The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline

Matthew Shardlow[1], **Fernando Alva-Manchego**[2], **Riza Batista-Navarro**[3],
Stefan Bott[4], **Saul Calderon Ramirez**[5], **Rémi Cardon**[6], **Thomas François**[6],
Akio Hayakawa[4], **Andrea Horbach**[7,8], **Anna Hülsing**[7], **Yusuke Ide**[9],
Joseph Marvin Imperial[10,15], **Adam Nohejl**[9], **Kai North**[11], **Laura Occhipinti**[12],
Nelson Peréz Rojas[5], **Nishat Raihan**[11], **Tharindu Ranasinghe**[13],
Martin Solis Salazar[5], **Sanja Štajner**[14], **Marcos Zampieri**[11], **Horacio Saggion**[4]
[1]Manchester Metropolitan University [2]Cardiff University [3]University of Manchester
[4]Universitat Pompeu Fabra [5]Tecnológico de Costa Rica [6]UCLouvain
[7]University of Hildesheim [8]FernUniversität in Hagen
[9]Nara Institute of Science and Technology [10]National University Philippines
[11]George Mason University [12]University of Bologna [13]Aston University
[14]Karlsruhe [15]University of Bath
m.shardlow@mmu.ac.uk

Abstract

We report the findings of the 2024 Multilingual Lexical Simplification Pipeline shared task. We released a new dataset[1] comprising 5,927 instances of lexical complexity prediction and lexical simplification on common contexts across 10 languages, split into trial (300) and test (5,627). 10 teams participated across 2 tracks and 10 languages with 233 runs evaluated across all systems. Five teams participated in all languages for the lexical complexity prediction task and 4 teams participated in all languages for the lexical simplification task. Teams employed a range of strategies, making use of open and closed source large language models for lexical simplification, as well as feature-based approaches for lexical complexity prediction. The highest scoring team on the combined multilingual data was able to obtain a Pearson's correlation of 0.6241 and an ACC@1@Top1 of 0.3772, both demonstrating that there is still room for improvement on two difficult sub-tasks of the lexical simplification pipeline.

1 Introduction

The lexical simplification pipeline is a family of systems designed to automatically identify and replace complex vocabulary with simpler alternatives (North et al., 2023b). The lexical simplification pipeline provides a more targeted approach to simplification than automated text simplification (Al-Thanyyan and Azmi, 2021; Alva-Manchego et al., 2020; Saggion, 2017) which directly rewrites entire sentences. The two core operations included in the lexical simplification pipeline are (1) lexical complexity prediction (LCP) and (2) the replacement of complex words with simple synonyms.

LCP (Shardlow et al., 2020, 2022; North et al., 2023b,c), a form of Complex Word Identification (CWI) (Shardlow, 2013), involves assigning continuous values (0-1) to given tokens in context, representing the difficulty that an intended reader population may associate with that target word.

The second task, often referred to just as lexical simplification (LS) (Saggion et al., 2022) involves generating simple substitutions for target words in context. This task has been explored for single words and multi-word expressions, and is related to the identification of simple paraphrases (Maddela et al., 2021).

We previously identified two shortcomings of current work on the lexical simplification pipeline (Shardlow et al., 2024) as follows:

1. Current datasets only explore one pipeline operation, but no dataset exist with multiple operations on the same target words in context. This means that systems that are trained on one task are unsuitable for the other. Systems trained using multiple datasets may experience 'genre drift', where the text type across datasets differs.

2. The existing data is overwhelmingly in the English language. Whereas recent efforts exist to provide open source data in languages other than English, there is no guarantee that these datasets are created using the same protocols.

[1] https://github.com/MLSP2024/MLSP_Data/

Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications, pages 571–589
June 20, 2024 ©2024 Association for Computational Linguistics

We introduce the Multilingual Lexical Simplification Pipeline (MLSP) shared task, which provides a newly annotated dataset across 10 languages for LCP and LS. The annotations for both these tasks are provided on common targets in common contexts allowing further exploration of the interplay between the two tasks and evaluation of the full pipeline on common datasets. We release data in English, Spanish, French, Portuguese, Sinhala, Filipino, Japanese, Italian, German and Catalan. Of these languages, there were previously no available LCP resources for Portuguese, Sinhala, Filipino, Italian or Catalan and no LS resources available for Sinhala, Filipino, Italian, German or Catalan.

In the remainder of this Findings paper, we overview previous related shared tasks (Section 2); give a description of the task (Section 3); overview the preparation of our shared task dataset (Section 4); the participating systems (Section 5) and the results (Section 6). We conclude with a discussion of wider factors affecting our task (Section 7).

2 Related Tasks

LS 2012 at SemEval: The first shared task in LS was proposed for SemEval 2012. It addressed English LS (Specia et al., 2012) and offered the opportunity to evaluate systems able to rank substitution candidates in relation to their simplicity. The dataset used was taken from the Lexical Substitution task at SemEval 2007 (McCarthy and Navigli, 2007) which was enriched with simplicity rankings provided by second language learners with high proficiency levels in English. The task attracted five different institutions which provided nine systems in total.

CWI 2016 at SemEval: At SemEval 2016, the CWI task (Paetzold and Specia, 2016) requested participants predict which words in a given sentence would be considered complex by a non-native English speaker. A new dataset composed of 9,200 instances was created. The task attracted 21 teams which produced a total of 42 systems. A post-completion analysis (Zampieri et al., 2017) highlighted the difficulty of the shared-task. The authors claimed that a disproportionate train/test split with over 40 times more test data, together with low inter-annotator agreement, was to blame for poor system performances.

CWI 2018 at BEA: The BEA 2018 CWI shared task (Yimam et al., 2018) proposed to tackle CWI in English, German, and Spanish (training and test data were provided), together with a multilingual task with French as a target language without training data. Teams were asked to classify words as either complex or simple (binary) and/or provide a probability for the complexity of each word. The shared task attracted eleven teams.

ALexS 2020 at IberLEF: Additionally, the IberLef 2020 forum proposed a shared task on Spanish CWI(Ortiz-Zambranoa and Montejo-Ráezb, 2020). This workshop attracted seven teams, of which three submitted to the final task. The teams competed on the newly annotated VYTEDU-CW corpus which provided binary complexity judgments over educational texts.

LCP 2021 at SemEval: The SemEval 2021 shared task on LCP (Shardlow et al., 2021) also provided a new dataset for complexity detection for single words and multi-word expressions in English attracting 55 teams. Annotations were provided as continuous complexity judgements as opposed to binary complexity values. Teams made use of deep learning based approaches to predict lexical complexity values across the corpus.

