Archaeology at MLSP 2024: Machine Translation for Lexical Complexity Prediction and Lexical Simplification

Petru Theodor Cristea
petru-theodor.cristea@s.unibuc.ro

Sergiu Nisioi
sergiu.nisioi@unibuc.ro

Human Language Technologies Research Center
Faculty of Mathematics and Computer Science
University of Bucharest

Abstract

We present the submissions of team Archaeology for the Lexical Simplification and Lexical Complexity Prediction Shared Tasks at BEA2024. Our approach for this shared task consists in creating two pipelines for generating lexical substitutions and estimating the complexity: one using machine translated texts into English and one using the original language. For the LCP subtask, our xgb regressor is trained with engineered features (based primarily on English language resources) and shallow word structure features. For the LS subtask we use a locally-executed quantized LLM to generate candidates and sort them by complexity score computed using the pipeline designed for LCP. These pipelines provide distinct perspectives on the lexical simplification process, offering insights into the efficacy and limitations of employing Machine Translation versus direct processing on the original language data.

Our results and experiments are released at https://github.com/senisioi/MLSP_Participants

1 Introduction

In the realm of Natural Language Processing (NLP), the twin challenges of lexical complexity prediction and language simplification play pivotal roles in advancing text comprehension and promoting accessibility. Lexical complexity prediction refers to the difficulty of understanding phrases based on their lexical features, while simplification aims to enhance accessibility by offering simplified, easier-to-understand alternatives. The importance of addressing these challenges is underscored by their wide-ranging implications across various domains (Gooding, 2022; North et al., 2023; Saggion et al., 2023).

Our approach is guided by the idea to extend such methods beyond the languages that currently have available data sets or corpora; thus, our first set of submissions to the 2024 MLSP Shared Task (Shardlow et al., 2024a) uses machine translation to translate all datasets and languages into English, which has been the central language of text simplification and complexity research in recent years (North et al., 2023). Both the lexical simplification (LS) and the lexical complexity prediction (LCP) pipelines are using only data in English in this case [1].

The second approach is trained on the original texts as released by Shardlow et al. (2024b) and uses an LCP pipeline trained with language-independent hand-crafted features such as word length, syllables, vowels, etc. and a regression method trained on the small trial data from the original language.

For generating candidates for lexical simplification, we have opted for an LLM that can be run locally using a quantized version of OpenHermes 2.5 based on Mistral (Jiang et al., 2023) that has been fine-tuned on code. According to the authors[2], the model was trained on a good ratio of code instruction (7-14% of the total dataset) that boosted several noncode benchmarks, including TruthfulQA, AGIEval, and GPT4All suite. The quantized LLM is not inherently multilingual, however, in our small-scale tests we have observed some ability to generate simplification candidates for non-English language,

The LLM we used to generate the alternatives does not guarantee the correct form of the generated alternative and this problem is amplified by using Machine Translation to get the phrases in original languages, which could incorrectly translate words without context. Regarding Machine

[1] Because of a bug in our submission code, the first LCP submission was run with a LCP regression model trained purely on English data with no other language involved.

