency-based prompts to differentiate the complexities of words taken from the news genre more easily than they were for words taken from the Bible or biomed genres. For the news genre, GPT 3.5 attained a F1-score of 0.7474 and Mistral-8x7B produced a F1-score of 0.6986. However, for the Bible and biomed genre, GPT 3.5 produced F1-scores of 0.6555 and 0.6063 respectively, whereas as Mistral-8x7B achieved F1-scores of 0.5907 and 0.6087 respectively. Interestingly, Falcon-40B (#12) produced it's highest F1-score when using sentence-based prompts (Table 6) for ranking words from the biomed genre. A probable explanation likely stems from the varying lexical diversity of each genre. The news genre was found to contain a greater combination of everyday and jargon-specific vocabulary making its complex and non-complex words easier to differentiate. The vocabulary of the Bible and biomed genres, on the other hand, were more jargon-specific making binary comparative LCP a harder task when considering word frequency, yet an easier task when comparing two target sentences since surrounding words are also taken into consideration.

8 Conclusion

The MultiLS framework provides a guide for the creation of a multi-purpose and multi-genre LS dataset. The MultiLS framework is unique in that it provides gold continuous complexity values and gold candidate substitutions, a feat not achieved by previous LS datasets (Sections 2 and 4). The resulting dataset can be used to train and evaluate all LS sub-task, including LCP.

We introduce MultiLS-PT, the first Portuguese LS dataset to be created using the MultiLS framework. By experimenting on MultiLS-PT, we were able to theorize the optimum LS pipeline for Portuguese given current state-of-the-art models and make several observations regarding the impact of genre and context on LS. Performances indicate that LLMs are incapable of rating lexical complexity for a specific target demographic, but are able to generate and rank possible simplifications. This provides insight into the role LLMs will have in future LS systems.

8.1 Future Work

In this paper, we provided empirical evidence of the MultiLS framework's potential to be used as an all-in-one simplification framework. We have trained models and conducted several experiments using MultiLS-PT. However, there are multiple research questions left outstanding that the MultiLS framework and MultiLS-PT can be used to answer. Future work will utilize the MultiLS framework to explore three open research areas as follows.

Full Pipeline Evaluation LLMs are able to simplify an entire text as a response to a single prompt and are even state-of-the-art for substitute generation (Section 7). This questions the need for models trained on individual sub-tasks of the LS pipeline. Comparisons need to be made between the readability and accessibility of texts simplified by a general LLM compared to texts simplified by end-to-end LS systems. To make this possible, we aim to perform an empirical comparison of the performance of a LS pipeline trained on a MultiLS dataset to a generalized LLM for text simplification.

Multilingual LS and Cross-lingual Transfer Cross-lingual models with transfer learning from a high-resource to a low-resource language is a successful strategy widely used in various NLP tasks. However, there is conflicting evidence regarding the performance of cross-lingual models for LS (North et al., 2022a; Štajner et al., 2022; North, Kai and Zampieri, Marcos, 2023). Further research is needed to establish whether cross-lingual transfer is viable for LS, especially for which LS sub-tasks. In this endeavour, we plan to apply the MultiLS framework to other languages whereby the complexities of shared and hand-selected words can be used to research the effects of multilingual LS and cross-lingual transfer on LS performance.

Domain Generalization LS systems are commonly trained on a single dataset containing either a specific genre, including newspaper extracts (Leal et al., 2018; North et al., 2022b) or educational materials (Hartmann and Aluísio, 2020; Merejildo, 2021), or for an undefined mix of genres, such as Wikipedia extracts on a range of topics (Shardlow, 2013; Horn et al., 2014). The lack of datasets containing multiple types of texts separated by genre limits the development of LS systems capable of domain generalization. The results presented in this paper account for different genres. As such, researchers can see what does and does not work well for specific genres and use this information to develop their LS systems accordingly. We aim to continue to experiment with MultiLS-PT developing a fully generalizable LS system for Portuguese.

Lay Summary

Lexical Simplification (LS) is the task of automatically replacing difficult words for easier ones while preserving a sentence's original meaning. LS is an important component of text simplification system that are developed to simplify texts aiming to improve accessibility to various populations such as individuals with learning disabilities.

Datasets containing hundreds or thousands of excerpts of texts annotated with human judgments are needed to train LS systems. Several datasets exist for LS but each of them specialize in a step of the traditional LS pipeline such as recognizing complex words, generating substitute words, or selecting the best substitute word. To the best of our knowledge no single LS dataset represents all steps of the pipeline.