SimpleText 2021 at CLEF: The SimpleText workshop (Ermakova et al., 2022) has been running at CLEF since 2021. This workshop aims to provide benchmarks and datasets for the improvement of the accessibility of scientific information. The workshop provides datasets that participants can compete on each year in the areas of: (1) passage selection for the creation of simplified extractive summaries; (2) identification of difficult concepts and (3) query-based simplified rewriting of scientific abstracts.

TSAR 2022 Shared Task on LS: The TSAR-2022 shared task (Saggion et al., 2022) provided annotations for LS in English, Spanish and Portuguese. Participants were required to predict up to 10 simple substitutions for a complex word in each language. Participants were free to contribute to one, two or all three languages. 14 Teams submitted 60 runs across the three languages. Successful systems made use of prompt engineering (Aumiller and Gertz, 2022; Vásquez-Rodríguez et al., 2022) with large language models, as well as incorporating feature-based approaches (Li et al., 2022).

3 Task Description

Our dataset consists of instances of marked words in context, where participants are required to develop systems that first identify the complexity level of the marked word and then provide suggestions for appropriate simplifications. This unites the two previous tasks of LCP and LS into a single task, executed on common data. We have provided test data in 10 languages (Catalan, English, Filipino, French, German, Japanese, Italian, Portuguese, Sinhala, Spanish) with our final dataset totalling 300 trial instances and 5627 test instances. Participants were free to choose which tasks and language tracks they participated in.

4 Data and Resources

We initially provided participants with labelled trial data only (30 instances across 10 contexts per language, designed to indicate the format of the task). We did not provide training data, but instead pointed participants to existing resources for LCP and Complex Word Identification arising from previous shared tasks. We have provided a simplified example of the task presented to participants below:

(1) That period of **intense** regulatory **scrutiny** is a routine part of the **purchasing** process.

Token	Complexity	Substitutions
intense	0.5	strong, forceful
scrutiny	0.8	examination, observation, inspection
purchasing	0.6	buying, acquiring, obtaining

In the table above, the first column shows the tokens that were selected by the organisers for annotation. The second column shows the complexity label assigned to each word, which is provided by the participant systems. The final column shows the substitutions for each word, also provided by the participant systems. Participants provided similar annotations across their chosen language tracks, which were compared to the gold evaluation data.

4.1 Dataset Collection

Each section of the dataset was provided by a team of organisers consisting of at least one native speaker for the given language. We collected annotations from a minimum of 10 annotators per instance. Annotators were required to annotate lexical complexity for each identified token on a scale of 1-5. Annotators were also asked to provide up to 3 possible simplifications for each instance. More information on the trial dataset creation is given at Shardlow et al. (2024) and the MultiLS protocol we used at North et al. (2024).

Depending on the availability of appropriate texts requiring simplification and target populations to provide annotations, the organisers responsible for each language made autonomous decisions on the most appropriate method to gather language specific LCP and LS annotation. Information on language-specific concerns are described below.

4.1.1 Catalan

The Catalan dataset is comprised of sentences selected from the news section on education of the TeCla corpus[2] (Armengol-Estapé et al., 2021) of Catalan news texts. Target words were annotated by proficient Catalan speakers, in part recruited from persons of the social environment of the data collectors (10 participants) and in part from workers recruited via Prolific[3] crowdsourcing platform (74 participants). Although only 22% of participants were native speakers, all annotators had a high level of Catalan proficiency. The annotation process in Prolific was monitored in order to detect workers who were not following the annotation guidelines, for example, annotators who always returned the same target word as the substitute, or provided synonyms in Spanish. Non-compliant annotators were given the chance to repeat the annotation and, if they failed again, excluded.

4.1.2 English

The English dataset takes WikiBooks as a source text. English targets were identified using frequency profiling for 200 contexts. 2 additional words were identified per context ensuring that all selected words in the set were unique. The lexical complexity annotations and LS annotations were completed jointly by 21 annotators (10 native speakers, 11 non-native), all of whom were registered as students at the Manchester Metropolitan University. Each annotator saw 300 instances,

[2] https://huggingface.co/datasets/projecte-aina/tecla

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

with a total of 10-11 annotations across 600 instances.

4.1.3 Filipino

The Filipino data is composed of sentences retrieved from early-grade level books accredited by the Department of Education in the Philippines and sampled from a larger collection of Filipino resource works (Imperial and Kochmar, 2023a,b; Imperial and Ong, 2021). The genre of the sentences varies and includes samples from fiction, biographies, and instructional reference books. The annotations for the dataset were provided by 10 university staff who were native speakers of Filipino and were asked to consider the reading level of a second-grade elementary student while annotating each sentence. Instances of borrowed English words in the data were transliterated to Filipino to preserve the uniformity of phonetics (e.g. *basketball* is converted to *basketbol*).

4.1.4 French

The French dataset was compiled from a collection of texts that are used in French as a Foreign Language (FFL) classes in France, which is still under construction. The corpus contains texts targeting learners with CEFR levels going from A1 to B2. Various genres are represented, including encylopedia articles, news articles, social media, commercial and professional communication, fiction and non fiction books, or legal and political texts. Sentences that appear in the shared task dataset contain at least one word marked as B2 in the FLELex graded lexicon (François et al., 2014). Two other words were chosen manually for each sentence. The complexity annotation was performed by 10 FFL students in Belgium, attending A2 and B1 classes (5 from each level). The substitutions were provided by 10 native French speakers – Belgian master's students attending literature or social science classes.

4.1.5 German

The German data consists of Wikipedia (50%) and literary texts (50%). The data was chosen based on topics and texts mandatory for German students in their last year of secondary education in history lessons (e.g. Berlin Wall) and German lessons (e.g. *Der goldene Topf* by E. T. A. Hoffmann). Annotations were provided by German native speakers employed at universities, who were asked to take the perspective of the target group: students in their last year before graduation with a first language other than German. Simplifications that required context changes were only considered acceptable if the gender or number of a simplification required agreement with a preceding determiner, pronoun, or adjective. Example for the simplification of *Tempo*, where the determiner (underlined) changes: *mit <u>dem</u> Tempo* ("at the pace") is substituted by *mit <u>der</u> Schnelligkeit* ("at the speed").

4.1.6 Japanese

The Japanese data targets non-native Japanese speakers, whose native language is neither Chinese nor Korean, as Chinese or Korean L1 background constitutes a considerable advantage in comprehension of Japanese due to partially shared vocabulary (Koda, 1989), and therefore affects perceived lexical complexity (Ide et al., 2023).