[2] https://huggingface.co/teknium/OpenHermes-2.5-Mistral-7B

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 610–617

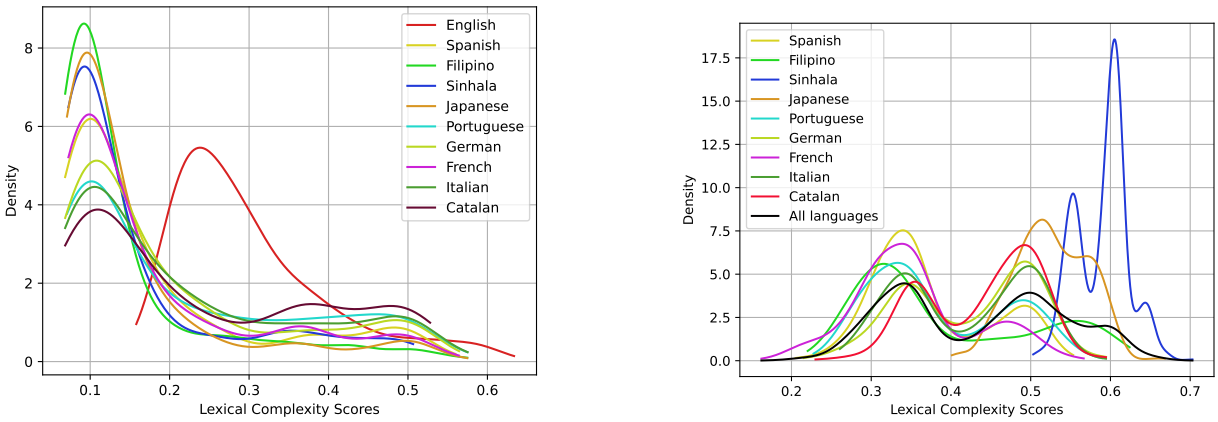

Figure 1: Density plots of LCP submitted scores predicted on the test set using (a) translated sentences and Englsh original sentences and (b) original-language texts. The model run on original texts generally observes two peaks, one with smaller values for simple words and one with larger values associated with complex words. Back-translated words show quite a different pattern with only a single peak of words marked as simple.

		es	fil	si	ja	pt	de	fr	it	ca	en
Trans.	**mean**	1.36	1.52	1.49	1.74	1.32	1.23	1.39	1.29	1.52	-
	max	8.2	6	12.71	8.75	8.25	5	6.5	4.3	6.3	-
	empty	0.2	0	1.3	0.4	0	0.4	0.5	0.5	0.7	-
Orig.	**mean**	1.27	1.33	1.14	2.3	1.22	1.11	1.29	1.19	1.39	1.31
	max	3.44	4.1	3	8.4	5.1	3.2	5.5	6.2	6	3.4
	empty	14.2	2.8	3.2	0.7	8.6	3.2	3.7	12.5	2	5.3

Table 1: Average lengths of multi-word expressions that our systems suggested as alternative lexical simplifications. Row empty indicate the percentage of empty suggestions for each language. The upper part of the table shows that the number of empty suggesions of OpenHermes2.5 are low for texts translated into English, but the average number of new words is higher than for prompts using texts in the original language.

Translation, we used DeepL[3] for French, Spanish, Japanese, German, Portuguese, Italian, and Google Translate for Sinhala, Catalan, and Filipino, thus obtaining only sentences in English to be able to effectively apply feature extraction.

In many cases, during the translation process, contextual information or expressions may be lost, significantly affecting the correlation between features. Table 1 shows the average number of multi-word expressions introduced by the translation step or by the predictions of the LLM model. Our LLM suggested in many cases empty strings, we did not check for those cases. As it stands, 14% of Spanish 12% of Italian are empty, however the overall scores with LLMs for these languages exceed the scores with MT (by a small margin). With MT, the number of empty suggestions is considerably smaller, but strangely enough 5% of original English sugestions are empty.

The LLM works better at generating candidates directly using English translations as the number of empty candidates is lower; however, the actual candidates generated tend to be multi-word expressions instead of simple lexical substitutions.

In summary, combining our approach for predicting lexical complexity and simplification in a unified framework may not be the best solution for text comprehension, but it can provide a source of interesting results for different languages.

2 Lexical Complexity Prediction

For lexical complexity prediction we reuse an approach that has been previously tested at the LCP2021 Shared Task (Shardlow et al., 2021) that obtained a Pearson correlation of .75 using a regression method trained on hand-crafted features.

