GMU at MLSP 2024: Multilingual Lexical Simplification with Transformer Models

Dhiman Goswami, Kai North, Marcos Zampieri
George Mason University, USA
dgoswam@gmu.edu

Abstract

This paper presents GMU's submission to the Multilingual Lexical Simplification Pipeline (MLSP) shared task at the BEA workshop 2024. The task includes Lexical Complexity Prediction (LCP) and Lexical Simplification (LS) sub-tasks across 10 languages. Our submissions achieved rankings ranging from 1^{st} to 5^{th} in LCP and from 1^{st} to 3^{rd} in LS. Our best performing approach for LCP is a weighted ensemble based on Pearson correlation of language specific transformer models trained on all languages combined. For LS, GPT4-turbo zero-shot prompting achieved the best performance.

1 Introduction

Understanding LCP and LS is crucial for enhancing communication accessibility and readability across diverse linguistic contexts. LCP involves analyzing linguistic features to understand text difficulty, while LS focuses on making complex language more readable without losing its meaning. Therefore, LCP and LS provide inclusive communication and broadening access to information. Nowadays, NLP research is interested in identifying complex words which may be difficult for certain readers (Shardlow, 2013; Paetzold and Specia, 2016a). These difficult words requires various types of intervention, such as direct replacement in the setting of LS (Gooding and Kochmar, 2019), or generating further explanation (Rello et al., 2015)

Previously, the task of LCP involved labelling the complex words by binary classification (Paetzold and Specia, 2016a; Zampieri et al., 2017; Yimam et al., 2018). This approach was referred to as Complex Word Identification (CWI) which means a word can either be complex or not. However, in practice, word complexity should be a continuous value representing from the least to the most complex. Shardlow et al. (2021) and Shardlow et al. (2020b) were the first to introduce the task of LCP where a continuous value is assigned to identify a word's complexity. LS is the task of replacing difficult words with easier synonyms while preserving the information and intelligibility of the original text. This is a sub-task of Automatic Text Simplification (ATS) (Saggion and Hirst, 2017). Recently, similar to LCP, this task has also gained considerable amount of attention (Štajner et al., 2022).

In this paper, we use a cross-lingual weighted ensemble of transformer models to find LCP of a word in context of a sentence for 10 languages. For LS, we use GPT4-turbo (OpenAI, 2023) zero-shot prompting and also top 10 suggestions of GPT4-turbo and transformers models in terms of cosine similarity for 10 languages.

2 Related Work

2.1 Lexical Complexity Prediction

North et al. (2023c) is considered a comprehensive survey on LCP which provides us with a chronological journey of this task. LCP researchers traditionally used lexical features like word2vec, POS tag, frequency features including maximum entropy as traditional approaches (Paetzold and Specia, 2016a). Moreover, features like word length, frequency, n-gram features and word embeddings were also explored (Yimam et al., 2018) for LCP. On top of that, Binary classifiers such as SVMs, Decision Trees, Random Forests and threshold based metrics, variety of traditional machine learning classifiers and Neural Networks were used in different LCP systems. For example, the winning system CWI shared task of 2016 used a threshold-based methods and features extracted from Simple Wikipedia (Paetzold and Specia, 2016b) and Adaboost with WordNet features, POS tags, dependency parsing relations and psycholinguistic features were used by the winning system (Gooding and Kochmar, 2018) of BEA 2018.

From the approach of binary classification, LCP

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 627–634
June 20, 2024 ©2024 Association for Computational Linguistics

gradually shifted towards regression or probabilistic classification and thus transformer based models show better performance. A few years later, the idea of expressing complexity of words with a continuous value was first introduced on LCP shared task 2021 (Shardlow et al., 2021). A pre-trained transformer models fine-tuned for LCP (Pan et al., 2021) and a weighted ensemble of BERT and RoBERTa (Yaseen et al., 2021) respectively won the single word multi-word expressions sub-task of the shared task of 2021.

2.2 Lexical Simplification

LS research has utilized the word embedding models for retrieval or substitution generation (Glavaš and Štajner, 2015; Paetzold and Specia, 2016b). A pipeline of Substitute Generation (SG), Substitute Selection (SS) and Substitute Ranking (SR) was developed for this task. SG returns top-k most appropriate substitution of the complex word which are easy to understand and also preserve the original complex word's meaning and context. SS filters the generated top-k candidate substitutions and removes the unsuitable substitutions. SR orders the remaining top-k candidate substitutions by the decreasing order of simplicity and replace the complex word with the most suitable substitution (North et al., 2023b). Such approaches have proven better compared to earlier systems.