The Japanese sentences were extracted from Wikipedia (50%), web pages with practical information, e.g. from local authorities (21%), literary fiction (19.5%), news texts (5.5%), and texts about Japanese culture and history (4%). The target words were selected to represent a wide range of word frequencies and character (*kanji*) frequencies, as well as diverse parts of speech (nouns, verbs, adjectives, adverbs, particles, and auxiliaries). Additionally, the targets include specific types of words known to be difficult for learners (compound verbs, compound particles, and onomatopoeia).

We recruited 10 non-native annotators for LCP annotation, and 10 native annotators for LS annotation. The LCP annotators were holders of Japanese Language Proficiency Test (JLPT)[4] levels 1 (N1) or 2 (N2) and their native language was neither Chinese nor Korean. The LS annotators had at least one year of experience teaching Japanese as a second language.

4.1.7 Italian

The Italian dataset comprises texts related to Italian literature, a subject taught across all school levels and grades. Specifically, 50% of the sentences have been extracted from Wikibooks, while the remaining 50% consist of sentences from 20th-century Italian authors sourced from Wikisource. We selected modern authors to avoid words considered too arcane for contemporary speakers. The task was designed for a 'general Italian speaker',

[4] `https://www.jlpt.jp`

and therefore, annotations were provided by native speakers with varying levels of education and literacy. A total of 215 individuals participated in the annotation process, ensuring a minimum of 10 annotations per sentence. For the substitution task, it was specified that annotators could replace target terms with words of different genders, thus not limiting the choice of possible substitutes. Additionally, annotators were instructed to treat pronominal verbs as single entities, which could also be replaced with other verbs, for example, replacing "mobilitarsi" with "agire".

4.1.8 Portuguese

The Portuguese dataset contains sentences taken from Bible extracts (47%), news articles (35%), and biomedical papers (17%). Bible instances were obtained from the Bíblia Sagrada (North et al., 2024). News instances were taken from the PorSimplesSent dataset (Leal et al., 2018) and from the CC-News (Common Crawl-News) corpus (North et al., 2022, 2023a). Biomedical instances were extracted from abstracts of biomedical literature provided by WMT-2019 (Bawden et al., 2019). Only one target word per sentence was annotated, rather the three target words per context. 21 Portuguese annotators were crowdsourced using Amazon Mechanical Turk (MTurk) and were selected from Brazil.

4.1.9 Sinhala

The Sinhala data consists of sentences extracted from a recent Sinhala news corpus (Hettiarachchi et al., 2024) and Sinhala translation of Tripitaka; the standard collection of scriptures in the Theravada Buddhist tradition written originally in Pali. Approximately 30% of the sentences were extracted from Tripitaka, and the rest of the sentences were from the news corpus. We recruited ten university students who were studying for a BA in Sinhala and were also native speakers of Sinhala for the annotation process.

4.1.10 Spanish

The Spanish dataset derives from a corpus of over 5K sentences for sentence simplification currently under development. The sentences were extracted from four online university educational books in the area of finance and were simplified following a set of simplification guidelines borrowed from the Simplext project (Saggion et al., 2015). The annotation was undertaken by 60 students who are native Spanish speakers and by 10 persons from social contacts of the data collectors, half of whom were native speakers. Out of all annotators, 8% were non-native speakers with high Spanish language proficiency.

4.2 Evaluation Metrics

For the evaluation of the LCP task we use **Pearson's correlation**, **Spearman's rank**, and the coefficient of determination (R^2) in line with the 2021 shared task on LCP.

For the evaluation of the LS task (see (Štajner et al., 2022)) we use Accuracy@k@top1 and MAP@K defined as follows the 2022 shared task on LS: **Accuracy@k@top1** is the percentage of instances where at least one of the k top-ranked substitutes matches the most frequently suggested synonym in the gold data. **MAP@k** uses a ranked list of generated substitutes, which can either be matched (relevant) or not matched (irrelevant) against the set of the gold-standard substitutes.

As some of the instances are not simplifiable or have less than k gold standard simplifications, the maximum achievable results in Accuracy@k@top1 and MAP@k are less than 1. Appendix A shows the number of unsimplifiable instances as well as maximum achievable values in all metrics.

4.3 Baselines

For LCP, we provide a baseline modelled as a linear regression on log-frequency. The frequency baseline is trained using log-frequency (minimum value if the target consists of multiple tokens) on the trial set for each language. We use frequencies provided by the `wordfreq` package[5] when possible. Additionally, since the package uses an incompatible tokenization for Japanese and does not provide any data for Sinhala, we use TUBELEX-JA[6] for Japanese, and a word frequency list for Sinhala[7] by Fernando and Dias (2021).

For LS, we provide a baseline based on zero-shot prompting a large language model. We employ the chat-finetuned Llama 2 70B model[8] (Touvron et al., 2023) in 4-bit quantisation. We use the following zero-shot prompt template and tem-

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

[6] https://github.com/adno/tubelex

[7] https://github.com/nlpcuom/Word-Frequency-List-for-Sinhala

[8] https://huggingface.co/meta-llama/Llama-2-70b-chat-hf

perature 0.3 to generate a maximum of 256 new tokens.

```
Context: {context}↵
Question: Given the above context, list ten
alternative {language} words for "{word}"
that are easier to understand. List only the
words without translations, transcriptions
or explanations.↵
Answer:
```

Only the ↵ symbols represent line breaks. To construct the prompt, the placeholders in curly braces are replaced by the context, the language of the instance, and the target word to be simplified. For English, the placeholder **{language}** and the subsequent space is omitted. The prompt is identical to a zero-shot prompt employed for LS using a ChatGPT model by Aumiller and Gertz (2022), except for the the underlined sentence (`List only...`), which we have added to reduce unnecessary translations to English, transcriptions to Latin alphabet, or explanations. Such extra input was generated frequently when we applied the original prompt to trial data. The addition of the sentence results in both faster inference and higher accuracy.

Our postprocessing also builds on the work by Aumiller and Gertz (2022). Based on an examination of outputs using the trial data, we made minor changes reflecting a broader array of languages and scripts as well as a different model. For instance, we allow words to be separated by ideographic commas (、) commonly used in Japanese, or lists enumerated using letters (e.g. `a)`, `b)`, ...), which occurred in Llama 2 output.

5 Participating Systems

ANU (Seneviratne and Suominen, 2024) The ANU team relied on a prompting strategy with GPT-3.5 (i.e. GPT-3.5-turbo-instruct) for both tasks using zero, one, and few-shot strategies. The zero-shot strategy included the context and target word while the non-zero strategies relied on instructing the model with one or three random samples from the trial data according to the prompting template. For LS, a combination of filtering and substitution was applied. Overall, the authors indicate under-performance for the LCP task while strong performance for English in LS.