Shallow Word Structure Features

We believe that this set of characteristics is as much as possible language independent when additional Latin-alphabet transliterations are used:

- character length of word

[3] https://deepl.com

- zipf_frequency from wordfreq library (Speer, 2022) (except for Sinhala)
- is title (not applicable for non-Latin glyphs)
- number of vowels (not applicable for non-Latin glyphs)
- number of syllables from pyphen library[4] (not applicable for non-Latin glyphs)

Medical Research Council Psycholinguistic Database

The MRC database (Wilson, 1988) is one of the most widely used feature source for LCP (Devlin, 1998; Yimam et al., 2018; Shardlow et al., 2021; North et al., 2023) demonstrating over three decades of high usability (Scott et al., 2019) built on top of word annotations (Thorndike and Lorge, 1944) and highlighting the necessity of such databses beyond the English language. Each lexical item is lemmatized using the spacy English large model (Montani et al., 2023) and searched in the database. The features we employ are:

- aoa - age of acquisition 1-7 Likert scale multiplied by 100 (Carroll and White, 1973; Gilhooly and Logie, 1980)
- conc - concreteness rating from the methodology of Spreen and Schulz (1966); Gilhooly and Logie (1980): "words referring to objects, materials, or persons were to receive a high concreteness rating, and words referring to abstract concepts that could not be experienced by the senses were to receive a low concreteness rating"
- fam - (Noble, 1953; Gilhooly and Logie, 1980) familiarity rating (100-700)
- imag - imagability / imagery rating (Paivio et al., 1968; Gilhooly and Logie, 1980): "words arousing images most readily were to be rated 7, and words arousing images with great difficulty or not at all were to be rated 1" scores multiplied by 100
- meanp - meaningfulness - defined as "the mean number of associations given in a 30-sec production period" from the Paivio et al. (1968)
- meanc - meaningfulness - Colorado Norms (Toglia and Battig, 1978) obtained using a different methodology from meanp (Wilson, 1988)
- brown_freq - Brown verbal frequency (Brown, 1984)
- Kucera-Francis number of categories, samples and frequency (Kučera et al., 1967)
- tl_freq - Thorndike-Lorge written frequency (Thorndike and Lorge, 1944)

Syntactic Features

For all lanugages except Filipino and Sinhala, we load spacy medium-sized models (Montani et al., 2023) using the latest version available. The only syntactic features are the number of immediate children in syntactic dependency parse. We use spacy here to introduce additional boolean features such as: is entity, is sentence start, is sentence end. Such words could be markers of conceptual complexity (Stajner et al., 2020).

WordNet Features

Similar to the MRC features, these are only available for English. We access WordNet (Miller, 1994) from NLTK (Bird et al., 2009) to extract the number of synsets, hypernyms, and hyponyms.

External Lists

The system also incorporates external datasets, such as the Dale-Chall (Dale and Chall, 1948) list to create a boolean feature set. Furthermore, additional frequency data is derived from non-native speakers in the European Parliament (Nisioi et al., 2016).

Similar features to ours have been used for the CWI identification Shared Task in 2018 (Gooding and Kochmar, 2018) obtaining excellent results on a related task.

Regression Model

We use an XGBoost Regressor (Chen and Guestrin, 2016), which operates within a gradient boosting framework, sequentially training weak learners to minimize a specified loss function. For this task, we do not employ hyperparameter tuning. All features are passed through a scikit learn standard scaler (Pedregosa et al., 2011) which standardizes the features to zero mean and a standard deviation of one. Although it might have been advisable to check which features are good for scaling, we did not proceed with this step, but rather passed all the features (including the Boolean ones) through the scaler.

We train our model on the English dataset released during LCP2021 (Shardlow et al., 2021) concatenated with all the languages from the current year's shared task (Shardlow et al., 2024b).

[4] https://doc.courtbouillon.org/pyphen

We use the same amount of features for all languages, which presents an interesting corner case where words with similar forms are found in the English-only resources. Such examples can be in Filipino: *amin* (En. *us*), *ate* (En.: *sister*) or the French words: *notice, question, coach*, Portuguese words such as: *bases, rigor*, and Catalan: *decimals*. This idea might point to a future research direction to explore where false friends, borrowings, and cognates (Dinu et al., 2023) could have the ability to preserve lexical traits across languages that have a history of contact.