To address this limitation, we propose MultiLS, the first framework that allows for the creation of all-in-one LS datasets representing all steps of the pipeline. We present MultiLS-PT, a Portuguese dataset dataset created using the MultiLS framework. MultiLS-PT contains texts from the Bible, news articles, and biomedical texts. Finally, we carry out various experiments that demonstrate the potential of the MultiLS framework and the MultiLS-PT dataset of improving LS systems and related assistive technologies.

References

Rachel Bawden, Kevin Bretonnel Cohen, Cristian Grozea, Antonio Jimeno Yepes, Madeleine Kittner, Martin Krallinger, Nancy Mah, Aurelie Neveol, Mariana Neves, Felipe Soares, Amy Siu, Karin Verspoor, and Maika Vicente Navarro. 2019. Findings of the WMT 2019 Biomedical Translation Shared Task: Evaluation for MEDLINE Abstracts and Biomedical Terminologies. In *Proceedings of WMT*.

Marc Brysbaert and Andrew Biemiller. 2017. Test-based age-of-acquisition norms for 44 thousand English word meanings. *Behavioural Research*, 49:1520–1523.

Marc Brysbaert, Pawel Mandera, Samantha McCormick, and Emmanuel Keuleers. 2019. Word prevalence norms for 62,000 English lemmas. *Behavior Research Methods*, 51:467–479.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, and Others. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In *Proceedings of ACL*.

Abhinandan Desai, Kai North, Marcos Zampieri, and Christopher Homan. 2021. LCP-RIT at SemEval-2021 Task 1: Exploring Linguistic Features for Lexical Complexity Prediction. In *Proceedings of SemEval*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL*.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, twenty-sixth edition. SIL International, Dallas, Texas.

Daniel Ferres and Horacio Saggion. 2022. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of LREC*.

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.

Colby Horn, Cathryn Manduca, and David Kauchak. 2014. Learning a Lexical Simplifier Using Wikipedia. In *Proceedings of ACL*.

Timothy D. Ireland. 2008. Literacy in brazil: From rights to reality. *International Review of Education*, 54(5/6):713–732.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv: 2310.06825*.

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 COLING*.

Mounica Maddela and Wei Xu. 2018. A Word-Complexity Lexicon and A Neural Readability Ranking Model for Lexical Simplification. In *Proceedings of EMNLP*.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 Task 10: English Lexical Substitution Task. In *Proceedings of SemEval*.

Borbor Merejildo. 2021. Creación de un corpus de textos universitarios en español para la identificación de palabras complejas en el área de la simplificación léxica. Master's thesis, Universidad de Guayaquil.

Kai North, Alphaeus Dmonte, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023a. ALEXSIS+: Improving Substitute Generation and Selection for Lexical Simplification with Information Retrieval. In *Proceedings of BEA*.

Kai North, Alphaeus Dmonte, Tharindu Ranasinghe, and Marcos Zampieri. 2022a. GMU-WLV at TSAR-2022 Shared Task: Evaluating Lexical Simplification Models. In *Proceedings of TSAR*.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023b. Deep Learning Approaches to Lexical Simplification: A Survey.

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022b. ALEXSIS-PT: A New Resource for Portuguese Lexical Simplification. In *Proceedings of COLING*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2022c. An Evaluation of Binary Comparative Lexical Complexity Models. In *Proceedings of BEA*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2022d. Lexical Complexity Prediction: An Overview. *ACM Computing Surveys*, 55(9).

North, Kai and Zampieri, Marcos. 2023. Features of lexical complexity: insights from L1 and L2 speakers. *Frontiers in Artificial Intelligence*.

Gustavo Paetzold and Lucia Specia. 2016a. PLUMBErr: An Automatic Error Identification Framework for Lexical Simplification. In *Proceedings of LREC*.

Gustavo Paetzold and Lucia Specia. 2016b. SemEval 2016 Task 11: Complex Word Identification. In *Proceedings of SemEval*.

Gustavo H. Paetzold and Lucia Specia. 2017. A Survey on Lexical Simplification. *J. Artif. Int. Res.*, 60(1):549–593.

Gustavo Henrique Paetzold and Lucia Specia. 2015. LEXenstein: A Framework for Lexical Simplification. In *ACL 2015 System Demonstrations*, pages 85–90.

Gustavo Henrique Paetzold and Lucia Specia. 2016c. Benchmarking Lexical Simplification Systems. In *Proceedings of LREC*.