Archaeology (Cristea and Nisioi, 2024) The Archaeology team participated in both LCP and LS. For both tasks, they make use of machine translation software to convert all texts to English. The LCP values are generated using a feature-based approach with word-level, syntactic-level and semantic-level features. An XGBoost regressor is trained on the Semeval 2021 English test dataset and used to predict lexical complexity values for all languages. The simplifications for the LS task were generated in English using the translated data by prompting a large language model (OpenHermes 2.5) to produce JSON data containing the candidate replacements and back-translated to the target language.

CocoNut The CocoNut team submitted LAE-LS, which introduced a novel method for LS, trained without the use of parallel corpora or external linguistic resources. LAE-LS employed an Adversarial Editing System with guidance from a confusion loss and an invariance loss to predict lexical edits in the original sentences. An LLM-enhanced loss was tailored to distill high-quality knowledge from LLMs into the Edit Predictor. Complex words within sentences were masked and a Difficulty-aware Filling module crafted to replace masked positions with simpler words. For LCP, the team used the probability of a word being masked by the Edit Predictor as the complexity value of the word in context. For LS, complex words were masked and the Difficulty-aware Filling module was used to predict substitute words.

GMU (Goswami et al., 2024) The GMU team participated in both subtasks. For LCP they employed a weighted ensemble of mBERT, XLM-R and language specific BERT models. All trial data was used for cross-lingual training and evaluation. For the combined track, an ensemble of language specifc models was used. For LS GPT4-turbo zero shot prompting was used, as well as mBERT, XLM-R and language specific BERT models. Cosine similarity between the target token and the substitutions generated by all the models were generated. Sentence transformer LaBSE is used to find the embeddings of the substitutions. The top 10 substitutions with the highest cosine similarity are selected for the output.

ISEP Presidency University (Dutilleul et al., 2024) The ISEP team also relied on a GPT-3 language model (i.e. GPT-3.5-turbo-instruct) and prompt engineering to solve the LS task. More concretely, several prompt generation strategies are used: a context-free strategy asks for ten simpler substitutes for the target word without specifying the context, a zero-shot strategy instead provides the context and the target word, a one-shot strategy is similar to zero-shot but provides one example of how to answer, and finally a few-shots strategy provide several examples to the model before testing. Responses from all strategies are aggregated and answers ranked to produce the final list of substitutes. The team reports satisfactory aggregated performance in most languages they applied this method to.

ITEC (Tack, 2024) The ITEC team participated only in the LCP subtask for French. They relied on two pre-trained models, previously developed for personalised LCP. Due to the characteristics of the shared task data, the personalisation component was removed. The team employed two models of similar architectures: a mix of character and FastText embeddings that are fed to either a BiLSTM or a feed-forward network, in order to consider contextual information or not, respectively, for predictions.

RETUYT-INCO (Sastre et al., 2024) The RETUYT-INCO team make use of a range of methods for their submitted runs, including word embeddings and frequency baselines for Spanish, English and Portuguese (LS). Feed forward networks with BERT-based embeddings for Spanish and English (LCP). Fine-tuning Mistral-7B for English (LCP) and with synthetic data and self-consistency for English, Spanish, Catalan and Portuguese (LCP and LS) and finally, prompting strategies using models available in the Groq API for Spanish (LS).

SCaLAR The SCaLAR team participated across both tasks, employing Mistral-7B for LS in a few shot learning setup with post-processing. Similarity scores were obtained through Word2Vec to identify the the top 10 similar words for each complex word. For LCP, the team used a weighted sum of 2 approaches: (1) MPNet Hidden State to Image Regression with EfficientNet: Transforms MPNet hidden states into image format and employs EfficientNet for image regression, bridging text data to convolutional neural networks. (2) XGBoost Regressor with TF-IDF and Zipf Frequency Features: Utilizes XGBoost regressor with features derived from TF-IDF and Zipf frequency.

SDJZUandUU The Complex Word Identification (CWI) model of team SDJZUandUU comprises of three integral modules: the Feature Collection Module, Feature Fusion Module, and Regression Model. The Feature Collection Module is designed to gather diverse feature sets including 16 commonly utilized handcrafted features, GloVe embeddings, and dynamic dependency embeddings. This module incorporates Gaussian vectorization techniques to vectorize the handcrafted features effectively. Subsequently, the Feature Fusion Module combines the aforementioned feature types into a vector representation, which is then passed to the Regression Model. The Regression Model is composed of three layers: two Support Vector Regression (SVR) polynomial layers for feature refinement within the feature vectors, and one feedforward layer aimed at predicting the final complexity value.

TMU-HIT (Enomoto et al., 2024) TMU-HIT employed a GPT-4 based approach in both tasks. In System 1, the team used GPT-4 to generate 10 alternative words for the target word in a zero-shot setting. In the case of Japanese, rather than solely generating alternative words, the team directed GPT-4 to generate sentences wherein the target words were substituted with each alternative word. This approach was necessary to ensure that the "katsuyou" (inflection) appropriately suited the context in Japanese. substitues were reranked through (a) prompting and (b) fine-tuned XGLM. For LCP, the team use a chain-of-thought based prompting method employing GPT-4 to generate an instruction in English, and subsequently assigning complexity scores to target words across all languages based on the English instruction.

6 Results

The full results for LCP and LS are displayed in Appendix B and Appendix C respectively. Each team was permitted to submit up to 3 runs per language track, with teams permitted to submit to both the combined track and the individual language tracks. The ID field indicates the run ID of the participants systems. Where teams submitted a separate system to the combined track, the results for each individual language were also separately processed and included in the results tables for the individual language tracks, these are indicated by a run ID preceding with 'A'. All team outputs can be found via GitHub[9].

Whilst all systems provide interesting insights into the nature of the lexical simplification pipeline, we have chosen to highlight a small number of systems below. The full descriptions of each system are available in the proceedings.

The results demonstrate that the GPT-4 based approach of the *TMU-HIT* team performed well across both tasks and all language tracks. This system consistently outperforms the baseline and is consistently the first or second highest ranked system. Prompt-based strategies have previously proved to be effective for LS, but not for the LCP task.

The *Archaeology* submission based on machine translation performs well for LCP, ranking as the second team in the combined track. This system uses a feature-based regression, demonstrating that this is still a competitive approach. The system does not perform as well on the LS task, and this is likely due to the challenge of correctly identifying targets after back-translation.