3 Lexical simplification

For lexical simplification, we employ the locally run quantized OpenHermes 2.5 based on Mistral (Jiang et al., 2023) using llama-cpp[5] and langchain (Chase, 2022) libraries. The context contains the entire sentence and the target word, and the model is prompted to generate a json with potential replacement candidates. We run the model on the English-translated texts and the results are then back-translated into the initial language. The model prompt is as follows: *This sentence "TRANSLATION" is a translation of "ORIGINAL" and the word "TRANSLATED_WORD" is a translation of "ORIGINAL_WORD" Provide a list of 10 alternative simpler words (as a json object) that a child would understand easily to replace the word ""TRANSLATED_WORD"" in the following sentence. It is mandatory for pattern of the answer to be displayed as a JSON with words as keys and complexity scores as values with all the 10 alternatives.*

The second set of submissions is generated with the model running on the original language date. Nevertheless, it is imperative to acknowledge that the model's capability is constrained when handling multilingual data, often leading to hallucinations. The prompt used for original language data is: *Provide a list of 10 alternative simpler words (as a json object) that a child would understand easily to replace the word "ORIGINAL_WORD" in the context of the following sentence. It is mandatory to use suitable meanings for the context of the sentence and for the pattern of the answer to be displayed as a JSON with words as keys and complexity scores as values with all the 10 alternatives. Provide only words in "LANGUAGE". Sentence: "ORIGINAL. Here are some possible synonyms: "SYNONYMS"* The synonyms are given in the context extracted from ConceptNet (Speer et al., 2017) with a quick request to the API.

4 Results

Our first set of submissions (suffixed with "*_1.tsv") contain LCP only run with English-only models and LS predictions run on translated texts. The translation model tends to increase the number of words, as seen in Table 1 because we translate words out of their context, and some translations might not end up being found in the text mot-à-mot. We identify a target word in the context of the sentence (which will become our target for LCP) by doing a proximal cosine similarity search using spacy embeddings.

Our second set of submissions (suffixed with "*_2.tsv") are LCP predictions run on the original target words. Figure 1 shows the density plots of the predictions on the test set. The translations-to-English complexity scores (a) are in the same range for all langauges (except for Sinhala) while the predictions on the original texts (b) show more divergent patterns due to different features available for each language. Here we only report the results on the LCP task as these are the only ones that proved to be competitive in the shared task. For a complete set of results we point the reader to the official task page[6].

We perform several experiments on the trial data to verify which features of the original language have the strongest correlation with the complexity scores provided. This should give us a rough idea of the features that contribute the most to the final prediction. The correlations computed on the trial data are reported as a proxy for potential feature impact; given the small sample size, they may also show accidentally high values.

For **English**-language predictions (originals and translations included) word frequency achieves between -.64 and -.7 ρ, followed by MRC features, such as the Kucera-Francis (Kučera et al., 1967) number of categories feature (-.55 ρ). MRC features are generally well correlated among each other. Sparse features such as Dale-Chall, EuroParl frequency, hyponyms, synsets, and other MRC-based features contribute significantly because they create a boundary between words that are not in the external resource (feature value 0) and words

[5] https://github.com/ggerganov/llama.cpp

[6] https://sites.google.com/view/mlsp-sharedtask-2024/home

that are (value > 0). For **French**: word frequency (-.4 ρ) and the number of immediate children in syntactic dependency parse (.4 ρ) show the best correlation with the complexity annotations. **German** trial data shows that word frequency is at -.76 ρ followed by character length .46 ρ, this could be a lucky coincidence from the small size or the distribution in the trial data. Similarly, **Filipino** (-.61 ρ), **Spanish** (-.6 ρ), **Portuguese** (-.71 ρ), and Italian (-.63 ρ) show relatively good correlations between complexity scores and word frequency. **Catalan** shows weak correlations of all individual characteristics (-.2 ρ on frequency), which also confirms our overall scores (Shardlow et al., 2024a), and so is **Japanese** (-.58 ρ). **Sinhala** is a special case of language where we do not use word frequency nor other resource and the only relevant features is the character length (-.3 ρ).