The state of the art for English LS was the LS-BERT system (Qiang et al., 2020) before 2022. It used a BERT (Devlin et al., 2018) based masking technique to find suitable simplifications for complex words and employed unsupervised ranking using various feature combinations. In 2022, Ferrés and Saggion (2022) introduced a benchmark dataset for LS in Spanish named ALEXSIS, and conducted experiments with various neural and unsupervised systems. They also evaluated an adaptation of LSBERT for Spanish, achieving state-of-the-art performance. Similarly, North et al. (2022b) developed and evaluated transformer models for Portuguese in 2022, based on a new corpus derived from ALEXSIS, following the BERT masked approach for substitute generation.

The first multilingual LS shared task was TSAR-2022 (Saggion et al., 2023). On this shared task, the best ranking for English was achieved using GPT-3 zero shot and few shot prompting (Aumiller and Gertz, 2023). For Portuguese, two customized pre-trained monolingual transformers and a large pre-trained monolingual model BERTimbau for masked language modeling achieved the best performance (North et al., 2022a). This prompting technique was further introduced in ALEXSIS+ (North et al., 2023a). Likewise, a masked language model followed by candidate token generation, candidate word selection and candidate word pruning along cosine similarity and parts of speech checking for substitution ranking (Whistely et al., 2022) was used for Spanish LS. Recently, a detailed Multi-task LS framework has been proposed by (North et al., 2024) which enables the creation of a multi-task LS dataset and training of a full LS pipeline.

3 Datasets

The MLSP shared-task (Shardlow et al., 2024a) covers 10 different languages - Catalan, English, Filipino, French, German, Italian, Japanese, Portuguese, Sinhala, Spanish and it has two sub-tasks- LCP and LS. LCP data instances include a sentence of a specific language and a specific word from that language of various text genre like news, religious, educational, Wikibooks etc. (Shardlow et al., 2024b). Then a complexity value ranging from 0-1 of that specific word in the context of that sentence is given. LS also has similar types of data instances but instead of a complexity value 10 simplified substitutions of the target word are provided for each instance. Moreover, MultiLS SP/CA dataset was used for both the LCP and LS task for Spanish and Catalan language (Bott et al., 2024). For each language, the data annotators are from different age group and professions like students, language learners, university faculty, freelancers. The data was annotated by both native and non-native speakers of each specific language. The data count for all the languages are shown in Table 1.

Language	Test
Catalan	445
English	570
Filipino	570
French	570
German	570
Italian	570
Japanese	570
Portuguese	569
Sinhala	600
Spanish	593
All Combined	5,627

Table 1: Data Distribution of Lexical Complexity Prediction and Lexical Simplification Dataset

There is no training data for this task. 30 Trial data was provided for each of the languages. For both the tasks we used all the trial data for validation. We performed cross-lingual transfer learning for the target language for LCP task. Moreover, only for the LCP task in English, we used CompLex dataset (Shardlow et al., 2020a) as training set for additional experiment. We merged 421 trial, 7,662 train and 917 test instances of this dataset and used these 9,000 instances together for the training purpose of English. We used the English trial data provided for this shared task as validation data in this case.

4 Experiments

Trial data provided for all the languages of the LCP task is very small. In general, it is common to use data augmentation and back-translation techniques to increase the number of data instances in such conditions (Akhbardeh et al., 2021). However, it will not work here as these techniques can change the word or even the context of the word after augmentation and back-translation causing change to the complexity also. As such, we use the idea of cross-lingual weighted ensemble approach by using trial data of all the languages for training and validation. We used 80-20 train and validation split. After that we use the test data of the target language for predicting lexical complexity. For training we have used weighted ensemble of mBERT (Devlin et al., 2018), XLM-R (Conneau et al., 2020) and language specific BERT models. For Catalan, Filipino, French, German, Italian, Japanese, Sinhala and Spanish we used calBERT (Codegram, 2020), RoBERTa-tagalog (Cruz and Cheng, 2021), flauBERT (Le et al., 2020), germanBERT (Dbmdz, 2020b), italianBERT (Dbmdz, 2020a), japaneseBERT (Tohoku-NLP, 2020), sinhalaBERTo (Dhananjaya et al., 2022) and spanishBERT (Cañete et al., 2020) respectively. For English we used BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2021) as language specific model. For all language combined - ensemble of mBERT, XLM-R calBERT, DeBERTa, RoBERTa-tagalog, flauBERT, germanBERT, italianBERT, japaneseBERT, BERTimbau, sinhalaBERTo and spanishBERT were used. Pearson correlation coefficient was used as weight for the ensemble.