The *RETUYT-INCO* submission attains second place in LCP for Catalan, Filipino, Sinhala and Spanish. This submission made use of bespoke resources, including synthetic data for low-resource languages. The competitive performance of this submission on these tracks indicates that this approach may be appropriate for future low-resource languages that cannot be handled through a conventional prompt-based approach.

The *GMU* team attained first place for the EN-LCP task, setting a new hard to beat baseline for this dataset. Their approach also attains strong LS results for all languages, consistently attaining the 2nd or third ranked team in each language and ranking as the second team on the combined track.

Finally, the *ISEP* team chose to only compete in a reduced set of languages for the LS task. This focus allowed them to submit a competitive system for Catalan (1st place), Portuguese (1st place), French (2nd team) as well as English (4th Team) and German (4th team), outperforming the baseline in all cases.

We provided a simple baseline for LCP based on word-frequency and for LS based on a simple LLM-prompting strategy following prior work. The baseline is included in all results tables as 'Baseline', except for the combined results table, where we have not included a baseline result. We have sorted each results table, including the baselines, according to the Pearson's Correlation for LCP and Acc@1@Top1 for LS and we refer to systems 'above the baseline' in this context.

For LCP our baseline system was generally competitive, expect for Sinhala. The system was based on word frequencies and the frequencies we had available for Sinhala were not suitable for the task. Our baseline received a negative correlation to the gold labels for Sinhala (as did several participant systems). For other systems, our baseline performs strongly (ranking between the 2nd and 4th system for all languages except for English and Sinhala) confirming our hypothesis that word frequency would be a strong indicator of lexical difficulty. For English, the baseline system attains a strong correlation of 0.7480, but is outperformed by 9 other systems. The English LCP track was more subscribed than any other.

For LS, our baseline system received mixed results, generally attaining a mid-table ranking. Our approach was to reuse the prompt from the previous LS shared-task winner, which is a similar strategy to many of the submitted systems which also further improved on this same approach. Our system performs particularly poorly for Filipino and for Sinhala, and this is likely the result of the base language model lacking training data for these languages.

Although we have ranked our systems according to Pearson's correlation for LCP, it is also interesting to observe the R^2 metric of each system as compared to the baseline. The R^2 metric describes the proportion of variance captured by the system's results, i.e., how well do the LCP values returned by the system describe the LCP values in the gold labels. A negative R^2 indicates that the returned values are a poor fit to the gold values, whereas a positive R^2 indicates a good fit.

Our baseline attains a positive R^2 for all systems, except for Catalan and Sinhala. Notably, for English the baseline system attains the highest R^2 of any system. This is also true for Filipino (all other systems have negative R^2), German and Portuguese. This indicates that although systems are able to provide correlative LCP judgements, additional factors are still required to fully represent the underlying data distributions.

[9] https://github.com/MLSP2024/MLSP_Participants/

7 Discussion

We provided 10 languages for the evaluation of LCP and LS. Unsurprisingly, the most subscribed language track was English, with the most prior work and existing resources in NLP concentrated on English. We hope to address the imbalance in LCP/LS research by providing equal amounts of data for all languages that we have included. The English submissions attained the highest scores overall for LCP and LS, demonstrating that the English task is better resourced. Further developments in multilingual NLP and in bespoke resources for individual target languages will help to improve the performance of other systems on the tasks in our dataset.

Our dataset covers widespread global languages such as English, Spanish and French. There are a disproportionate number of languages in our dataset that are influenced from the romance family (Spanish, Catalan, French, Italian, Portuguese). We hope to extend the dataset in further iterations to include other widespread languages such as Mandarin Chinese, Hindi, Modern Standard Arabic and Bengali.

In addition to focussing future development on widespread languages, our work has also shown that LCP and LS can be effectively applied to low-resource languages. Future work to develop LCP/LS resources using the MultiLS framework (North et al., 2024) which we have followed will be incorporated into our dataset to enable the LS task for wider digital communities.

Whereas previous approaches to LCP have focussed on regression studies, e.g., using a language model with a regression head, it is interesting to note that many of the systems were able to use a prompting strategy to get good results for the LCP task. The TMU-HIT system relies on prompting to generate N judgements, effectively forcing the LM to undertake the annotation task. This proves effective across many languages. The use of language models to replicate the annotators is an interesting area of future exploration which may have significant repercussions across other similar lexical semantics tasks such as hate speech or sentiment analysis. Nonetheless, feature based systems such as the frequency baseline and the feature-based regression of the Archaeology team still performed competitively, demonstrating that this can be an effective method for LCP, especially when large language models are not available for the target language.

The principal strategy for the LS task employed by our participants was through prompt engineering. It is worth noting that several of the top-ranked submissions on this task used GPT4/GPT3.5, both of which are closed-source proprietary models. Whilst differing prompt engineering strategies were employed throughout the task, it is very difficult to separate the differences in performances that can be attributed to (a) the prompting strategies used and (b) the language models that they have been applied to. A possible future strategy to prevent model-variance may be to provide all teams access to some common model and enforce its use in a task.

8 Conclusion

We present the findings of the 2024 Multilingual Lexical Simplification Pipeline shared task hosted at the 19th Workshop on Innovative Use of NLP for Building Educational Applications. We provided the first multilingual dataset for LCP and LS on common targets, spanning ten languages and nearly 6,000 instances. Ten teams participated in our task employing a range of LLM-based strategies at the forefront of modern NLP. Seven teams submitted system description papers. Our shared task has progressed the forefront of lexical simplification research and the organisers look forward to seeing future multilingual lexical simplification research born of these efforts. All datasets, baselines and participant submissions are available through the MLSP2024 GitHub Organisation[10].

Acknowledgments

Joseph Imperial is supported by the National University Philippines (Project ID: 2023I-1T-05-MLA-CCIT-Computer Science) and the UKRI Centre for Doctoral Training in Accountable, Responsible and Transparent AI [EP/S023437/1] of the University of Bath.

Andrea Horbach's work is part of the research conducted at CATALPA – Center for Advanced Technology-Assisted Learning and Predictive Analytics of the FernUniversität in Hagen, Germany. Anna Hülsing is supported by the German Federal Ministry of Education and Research (grant no. FKZ 01JA23S03C).

[10] https://github.com/MLSP2024

The Japanese section of the dataset was created as a part a joint research project of Nara Institute of Science and Technology and Nikkei, Inc.

Horacio Saggion, Stefan Bott and Akio Hayakawa acknowledge funding from the European Uniont's Horizon Europe research and innovation program under the Grant Agreement No. 101132431 (iDEM Project) – views and opinions expressed are however those of the author(s) only and do necessarily reflect those of the European Union. Horacio Saggion, Stefan Bott and Akio Hayakawa also thank the support from Departament de Recerca i Universitats de la Generalitat de Catalunya (ajuts SGR-Cat 2021).