Jipeng Qiang, Yun Li, Zhu Yi, Yunhao Yuan, and Xindong Wu. 2020. Lexical Simplification with Pretrained Encoders. In *Proceedings of AAAI*.

João Rodrigues, Luís Gomes, João Silva, António Branco, Rodrigo Santos, Henrique Lopes Cardoso, and Tomás Osório. 2023. Advancing Neural Encoding of Portuguese with Transformer Albertina PT-*. *arXiv: 2305.06721*.

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 TSAR*.

Matthew Shardlow. 2013. The CW Corpus: A New Resource for Evaluating the Identification of Complex Words. In *Proceedings of ACL*.

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020. CompLex — a new corpus for lexical complexity prediction from Likert Scale data. In *Proceedings of READI*.

Matthew Shardlow, Richard Evans, Gustavo Paetzold, and Marcos Zampieri. 2021a. SemEval-2021 Task 1: Lexical Complexity Prediction. In *Proceedings of SemEval*.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2021b. Predicting lexical complexity in english texts. In *Proceedings of LREC*.

Matthew Shardlow, Richard Evans, and Marcos Zampieri. 2022. Predicting Lexical Complexity in English Texts: The Complex 2.0 Dataset. *Language Resources and Evaluation*, 56(4):1153–1194.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: pretrained BERT models for Brazilian Portuguese. In *Proceedings of BRACIS*.

Lucia Specia, Kumar Jauhar, Sujay, and Rada Mihalcea. 2012. SemEval - 2012 Task 1: English Lexical Simplification. In *Proceedings of SemEval*.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical Simplification Benchmarks for English, Portuguese, and Spanish. *Frontiers in Artificial Intelligence*.

Anaïs Tack, Thomas François, Anne-Laure Ligozat, and Cédrick Fairon. 2016. Evaluating Lexical Simplification and Vocabulary Knowledge for Learners of French: Possibilities of Using the FLELex Resource. In *Proceedings of LREC*.

Hugo Touvron, Louis Martin, and et al. Kevin Stone. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv: 2307.09288*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Luci Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A Report on the Complex Word Identification Shared Task 2018. In *Proceedings of BEA*.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex Word Identification: Challenges in Data Annotation and System Performance. In *Proceedings of NLP-TEA*.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. A Robustly Optimized BERT Pre-training Approach with Post-training. In *Proceedings of CCL*.

A Appendix

		All		Bible		News		Biomed	
#	Model-Prompt	A@1@Top1	Potential	A@1@Top1	Potential	A@1@Top1	Potential	A@1@Top1	Potential
1	Falcon-40B-Context	0.1708	0.5291	0.2086	0.5043	0.1329	0.5606	0.1428	0.5324
2	Falcon-40B-ZeroShot	0.1375	0.4333	0.1826	0.4521	0.0867	0.4219	0.1168	0.4025
3	Mistral-8x7B-Context	0.1375	0.4083	0.1695	0.4173	0.1040	0.3988	0.1168	0.4025
4	Mistral-8x7B-ZeroShot	0.1125	0.3187	0.1260	0.2739	0.0693	0.3294	0.1688	0.4285
5	Llama-2-13B-Context	0.1104	0.3208	0.1260	0.2869	0.0867	0.3526	0.1168	0.3506
6	GPT 3.5-ZeroShot	0.1083	0.3479	0.1217	0.3043	0.0867	0.3930	0.1168	0.3766
7	GPT 3.5-Context	0.1062	0.4250	0.1304	0.3956	0.0693	0.4797	0.1168	0.3896
8	Mistral-7B-Context	0.0937	0.2750	0.1086	0.2652	0.0693	0.2716	0.1038	0.3116
9	BERTimbau	0.0916	0.2645	0.1086	0.2434	0.0751	0.2890	0.0779	0.2727
10	Mistral-7B-ZeroShot	0.0729	0.2062	0.0826	0.1913	0.0578	0.2080	0.0779	0.2467
11	Llama-2-13B-ZeroShot	0.0520	0.1187	0.0739	0.1565	0.0231	0.0809	0.0519	0.0909
12	XLM-R	0.0458	0.1250	0.0478	0.0913	0.0462	0.1618	0.0389	0.1428
13	RoBERTa-PT-BR	0.0333	0.1229	0.0434	0.1000	0.0173	0.1329	0.0389	0.1688
14	mBERT	0.0229	0.1145	0.0217	0.0826	0.0231	0.1445	0.0259	0.1428
15	Llama-2-7B-ZeroShot	0.0229	0.1000	0.026	0.0782	0.0115	0.1213	0.0389	0.1168
16	MPT-7B-Context	0.0229	0.0958	0.0260	0.0826	0.0115	0.1040	0.0389	0.1168
17	BR-BERTo	0.0250	0.0770	0.0304	0.0391	0.0173	0.1098	0.0259	0.1168
18	MPT-7B-ZeroShot	0.0208	0.0750	0.0260	0.0652	0.0057	0.0867	0.0389	0.0779
19	Llama-2-7B-Context	0.0208	0.0541	0.0260	0.0391	0.0057	0.0635	0.0389	0.0779
20	Falcon-7B-Context	0.0166	0.0416	0.0173	0.0478	0.0057	0.0346	0.0389	0.0389
21	Albertina PT-BR	0.0145	0.0541	0.0173	0.0478	0.0115	0.0635	0.0129	0.0519
22	Albertina PT-PT	0.0145	0.0520	0.0173	0.0478	0.0115	0.0578	0.0129	0.0519
TSAR-2022 Benchmark (PT-BR)									
1	BERTimbau	-	-	-	-	0.2540	0.4812	-	-