We use GPT4-turbo (Achiam et al., 2023) zero shot prompting which provides the best result for LS on both trial and test phase. Additionally, we used the same set up of BERT based models like LCP for all the languages to find the best 10 simplified substitutions for trial and test data. Then for each instances of a language, we took the set of all the words suggested by the BERT based models and GPT4-turbo together. After that, we find the embeddings of those words and the target token by LaBSE sentence transformer (Feng et al., 2020). Furthermore, we find the cosine similarity of the target token to the set of suggested word embeddings. Lastly, we choose the best 10 words by the decreasing order of cosine similarity of the embeddings.

5 Results

For LCP in English, we used the English trial data merged with the CompLex dataset and performed weighted ensemble. We rank 1^{st} with this procedure with Pearson correlation coefficient 0.8497. For the other 8 languages and all language combined we used the cross-lingual weighted ensemble. For Sinhala, we secure 3^{rd} rank with Pearson coefficient score 0.1246. For all language combined, Italian, Filipino, Spanish, Japanese, Catalan and German our rank is 4^{th} with Pearson coefficient 0.3494, 0.2919, 0.2823, 0.2438, 0.1775, 0.1549 and 0.1402 respectively. Lastly, we rank 5^{th} for French with 0.3193 Pearson coefficient. Test results for LCP are shown in Table 2.

For LS, zero-shot prompting by GPT-4 turbo performs the best for the 9 languages and all language combined. For Sinhala, we ranked 1^{st} with Accuracy@1@Top1 score 0.4182. For German, Spanish, all language combined, Japanese and Filipino - we stand 2^{nd} with Accuracy@1@Top1 0.42, 0.4182, 0.3345, 0.2583 and 0.0562 respectively. Lastly in the 3^{rd} position, we have English, Italian, French and Catalan with 0.5157, 0.4042, 0.3661 and 0.2247 Accuracy@1@Top1 respectively. The detailed explanation of the evaluation metrics used for LS is available at (Saggion et al., 2023). Test results for LS are shown in Table 2.

Trial results of LCP and LS are available in Table 4 and 5 of Appendix.

6 Error Analysis

For LCP the highest mean absolute and squared error are 0.2089 and 0.0589 for French and the lowest mean absolute and squared error are 0.1018 and 0.0168 for Sinhala. This is an acceptable mar-

	Test Scores (Target Language)				
Language	**Pearson**	**Spearman**	**MAE**	**MSE**	**R2**
Catalan	0.1549	0.1574	0.1462	0.0318	-0.3378
English (CompLex)	0.8497	0.7984	0.1137	0.0175	0.5247
Filipino	0.2823	0.2767	0.1164	0.0227	-0.0457
French	0.3193	0.3207	0.2089	0.0589	0.0484
German	0.1402	0.1473	0.1567	0.0413	-0.5279
Japanese	0.1775	0.1827	0.1363	0.0270	0.0241
Sinhala	0.1246	0.1303	0.1018	0.0168	-0.0370
Spanish	0.2438	0.1984	0.1630	0.0379	-0.0731
All Combined	0.3494	0.3642	0.1464	0.0331	0.1094

Table 2: Test Results of LCP (Weighted Ensemble of the Models Used for Corresponding Languages in Trial Phase)