The English section of the dataset was annotated thanks to the Manchester Metropolitan University RISE programme. The organisers would like to extend their thanks to the organisers of the RISE programme and to the student annotators who participated in the annotation process for their timely completion and attendance at annotation meetings.

The organisers would like to thank all the annotators that participated across the ten languages we have provided, both paid and unpaid. This work would not have been possible without the kind support of friends and professional connections within the organiser's networks.

We would also like to thank the ten teams, consisting of 37 colleagues at various levels of their academic journeys, for their participation in the shared task. Particular thanks go to those that provided reviews for other participants shared task submissions.

Finally, we would like to thank the organisers of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024), who kindly hosted our shared task. Particular thanks go to Victoria Yaneva for serving as liaison to the BEA organisers and also to Andrea Horbach for her help in facilitating the Softconf track.

References

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

Fernando Alva-Manchego, Carolina Scarton, and Lucia Specia. 2020. Data-driven sentence simplification: Survey and benchmark. *Computational Linguistics*, 46(1):135–187.

Jordi Armengol-Estapé, Casimiro Pio Carrino, Carlos Rodriguez-Penagos, Ona de Gibert Bonet, Carme Armentano-Oller, Aitor Gonzalez-Agirre, Maite Melero, and Marta Villegas. 2021. Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4933–4946, Online. Association for Computational Linguistics.

Dennis Aumiller and Michael Gertz. 2022. UniHD at TSAR-2022 shared task: Is compute all we need for lexical simplification? In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 251–258, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

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 the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 29–53, Florence, Italy. Association for Computational Linguistics.

Petru Theodor Cristea and Sergiu Nisioi. 2024. Archaeology at MLSP 2024: Machine translation for lexical complexity prediction and lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Benjamin Dutilleul, Mathis Debaillon, and Sandeep Mathias. 2024. ISEP presidency university at MLSP 2024: Using GPT-3.5 to generate substitutes for lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Liana Ermakova, Patrice Bellot, Jaap Kamps, Diana Nurbakova, Irina Ovchinnikova, Eric SanJuan, Elise Mathurin, Sílvia Araújo, Radia Hannachi, Stéphane Huet, et al. 2022. Automatic simplification of scientific texts: Simpletext lab at clef-2022. In *European Conference on Information Retrieval*, pages 364–373. Springer.

Aloka Fernando and Gihan Dias. 2021. Building a linguistic resource : A word frequency list for Sinhala.

In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 606–610, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Thomas François, Núria Gala, Patrick Watrin, and Cédrick Fairon. 2014. Flelex: a graded lexical resource for french foreign learners. In *International conference on Language Resources and Evaluation (LREC 2014)*.

Dhiman Goswami, Kai North, and Marcos Zampieri. 2024. GMU at MLSP 2024: Multilingual lexical simplification. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Hansi Hettiarachchi, Damith Premasiri, Lasitha Uyangodage, and Tharindu Ranasinghe. 2024. NSINA: A News Corpus for Sinhala. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

Yusuke Ide, Masato Mita, Adam Nohejl, Hiroki Ouchi, and Taro Watanabe. 2023. Japanese lexical complexity for non-native readers: A new dataset. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 477–487, Toronto, Canada. Association for Computational Linguistics.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023a. Automatic readability assessment for closely related languages. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5371–5386, Toronto, Canada. Association for Computational Linguistics.

Joseph Marvin Imperial and Ekaterina Kochmar. 2023b. BasahaCorpus: An expanded linguistic resource for readability assessment in Central Philippine languages. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6302–6309, Singapore. Association for Computational Linguistics.

Joseph Marvin Imperial and Ethel Ong. 2021. Under the microscope: Interpreting readability assessment models for Filipino. In *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 1–10, Shanghai, China. Association for Computational Lingustics.

Keiko Koda. 1989. The effects of transferred vocabulary knowledge on the development of L2 reading proficiency. *Foreign Lang. Ann.*, 22(6):529–540.

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

Xiaofei Li, Daniel Wiechmann, Yu Qiao, and Elma Kerz. 2022. MANTIS at TSAR-2022 shared task: Improved unsupervised lexical simplification with pretrained encoders. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 243–250, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3536–3553, Online. Association for Computational Linguistics.

Diana McCarthy and Roberto Navigli. 2007. SemEval-2007 task 10: English lexical substitution task. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 48–53, Prague, Czech Republic. Association for Computational Linguistics.

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 the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 404–413, Toronto, Canada. Association for Computational Linguistics.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2023b. Deep Learning Approaches to Lexical Simplification: A Survey. *Preprint*, arXiv:2305.12000.

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

Kai North, Marcos Zampieri, and Tharindu Ranasinghe. 2022. ALEXSIS-PT: A new resource for Portuguese lexical simplification. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6057–6062, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

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

Jenny A Ortiz-Zambranoa and Arturo Montejo-Ráezb. 2020. Overview of alexs 2020: First workshop on lexical analysis at sepln. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020)*, volume 2664, pages 1–6.

Gustavo Paetzold and Lucia Specia. 2016. SemEval 2016 task 11: Complex word identification. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 560–569, San Diego, California. Association for Computational Linguistics.

Horacio Saggion. 2017. Automatic Text Simplification. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Horacio Saggion, Sanja Štajner, Stefan Bott, Simon Mille, Luz Rello, and Biljana Drndarevic. 2015. Making it Simplext: Implementation and Evaluation of a Text Simplification System for Spanish. *TACCESS*, 6(4):14.

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

Ignacio Sastre, Leandro Alfonso, Facundo Fleitas, Federico Gil, Andrés Lucas, Tomás Spoturno, Santiago Góngora, Aiala Rosá, and Luis Chiruzzo. 2024. RETUYT-INCO at MLSP 2024: Experiments on language simplification using embeddings, classifiers and large language models. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Sandaru Seneviratne and Hanna Suominen. 2024. ANU at MLSP 2024: Prompt-based lexical simplification for english and sinhala. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Matthew Shardlow. 2013. A comparison of techniques to automatically identify complex words. In *51st Annual Meeting of the Association for Computational Linguistics Proceedings of the Student Research Workshop*, pages 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

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

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

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

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.

Lucia Specia, Sujay Kumar Jauhar, and Rada Mihalcea. 2012. SemEval-2012 task 1: English lexical simplification. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 347–355, Montréal, Canada. Association for Computational Linguistics.