Table 8: Shows substitute generation performances on instances separated by genre with at least five gold candidate substitutions in MultiLS-PT. Models are ranked (#) from best to worst A@1@Top1. LLMs are separated by inputted prompt. The winning system from TSAR-2022 (Saggion et al., 2022) provided as a benchmark.

			All			Bible			News			Biomed		
Approach	#	Model	MSE	R	ρ	MSE	R	ρ	MSE	R	ρ	MSE	R	ρ
Transformers	1	BERTimbau	0.0664	0.8423	0.8081	0.0726	0.8260	0.8275	0.0558	0.7244	0.7047	0.0677	0.8959	0.8740
	2	BERTimbau-L	0.0681	0.8324	0.8086	0.0746	0.8144	0.8227	0.0533	0.7450	0.7308	0.0746	0.8720	0.8573
	3	XLM-R-L	0.0698	0.8295	0.8054	0.0777	0.8055	0.8224	0.0550	0.7212	0.7214	0.0724	0.8907	0.8586
	4	XLM-R	0.0706	0.8187	0.7974	0.0773	0.8012	0.8155	0.0595	0.6774	0.6995	0.0716	0.8824	0.8612
	5	mBERT	0.0743	0.7968	0.7724	0.0815	0.7746	0.7808	0.0585	0.6801	0.6941	0.0804	0.8502	0.8332
	6	RoBERTa-PT-BR	0.1469	0.7968	0.7539	0.1506	0.7440	0.7395	0.1169	0.7214	0.6834	0.1811	0.8768	0.8430
	7	BR-BERTo	0.1844	0.7522	0.6791	0.1842	0.6865	0.6500	0.1488	0.6518	0.5906	0.2340	0.8569	0.8111
LLMs	8	Mistral-8X7B	0.1810	0.4603	0.4816	0.1576	0.5608	0.5480	0.1953	0.4663	0.4566	0.2063	0.3762	0.3877
	9	Llama-2-13B	0.2249	0.2737	0.3441	0.2089	0.2226	0.3330	0.2289	0.2569	0.3233	0.2535	0.2687	0.2903
	10	Mistral-7B	0.4156	0.2758	0.3349	0.4117	0.3762	0.3880	0.4428	0.3379	0.3327	0.3739	0.1148	0.2261
	11	GPT 3.5	0.5050	0.0520	0.0895	0.5019	0.0134	0.0481	0.5286	0.0624	0.1197	0.4692	0.1411	0.1504
	12	Llama-2-7B	0.4031	0.0392	0.1535	0.4064	0.0394	0.1343	0.4199	0.1951	0.2084	0.3631	-0.0121	0.1287
	13	Falcon-7B	0.4273	0.0008	0.0353	0.4150	-0.019	-0.0132	0.4722	0.0718	0.0993	0.3703	0.0285	0.0613
LCP-2021 Benchmark (English)														
Transformers	1	BERT-Ensemble	0.0609	0.7886	0.7369	-	-	-	-	-	-	-	-	-

Table 9: LCP performances on instances separated by genre. Models are ranked (#) from best to worst Pearson Correlation (R) for all instances. Results produced by LLMs were from our highest performing prompt 1. ZeroShot-5-Likert (Table 6). The winning system from LCP-2021 (Shardlow et al., 2021a) provided as a benchmark.