Language	Models	A@1@Top1	A@2@Top1	A@3@Top1	MacAvgPrec@1	MacAvgPrec@3	MacAvgPrec@5	MacAvgPrec@10	MAP@3	MAP@5	MAP@10
Catalan	GPT4-turbo	**0.2247**	**0.3056**	**0.328**	**0.537**	**0.7101**	**0.7573**	**0.8044**	**0.362**	**0.2641**	**0.1582**
	Top10Suggestion	0.0651	0.1191	0.1595	0.2426	0.5191	0.6404	0.755	0.172	0.1408	0.0893
English	GPT4-turbo	**0.5157**	**0.635**	**0.6894**	**0.7491**	**0.8754**	**0.907**	**0.928**	**0.513**	**0.3691**	**0.2095**
	Top10Suggestion	0.1929	0.3228	0.4157	0.335	0.6315	0.7649	0.8649	0.2339	0.1869	0.1106
Filipino	GPT4-turbo	**0.0562**	**0.0632**	**0.0685**	**0.2934**	**0.3989**	**0.4358**	**0.4868**	**0.1395**	**0.0916**	**0.0491**
	Top10Suggestion	0.0157	0.0228	0.0245	0.0807	0.1842	0.2859	0.3859	0.0449	0.0338	0.0201
French	GPT4-turbo	**0.3661**	**0.4559**	**0.514**	**0.7411**	**0.8679**	**0.889**	**0.9154**	**0.5148**	**0.3946**	**0.2447**
	Top10Suggestion	0.0845	0.1672	0.2394	0.2271	0.5316	0.6971	0.8257	0.1725	0.149	0.1023
German	GPT4-turbo	**0.42**	**0.5043**	**0.5817**	**0.6414**	**0.7908**	**0.8312**	**0.8558**	**0.4002**	**0.2874**	**0.1631**
	Top10Suggestion	0.1192	0.2228	0.3	0.2578	0.5491	0.6666	0.7982	0.1852	0.1463	0.092
Italian	GPT4-turbo	**0.4042**	**0.5641**	**0.6309**	**0.7346**	**0.8822**	**0.9244**	**0.9402**	**0.4615**	**0.3328**	**0.1966**
	Top10Suggestion	0.1546	0.2724	0.3567	0.3567	0.6625	0.7855	0.8717	0.246	0.1965	0.1242
Japanese	GPT4-turbo	**0.2583**	**0.3708**	**0.4393**	**0.5413**	**0.6801**	**0.7223**	**0.7627**	**0.3618**	**0.2599**	**0.1529**
	Top10Suggestion	0.1195	0.2144	0.2847	0.3075	0.5817	0.6731	0.7469	0.2144	0.171	0.1107
Sinhala	GPT4-turbo	**0.2284**	**0.2829**	**0.3163**	**0.311**	**0.4165**	**0.4815**	**0.536**	**0.1387**	**0.0894**	**0.0469**
	Top10Suggestion	0.13	0.2372	0.3057	0.195	0.3848	0.4639	0.5272	0.1147	0.0759	0.0394
Spanish	GPT4-turbo	**0.4182**	**0.5362**	**0.6087**	**0.801**	**0.9173**	**0.9477**	**0.9612**	**0.5987**	**0.4653**	**0.2853**
	Top10Suggestion	0.236	0.3558	0.4704	0.5919	0.86	0.9106	0.9392	0.4371	0.3542	0.2244
All Combined	GPT4-turbo	**0.3345**	**0.4291**	**0.4828**	**0.5934**	**0.7276**	**0.7695**	**0.803**	**0.379**	**0.2754**	**0.1614**
	Top10Suggestion	0.1331	0.2261	0.2999	0.2876	0.5374	0.6467	0.7386	0.1981	0.1561	0.0971

Table 3: Test Results of LS (Top 10 Suggestions are Selected from the Output of GPT4-turbo and the Models Used for Corresponding Languages in Trial Phase)

gin of error when we are training a model with cross-lingual data and testing with language specific data. This is also a reason of getting negative R2 score for 4 languages which testifies that the data struggles to fit the regression model for those languages.

For LS, zero-shot prompting by GPT4-turbo alone provides the best result but when we try to find the best 10 suggestions from the set of suggestions generated by the BERT based models and GPT4-turbo together, the result significantly decreases. This was because the target token in the sentence varied be in different grammatical form. Therefore, finding proper simplified suggestions that fits the context proves to be a struggle for the BERT based model.

7 Conclusion

Our team *GMU*'s approaches in MLSP 2024 shared task achieved competitive results across multiple languages for both the LCP and LS sub-tasks. The weighted ensemble technique based on transformer models proved effective for LCP, while GPT-4 zero-shot prompting excelled at LS. The multilingual nature of this shared task also highlights the importance of developing techniques that can generalize across languages.

One key limitation of our approach is the reliance on cross-lingual transfer due to limited language-specific training data for most languages. While this allowed sharing resources across languages, having larger datasets to each language could potentially boost performance. Additionally, the error analysis revealed some remaining challenges in handling complex word expressions and phrases during LS. Further improvements in modeling could address these cases more effectively for MLSP in future.

Acknowledgements

We express our gratitude to the organizers of the BEA 2024 workshop for facilitating this insightful shared task and providing the multilingual datasets that enabled our research. We also thank the annotators for successfully annotating this large dataset.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Farhad Akhbardeh, Cecilia Ovesdotter Alm, Marcos Zampieri, and Travis Desell. 2021. Handling extreme class imbalance in technical logbook datasets. In *Proceedings of ACL (IJCNLP)*.