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*, 5:991242.

Anaïs Tack. 2024. ITEC at MLSP 2024: Transferring predictions of lexical difficulty from non-native readers. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *ArXiv preprint*, arXiv:2307.09288 [cs].

Laura Vásquez-Rodríguez, Nhung Nguyen, Matthew Shardlow, and Sophia Ananiadou. 2022. UoM&MMU at TSAR-2022 shared task: Prompt learning for lexical simplification. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pages 218–224, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

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.

Marcos Zampieri, Shervin Malmasi, Gustavo Paetzold, and Lucia Specia. 2017. Complex Word Identification: Challenges in Data Annotation and System Performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, Taipei, Taiwan. Asian Federation of Natural Language Processing.

A Dataset Statistics and Maximum Achievable Results

Language	# Test Instances	# Unsimplifiable	Max. MAP@1, Accuracy@k@Top1	Max. MAP@3	Max. MAP@5
All	5627	133	0.9763	0.9081	0.7963
Catalan	445	1	0.9977	0.9910	0.9793
English	570	0	1.0000	0.9491	0.8115
Filipino	570	130	0.7719	0.5222	0.3466
French	570	0	1.0000	0.9953	0.9673
German	570	0	1.0000	0.9309	0.7908
Italian	570	0	1.0000	0.9859	0.9228
Japanese	570	0	1.0000	0.9988	0.9957
Portuguese	568	1	0.9982	0.9241	0.7220
Sinhala	600	0	1.0000	0.8072	0.4873
Spanish	593	1	0.9983	0.9966	0.9885

B Lexical Complexity Prediction Results

Team Name	Language	ID	Pearson's	Spearman's	R^2
TMU-HIT	All	2	0.6241	0.6215	0.2456
TMU-HIT	All	1	0.5609	0.5697	-0.3111
Archaeology	All	2	0.5316	0.5415	0.2560
RETUYT-INCO	All	1	0.4858	0.4892	-0.6746
GMU	All	1	0.3494	0.3642	0.1094
SCaLAR	All	1	0.0979	-0.0104	-0.0301

Team Name	Language	ID	Pearson's	Spearman's	R^2
TMU-HIT	Catalan	A2	0.6158	0.5989	-0.1610
TMU-HIT	Catalan	A1	0.5279	0.5327	-0.9634
RETUYT-INCO	Catalan	1	0.3948	0.3862	-1.3972
RETUYT-INCO	Catalan	A1	0.3608	0.3564	-1.5394
Baseline	Catalan	1	0.3011	0.3106	-0.3698
Archaeology	Catalan	1	0.2960	0.3029	-0.0342
Archaeology	Catalan	2	0.2744	0.2649	0.0110
GMU	Catalan	1	0.1549	0.1574	-0.3378
GMU	Catalan	A1	0.1137	0.1081	-0.1453
SCaLAR	Catalan	A1	0.0424	0.0065	-0.2236

Team Name	Language	ID	Pearson's	Spearman's	R^2
GMU	English	1	0.8497	0.7984	0.5247
TMU-HIT	English	2	0.8198	0.7552	0.5147
SDJZUandUU	English	3	0.8123	0.7754	0.5245
SDJZUandUU	English	1	0.8111	0.7414	0.3731
RETUYT-INCO	English	1	0.8061	0.7596	0.3154
TMU-HIT	English	1	0.8036	0.7017	0.3161
Archaeology	English	2	0.7904	0.7547	0.4393
SDJZUandUU	English	2	0.7820	0.7182	0.3529
RETUYT-INCO	English	3	0.7599	0.7406	-0.1796
Baseline	English	1	0.7480	0.7451	0.5475
RETUYT-INCO	English	2	0.5502	0.4923	0.1062
ANU	English	1	0.3358	0.3591	-3.0241
GMU	English	A1	0.3118	0.3183	0.0585
CocoNut	English	1	0.1972	0.2160	-5.1596
ANU	English	3	0.1915	0.2402	-0.5842
ANU	English	2	0.1789	0.2285	-0.0917
SCaLAR	English	A1	0.0126	0.0139	-0.2984

Team Name	Language	ID	Pearson's	Spearman's	R^2
TMU-HIT	Filipino	A1	0.5692	0.5816	-0.3536
TMU-HIT	Filipino	A2	0.5013	0.5244	-2.4778
RETUYT-INCO	Filipino	A1	0.4640	0.4540	-1.4847
Archaeology	Filipino	2	0.4427	0.4476	-0.0763
Baseline	Filipino	1	0.3892	0.4178	0.0036
Archaeology	Filipino	1	0.3620	0.4133	-0.9131
GMU	Filipino	A1	0.2823	0.2767	-0.0457
GMU	Filipino	1	0.1942	0.1908	-0.0824
SCaLAR	Filipino	A1	-0.0700	-0.0792	-0.2649

Team Name	Language	ID	Pearson's	Spearman's	R^2
TMU-HIT	French	A1	0.6253	0.6302	0.2704
Archaeology	French	1	0.5335	0.5310	0.2136
TMU-HIT	French	A2	0.5278	0.5343	0.2391
Baseline	French	1	0.5166	0.5221	0.1458
RETUYT-INCO	French	A1	0.4868	0.4651	0.0279
Archaeology	French	2	0.4411	0.4188	0.1862
ITEC	French	2	0.3607	0.4972	-4.4459
ITEC	French	1	0.3253	0.3533	-3.3488
GMU	French	1	0.3193	0.3207	0.0484
GMU	French	A1	0.1557	0.1756	0.0039
SCaLAR	French	A1	0.1035	0.0674	0.0061

Team Name	Language	ID	Pearson's	Spearman's	R^2
TMU-HIT	German	A2	0.7177	0.7365	-0.5585
TMU-HIT	German	A1	0.6582	0.6813	-0.7654
Baseline	German	1	0.5912	0.6096	0.0727
Archaeology	German	2	0.5577	0.5774	-0.1320
Archaeology	German	1	0.5508	0.5726	0.0686
RETUYT-INCO	German	A1	0.3909	0.3981	-0.3463
GMU	German	A1	0.1402	0.1473	-0.5279
SCaLAR	German	A1	0.0310	0.0177	-1.2467
GMU	German	1	0.0123	0.0095	-1.1301