5 Conclusions

Translating documents into English and making lexical simplification predictions using translated texts introduces noise and severely limits the ability of the model to produce coherent substitutions, especially since our approaches with translations did not take into consideration the proper morphological form of the substitution or of the original word. Our results (reported in the Appendix) show that complexity prediction is significantly affected by the translation as much as lexical simplification. Our MT approach was surpassed by models trained only on English data, which appear to have a better ability to generate good LCP predictions on other languages (especially Latin-script languages or languages with a historical contact). This approach can yield decent results, achieving similar correlations to models trained directly on the source language or models using LLMs and transformer features (Shardlow et al., 2024a). Last but not least, we conclude that frequency- and string-based approaches might be a powerful alternative for LCP on low-resource languages.

6 Limitations

We observe several limitations of our approaches:

- potential innacuracies stemming from the translation system
- using MT for the low resource setting could be detrimental to the development of resources in the original language; translating all languages into English is not always feasible and depends on cultural factors, availability of resources and so on
- the performance of the MT systems themselves can vary depending on factors such as language pair, domain specificity, and the quality of the training data; in our case we have used closed-source models which is not desirable for open research
- our work is focused only on a single LLM that is English-centric, however the model was not able to generate suggestions that are in the correct tense or syntactic agreement with the rest of the sentence

7 Acknowledgements

This work has been supported by POCIDIF project in Action 1.2 "Romanian Hub for Artificial Intelligence".

References

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".

Gordon DA Brown. 1984. A frequency count of 190,000 words in the london-lund corpus of english conversation. *Behavior Research Methods, Instruments, & Computers*, 16(6):502–532.

John B Carroll and Margaret N White. 1973. Age-of-acquisition norms for 220 picturable nouns. *Journal of Verbal Learning and Verbal Behavior*, 12(5):563–576.

Harrison Chase. 2022. LangChain. Software. Released on 2022-10-17.

Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794.

Edgar Dale and Jeanne S Chall. 1948. A formula for predicting readability: Instructions. *Educational research bulletin*, pages 37–54.

Siobhan Devlin. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. *Linguistic databases*.

Liviu Dinu, Ana Uban, Alina Cristea, Anca Dinu, Ioan-Bogdan Iordache, Simona Georgescu, and Laurentiu Zoicas. 2023. RoBoCoP: A comprehensive ROmance BOrrowing COgnate package and benchmark

for multilingual cognate identification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7610–7629, Singapore. Association for Computational Linguistics.

Ken J Gilhooly and Robert H Logie. 1980. Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior research methods & instrumentation*, 12(4):395–427.

Sian Gooding. 2022. On the ethical considerations of text simplification. *arXiv preprint arXiv:2204.09565*.

Sian Gooding and Ekaterina Kochmar. 2018. Camb at cwi shared task 2018: Complex word identification with ensemble-based voting. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 184–194.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L'elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *ArXiv*, abs/2310.06825.

Henry Kučera, Winthrop Francis, William Freeman Twaddell, Mary Lois Marckworth, Laura M Bell, and John Bissell Carroll. 1967. Computational analysis of present-day american english. *(No Title)*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Ines Montani, Matthew Honnibal, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spaCy: v3.7.2: Fixes for APIs and requirements.

Sergiu Nisioi, Ella Rabinovich, Liviu P Dinu, and Shuly Wintner. 2016. A corpus of native, non-native and translated texts. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4197–4201.

Clyde E Noble. 1953. The meaning-familiarity relationship. *Psychological Review*, 60(2):89.

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

Allan Paivio, John C Yuille, and Stephen A Madigan. 1968. Concreteness, imagery, and meaningfulness values for 925 nouns. *Journal of experimental psychology*, 76(1p2):1.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

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*.

Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. 2019. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, 51:1258–1270.

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

Robyn Speer. 2022. rspeer/wordfreq: v3.0.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: an open multilingual graph of general knowledge. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, page 4444–4451. AAAI Press.

Otfried Spreen and Rudolph W Schulz. 1966. Parameters of abstraction, meaningfulness, and pronunciability for 329 nouns. *Journal of Verbal Learning and Verbal Behavior*, 5(5):459–468.

Sanja Stajner, Sergiu Nisioi, and Ioana Hulpuș. 2020. CoCo: A tool for automatically assessing conceptual complexity of texts. In *Proceedings of the*

Twelfth Language Resources and Evaluation Conference, pages 7179–7186, Marseille, France. European Language Resources Association.

Edward Lee Thorndike and Irving Lorge. 1944. The teacher's word book of 30,000 words.

Michael P Toglia and William F Battig. 1978. *Handbook of semantic word norms.* Lawrence Erlbaum.

Michael Wilson. 1988. Mrc psycholinguistic database: Machine-usable dictionary, version 2.00. *Behavior research methods, instruments, & computers*, 20(1):6–10.

Seid Muhie Yimam, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri. 2018. A report on the complex word identification shared task 2018. In *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

A Appendix

Language	Run ID	Pearson's R	Spearman's Rank	Mean Absolute Error	Mean Squared Error	R2
Catalan	1	0.2960	0.3029	0.1270	0.0246	-0.0342
Catalan	2	0.2744	0.2649	0.1236	0.0235	0.0110
Catalan	T	0.243333	0.200048	0.186026	0.050979	-1.146786
English	2	0.7904	0.7547	0.1225	0.0206	0.4393
Filipino	1	0.3620	0.4133	0.1729	0.0416	-0.9131
Filipino	2	0.4427	0.4476	0.1251	0.0234	-0.0763
Filipino	T	0.170322	0.200824	0.152792	0.039501	-0.817912
French	1	0.5335	0.5310	0.1898	0.0487	0.2136
French	2	0.4411	0.4188	0.1851	0.0504	0.1862
French	T	0.507726	0.502782	0.178938	0.046882	0.243141
German	1	0.5508	0.5726	0.1217	0.0252	0.0686
German	2	0.5577	0.5774	0.1369	0.0306	-0.1320
German	T	0.158362	0.18251	0.313923	0.129138	-3.779821
Italian	1	0.5341	0.5320	0.1705	0.0398	-0.4175
Italian	2	0.4790	0.4805	0.1426	0.0298	-0.0599
Italian	T	0.29937	0.309153	0.148348	0.03802	-0.353931
Japanese	1	0.2803	0.2648	0.2650	0.0894	-2.2358
Japanese	2	0.4851	0.5126	0.1440	0.0303	-0.0983
Japanese	T	0.038864	0.067513	0.181906	0.053068	-0.920658
Portuguese	1	0.7143	0.7102	0.1454	0.0276	-0.2612
Portuguese	2	0.6831	0.6923	0.1068	0.0166	0.2419
Portuguese	T	0.42688	0.446644	0.122814	0.026359	-0.206013
Sinhala	1	-0.0290	-0.0272	0.3920	0.1676	-9.3516
Sinhala	2	0.0437	0.0298	0.1239	0.0236	-0.4590
Sinhala	T	0.10023	0.065891	0.122526	0.028593	-0.76549
Spanish	1	0.5274	0.4793	0.1312	0.0265	0.2507
Spanish	2	0.5034	0.4588	0.1255	0.0272	0.2304
Spanish	T	0.326812	0.245494	0.20517	0.067601	-0.912674

Table 2: Lexical Complexity prediction of our models. The submissions marked with 1 are using a model trained only on the English language. The ones marked with 2 are trained on the entire multilingual data. And the ones marked with T are predictions only on translated data. It is clear from this table that translations significantly underperform predictions on original language even if the model was only trained on English data.