Dennis Aumiller and Michael Gertz. 2023. Unihd at tsar-2022 shared task: Is compute all we need for lexical simplification? *arXiv preprint arXiv:2301.01764*.

Stefan Bott, Horacio Saggion, Nelson Peréz Rojas, Martin Solis Salazar, and Saul Calderon Ramirez. 2024. Multils-sp/ca: Lexical complexity prediction and lexical simplification resources for catalan and spanish.

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

Codegram. 2020. Calbert: A catalan language model. https://huggingface.co/codegram/calbert-base-uncased.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Jan Christian Blaise Cruz and Charibeth Cheng. 2021. Improving large-scale language models and resources for filipino. *arXiv preprint arXiv:2111.06053*.

Dbmdz. 2020a. BERT-base italian cased model. https://huggingface.co/dbmdz/bert-base-italian-cased.

Dbmdz. 2020b. German bert model. https://huggingface.co/dbmdz/bert-base-german-uncased.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Vinura Dhananjaya, Piyumal Demotte, Surangika Ranathunga, and Sanath Jayasena. 2022. Bertifying sinhala–a comprehensive analysis of pre-trained language models for sinhala text classification. *arXiv preprint arXiv:2208.07864*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2020. Language-agnostic bert sentence embedding. *arXiv preprint arXiv:2007.01852*.

Daniel Ferrés and Horacio Saggion. 2022. Alexsis: A dataset for lexical simplification in spanish. In *Proceedings of LREC*.

Goran Glavaš and Sanja Štajner. 2015. Simplifying lexical simplification: Do we need simplified corpora? In *Proceedings of ACL-IJCNLP*.

Sian Gooding and Ekaterina Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of BEA*.

Sian Gooding and Ekaterina Kochmar. 2019. Recursive context-aware lexical simplification. In *Proceedings of EMNLP-IJCNLP*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *Proceedings of ICLR*.

Hang Le, Loïc Vial, Jibril Frej, Vincent Segonne, Maximin Coavoux, Benjamin Lecouteux, Alexandre Allauzen, Benoît Crabbé, Laurent Besacier, and Didier Schwab. 2020. Flaubert: des modèles de langue contextualisés pré-entraînés pour le français. In *Proceedings of TALN*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.

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. *arXiv preprint arXiv:2305.12000*.

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

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022b. Alexsis-pt: A new resource for portuguese lexical simplification. *arXiv preprint arXiv:2209.09034*.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023c. Lexical complexity prediction: An overview. *ACM Computing Surveys*, 55(9):1–42.

OpenAI. 2023. GPT-4 technical report. https://arxiv.org/abs/2303.08774.

Gustavo Paetzold and Lucia Specia. 2016a. Semeval 2016 task 11: Complex word identification. In *Proceedings of SemEval*, pages 560–569.

Gustavo Paetzold and Lucia Specia. 2016b. Sv000gg at semeval-2016 task 11: Heavy gauge complex word identification with system voting. In *Proceedings of SemEval*.

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. Deepblueai at semeval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of SemEval*.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the AAAI*.

Luz Rello, Roberto Carlini, Ricardo Baeza-Yates, and Jeffrey P Bigham. 2015. A plug-in to aid online reading in spanish. In *Proceedings of W4A*.

Horacio Saggion and Graeme Hirst. 2017. *Automatic text simplification*. Springer.

Horacio Saggion, Sanja Štajner, Daniel Ferrés, Kim Cheng Sheang, Matthew Shardlow, Kai North, and Marcos Zampieri. 2023. Findings of the tsar-2022 shared task on multilingual lexical simplification. *arXiv preprint arXiv:2302.02888*.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *Proceedings of SRW*.

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 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 READI*.

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

Matthew Shardlow, Michael Cooper, and Marcos Zampieri. 2020b. Complex: A new corpus for lexical complexity prediction from likert scale data. *arXiv preprint arXiv:2003.07008*.

Matthew Shardlow, Richard Evans, Gustavo Henrique Paetzold, and Marcos Zampieri. 2021. Semeval-2021 task 1: Lexical complexity prediction. *arXiv preprint arXiv:2106.00473*.

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*.

Tohoku-NLP. 2020. BERT-base japanese model. https://huggingface.co/tohoku-nlp/bert-base-japanese.

Peniel Whistely, Sandeep Mathias, and Galiveeti Poornima. 2022. Presiuniv at tsar-2022 shared task: Generation and ranking of simplification substitutes of complex words in multiple languages. In *Proceedings of TSAR*.

Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and Malak Abdullah. 2021. Just-blue at semeval-2021 task 1: Predicting lexical complexity using bert and roberta pre-trained language models. In *Proceedings of SemEval*.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo H Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. *arXiv preprint arXiv:1804.09132*.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex word identification: Challenges in data annotation and system performance. In *Proceedings of NLPTEA*.

A Appendix

Language	Models	Validation Scores (Combined Dataset)				
		Pearson	**Spearman**	**MAE**	**MSE**	**R2**
Catalan	mBERT	0.5839	0.5965	0.1296	0.0272	0.2676
	XLM-R	0.4131	0.3881	0.1496	0.0333	0.1051
	calBERT	0.4724	0.4868	0.1384	0.031	0.1653
English (All Trial)	mBERT	0.5926	0.57	0.1422	0.0297	0.2012
	XLM-R	0.4245	0.4345	0.1496	0.0335	0.0982
	BERT	0.4396	0.4489	0.1593	0.0349	0.0621
	RoBERTa	0.5418	0.5525	0.1375	0.0266	0.2848
	DeBERTa	0.5437	0.5234	0.138	0.0276	0.257
English (CompLex)	BERT	0.7732	0.751	0.1478	0.0288	0.2604
	RoBERTa	0.6454	0.7072	0.156	0.0325	0.1635
	DeBERTa	0.8144	0.7434	0.1486	0.0269	0.3094
Filipino	mBERT	0.5814	0.577	0.1299	0.027	0.2744
	XLM-R	0.4447	0.4368	0.1453	0.031	0.1665
	RoBERTa-tagalog	0.4162	0.3686	0.1504	0.0342	0.0807
French	mBERT	0.5703	0.6264	0.1402	0.027	0.2734
	XLM-R	0.4588	0.4576	0.1466	0.0306	0.1778
	flauBERT	0.3742	0.3068	0.1485	0.0322	0.1345
German	mBERT	0.6061	0.6159	0.1382	0.0263	0.29933
	XLM-R	0.4586	0.4481	0.1469	0.0296	0.2043
	germanBERT	0.4511	0.4669	0.1415	0.0306	0.1778
Italian	mBERT	0.6196	0.5757	0.1225	0.0244	0.3441
	XLM-R	0.4934	0.4625	0.144	0.0297	0.2003
	italianBERT	0.5577	0.5419	0.1353	0.0262	0.2946
Japanese	mBERT	0.5551	0.5568	0.1378	0.0301	0.1914
	XLM-R	0.5479	0.5355	0.1422	0.028	0.2462
	japaneseBERT	0.4286	0.4285	0.1521	0.0341	0.083
Sinhala	mBERT	0.5948	0.6375	0.1333	0.0263	0.2929
	XLM-R	0.4396	0.4569	0.1414	0.0304	0.181
	sinhalaBERTo	0.3766	0.4027	0.1568	0.0337	0.0923
Spanish	mBERT	0.5412	0.5861	0.136	0.0282	0.2428
	XLM-R	0.5119	0.5022	0.1391	0.0289	0.2225
	spanishBERT	0.4141	0.3909	0.1559	0.0328	0.1188
All Combined	mBERT	0.4511	0.495	0.1546	0.0326	0.1223
	XLM-R	0.4588	0.4576	0.1466	0.0306	0.1778
	calBERT	0.4044	0.4069	0.1517	0.0326	0.1226
	DeBERTa	0.4511	0.4745	0.1482	0.0318	0.1454
	RoBERTa-tagalog	0.4626	0.4588	0.1564	0.0354	0.0489
	flauBERT	0.4416	0.4236	0.1583	0.036	0.0306
	germanBERT	0.4383	0.4261	0.1531	0.0345	0.718
	italianBERT	0.5577	0.5419	0.1353	0.0262	0.2946
	japaneseBERT	0.4183	0.4461	0.1543	0.0347	0.0675
	BERTimbau	0.5274	0.5701	0.1306	0.0271	0.2697
	sinhalaBERTo	0.4249	0.4622	0.156	0.0338	0.0919
	spanishBERT	0.4679	0.4499	0.1535	0.0356	0.0433

Table 4: Trial Results of LCP

Language	Models	A@1@Top1	A@2@Top1	A@3@Top1	MacAvgPrec@1	MacAvgPrec@3	MacAvgPrec@5	MacAvgPrec@10	MAP@3	MAP@5	MAP@10
	mBERT	0.0666	0.1333	0.1333	0.1	0.1666	0.1666	0.2	0.0999	0.0866	0.0437
Catalan	XLM-R	0.0066	0.2333	0.3	0.1333	0.3666	0.4	0.4	0.1351	0.1094	0.0547
	calBERT	0.0666	0.1666	0.1666	0.1	0.1666	0.2333	0.2333	0.0999	0.0893	0.0446
	GPT4-turbo	**0.4666**	**0.4666**	**0.5**	**0.4666**	**0.6**	**0.7333**	**0.7333**	**0.2407**	**0.1634**	**0.0888**
	Top10Suggestion	0.2	0.2666	0.3333	0.2333	0.5	0.5666	0.6666	0.1259	0.0932	0.0524
	mBERT	0.1	0.2	0.26	0.2	0.4666	0.5	0.6333	0.1481	0.1012	0.0577
	XLM-R	0.1	0.1666	0.2666	0.1666	0.4333	0.5666	0.6333	0.1222	0.0796	0.0484
English	BERT	0.1666	0.1666	0.2	0.3	0.5333	0.5666	0.7	0.174	0.1297	0.0766
	RoBERTa	0.066	0.2	0.2333	0.2	0.4666	0.6333	0.7333	0.1648	0.1335	0.0815
	DeBERTa	0.1666	0.1666	0.1666	0.2333	0.2333	0.2333	0.2333	0.2	0.1446	0.0733
	GPT4-turbo	**0.4**	**0.5**	**0.5666**	**0.7**	**0.8**	**0.8666**	**0.8666**	**0.4444**	**0.3136**	**0.1728**
	Top10Suggestion	0.1333	0.2666	0.3666	0.2666	0.6666	0.6666	0.6666	0.174	0.1224	0.0612
	mBERT	0.0333	0.0666	0.0666	0.0333	0.0666	0.0666	0.1	0.0166	0.01	0.0054
Filipino	XLM-R	0.1333	0.2	0.2	0.1333	0.2	0.2	0.2333	0.0888	0.0533	0.0271
	RoBERTa-tagalog	0.2	0.2666	0.3	0.2333	0.3333	0.4	0.4333	0.1037	0.0652	0.0352
	GPT4-turbo	**0.3666**	**0.3666**	**0.3666**	**0.4**	**0.4333**	**0.4666**	**0.5**	**0.1611**	**0.1053**	**0.055**
	Top10Suggestion	0.0666	0.1333	0.2333	0.0666	0.2333	0.3333	0.4	0.0555	0.0373	0.0206
	mBERT	0.2	0.3333	0.4	0.2666	0.4333	0.5	0.5	0.1611	0.0996	0.052
French	XLM-R	0.1666	0.3	0.3666	0.2333	0.4333	0.5	0.5333	0.1185	0.0711	0.0402
	flauBERT	0.0166	0.0266	0.0366	0.0166	0.0266	0.0366	0.0366	0.0107	0.0071	0.0046
	GPT4-turbo	**0.5**	**0.6333**	**0.6666**	**0.7**	**0.8**	**0.8**	**0.8**	**0.3759**	**0.2305**	**0.1169**
	Top10Suggestion	0.2	0.2333	0.2333	0.2333	0.4333	0.5666	0.7333	0.1296	0.0927	0.0518
	mBERT	0.0333	0.0666	0.0666	0.0333	0.0666	0.0666	0.1	0.0287	0.019	0.0111
German	XLM-R	0.0333	0.0666	0.1333	0.1	0.1666	0.3	0.3333	0.0446	0.03	0.0168
	germanBERT	0.1666	0.2	0.2333	0.1333	0.2333	0.2333	0.2333	0.0814	0.0592	0.0299
	GPT4-turbo	**0.6**	**0.8666**	**0.9666**	**0.7333**	**0.9**	**0.9**	**0.9**	**0.3944**	**0.2603**	**0.137**
	Top10Suggestion	0.0333	0.0666	0.1666	0.0666	0.2666	0.4333	0.7	0.0555	0.053	0.0337
	mBERT	0.0333	0.0666	0.0666	0.0333	0.1	0.1333	0.2	0.0222	0.015	0.0092
Italian	XLM-R	0.1	0.1	0.1	0.1333	0.1333	0.1666	0.2333	0.0444	0.028	0.0158
	italianBERT	0.2333	0.3666	0.4	0.2666	0.4666	0.5333	0.6	0.1537	0.1038	0.0518
	GPT4-turbo	**0.5**	**0.6**	**0.7**	**0.3518**	**0.2334**	**0.1267**	**0.6**	**0.3518**	**0.2334**	**0.1267**
	Top10Suggestion	0.1666	0.2	0.2333	0.2	0.3666	0.5666	0.7666	0.1259	0.0905	0.0566
	mBERT	0.0666	0.0666	0.0666	0.0666	0.0666	0.0666	0.0666	0.0518	0.0427	0.0213
Japanese	XLM-R	0.0666	0.0666	0.0666	0.0666	0.0666	0.0666	0.0666	0.0518	0.0427	0.0213
	japaneseBERT	0.1	0.1333	0.1666	0.1333	0.1666	0.1666	0.1666	0.137	0.0955	0.0477
	GPT4-turbo	**0.4333**	**0.4666**	**0.4666**	**0.5333**	**0.6333**	**0.7333**	**0.8**	**0.2629**	**0.1767**	**0.0936**
	Top10Suggestion	0.0333	0.0666	0.1	0.0666	0.1333	0.2333	0.5	0.0407	0.0321	0.0226
	mBERT	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.0833	0.0599	0.0299
Sinhala	XLM-R	0.0333	0.0666	0.0666	0.1	0.0133	0.0133	0.0233	0.0481	0.0288	0.0159
	sinhalaBERTo	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0333	0.0111	0.0066	0.0033
	GPT4-turbo	**0.3666**	**0.5**	**0.5666**	**0.5333**	**0.7333**	**0.7666**	**0.8333**	**0.2851**	**0.1757**	**0.0961**
	Top10Suggestion	0.2	0.3333	0.4	0.3666	0.5666	0.6	0.7666	0.2037	0.1412	0.0786
	mBERT	0.1	0.1333	0.1333	0.1333	0.1333	0.1333	0.1333	0.1018	0.0744	0.0418
Spanish	XLM-R	0.1666	0.3	0.3666	0.2333	0.4333	0.5	0.5333	0.1185	0.0711	0.0402
	spanishBERT	0.2666	0.3333	0.4333	0.3333	0.5666	0.6666	0.7333	0.2055	0.133	0.0698
	GPT4-turbo	**0.4**	**0.6333**	**0.7666**	**0.6333**	**0.8666**	**0.9333**	**0.9333**	**0.4018**	**0.2721**	**0.1433**
	Top10Suggestion	0.2666	0.3333	0.4666	0.3	0.6333	0.7333	0.7666	0.1888	0.132	0.0716
	mBERT	0.0233	0.04	0.0566	0.0433	0.1	0.11	0.1533	0.0287	0.019	0.0111
	XLM-R	0.0466	0.0833	0.1033	0.0866	0.1533	0.2033	0.2366	0.0446	0.03	0.0168
	calBERT	0.0333	0.0366	0.0366	0.0333	0.0366	0.0366	0.0366	0.03	0.02	0.01
	DeBERTa	0.0333	0.0366	0.0366	0.0333	0.0366	0.0366	0.0366	0.03	0.02	0.01
	flauBERT	0.0166	0.0266	0.0366	0.0166	0.0266	0.0366	0.0366	0.0107	0.0071	0.0046
All Combined	germanBERT	0.0166	0.02	0.0233	0.02	0.03	0.0333	0.0333	0.0081	0.0056	0.0028
	italianBERT	0.0266	0.04	0.0433	0.0333	0.0533	0.0666	0.0766	0.0175	0.0118	0.006
	japaneseBERT	0.0333	0.0366	0.04	0.0366	0.04	0.04	0.04	0.0303	0.0202	0.0101
	BERTimbau	0.0066	0.0066	0.0066	0.0066	0.0066	0.0066	0.0066	0.0022	0.0013	0.0006
	sinhalaBERTo	0.0033	0.0033	0.0033	0.0166	0.03	0.04	0.0533	0.0083	0.0062	0.0035
	spanishBERT	0.0033	0.0033	0.0033	0.0066	0.0166	0.02	0.03	0.0033	0.0022	0.0012
	GPT4-turbo	**0.39**	**0.48**	**0.5333**	**0.5966**	**0.7433**	**0.7933**	**0.8366**	**0.3122**	**0.2088**	**0.1111**
	Top10Suggestion	0.1166	0.2166	0.2933	0.1833	0.45	0.5833	0.6833	0.1248	0.0942	0.0526

Table 5: Trial Results of LS