Team Name	Language	ID	Pearson's	Spearman's	R^2
TMU-HIT	Italian	A2	0.6011	0.6220	0.2425
TMU-HIT	Italian	A1	0.5391	0.5557	-1.7874
Archaeology	Italian	1	0.5341	0.5320	-0.4175
Baseline	Italian	1	0.5186	0.5417	0.2265
RETUYT-INCO	Italian	A	0.4945	0.5128	-2.6399
Archaeology	Italian	2	0.4790	0.4805	-0.0599
GMU	Italian	1	0.2919	0.2961	0.0770
GMU	Italian	A1	0.1797	0.1706	-0.0064
SCaLAR	Italian	A1	-0.0234	-0.0425	-0.0643

Team Name	Language	ID	Pearson's	Spearman's	R^2
TMU-HIT	Japanese	2	0.7333	0.7305	0.4129
TMU-HIT	Japanese	1	0.6448	0.6479	-0.0958
Baseline	Japanese	1	0.6420	0.6684	0.3395
Archaeology	Japanese	2	0.4851	0.5126	-0.0983
RETUYT-INCO	Japanese	A1	0.4054	0.4073	-0.5215
Archaeology	Japanese	1	0.2803	0.2648	-2.2358
GMU	Japanese	A1	0.1775	0.1827	0.0241
GMU	Japanese	1	0.0350	0.0408	-0.0393
SCaLAR	Japanese	A1	-0.0660	-0.0784	-0.1007

Team Name	Language	ID	Pearson's	Spearman's	R^2
TMU-HIT	Portuguese	A2	0.7858	0.7988	0.1533
TMU-HIT	Portuguese	A1	0.7638	0.7729	-0.4987
Baseline	Portuguese	1	0.7126	0.7427	0.4890
Archaeology	Portuguese	1	0.7143	0.7102	-0.2612
Archaeology	Portuguese	2	0.6831	0.6923	0.2419
RETUYT-INCO	Portuguese	1	0.6772	0.7121	-1.5487
RETUYT-INCO	Portuguese	A1	0.6571	0.6899	-1.5931
SCaLAR	Portuguese	A1	0.0490	0.0270	-0.1825

Team Name	**Language**	**ID**	**Pearson's**	**Spearman's**	R^2
TMU-HIT	Sinhala	A2	0.3081	0.3343	-1.6030
TMU-HIT	Sinhala	A1	0.2482	0.3261	-3.0794
RETUYT-INCO	Sinhala	A1	0.1344	0.1094	-7.2755
GMU	Sinhala	1	0.1246	0.1303	-0.0370
ANU	Sinhala	2	0.0534	0.0866	-2.3263
SCaLAR	Sinhala	A1	0.0450	0.0279	-0.9819
Archaeology	Sinhala	2	0.0437	0.0298	-0.4590
GMU	Sinhala	A1	0.0263	0.0284	-0.1142
ANU	Sinhala	1	-0.0108	-0.0105	-15.5689
ANU	Sinhala	3	-0.0162	0.0487	-1.5636
Archaeology	Sinhala	1	-0.0290	-0.0272	-9.3516
Baseline	Sinhala	1	-0.1955	-0.2564	-0.2875

Team Name	**Language**	**ID**	**Pearson's**	**Spearman's**	R^2
TMU-HIT	Spanish	A2	0.7616	0.7460	0.4940
TMU-HIT	Spanish	A1	0.7201	0.6796	-0.0991
RETUYT-INCO	Spanish	2	0.6641	0.6547	0.2808
RETUYT-INCO	Spanish	A1	0.6397	0.6296	0.2541
Baseline	Spanish	1	0.5513	0.5299	0.2556
Archaeology	Spanish	1	0.5274	0.4793	0.2507
Archaeology	Spanish	2	0.5034	0.4588	0.2304
RETUYT-INCO	Spanish	1	0.3126	0.2369	0.0131
GMU	Spanish	1	0.2438	0.1984	-0.0731
GMU	Spanish	A1	0.1957	0.1772	-0.0806
SCaLAR	Spanish	A1	-0.0009	0.0180	-0.0367

C Lexical Simplification Results

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
TMU-HIT	All	1	0.3772	0.5498	0.4652	0.3421
TMU-HIT	All	2	0.3573	0.5498	0.457	0.3371
GMU	All	1	0.3345	0.4828	0.379	0.2754
TMU-HIT	All	3	0.2933	0.5498	0.4461	0.3306
RETUYT-INCO	All	1	0.2156	0.3324	0.2412	0.165
RETUYT-INCO	All	2	0.2074	0.3216	0.2351	0.1608
GMU	All	2	0.1331	0.2999	0.1981	0.1561
Archaeology	All	A1	0.0538	0.134	0.0882	0.0713

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
ISEP	Catalan	1	0.2719	0.3932	0.5003	0.3759
TMU-HIT	Catalan	A1	0.2584	0.3707	0.469	0.3547
TMU-HIT	Catalan	A2	0.2516	0.3707	0.4578	0.348
GMU	Catalan	1	0.2247	0.328	0.362	0.2641
RETUYT-INCO	Catalan	A1	0.1977	0.2943	0.3024	0.21
Baseline	Catalan	1	0.1977	0.2898	0.3000	0.2121
TMU-HIT	Catalan	A3	0.1955	0.3707	0.4528	0.345
RETUYT-INCO	Catalan	A2	0.1932	0.2831	0.3077	0.2106
GMU	Catalan	2	0.0651	0.1595	0.172	0.1408
Archaeology	Catalan	2	0.0404	0.1101	0.1203	0.0972
Archaeology	Catalan	1	0.0292	0.0651	0.069	0.0556

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
TMU-HIT	English	1	0.5245	0.7456	0.5762	0.4142
GMU	English	1	0.5157	0.6894	0.513	0.3691
ANU	English	3	0.5105	0.6649	0.5324	0.3744
ANU	English	1	0.4684	0.6561	0.5069	0.3652
ISEP	English	1	0.4684	0.6754	0.5351	0.3877
ANU	English	2	0.4631	0.6421	0.4978	0.3524
TMU-HIT	English	2	0.4438	0.7456	0.5595	0.4042
Baseline	English	1	0.3877	0.5631	0.4241	0.2956
RETUYT-INCO	English	3	0.3789	0.5701	0.3832	0.2634
RETUYT-INCO	English	2	0.3438	0.5526	0.3718	0.2542
CocoNut	English	1	0.2298	0.3877	0.2303	0.1674
GMU	English	A2	0.1929	0.4157	0.2339	0.1869
GMU	English	2	0.1859	0.3561	0.1945	0.1454
Archaeology	English	2	0.0947	0.2578	0.151	0.1272

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
TMU-HIT	Filipino	A1	0.065	0.0878	0.1807	0.1189
TMU-HIT	Filipino	A2	0.0615	0.0878	0.1736	0.1147
GMU	Filipino	A1	0.0562	0.0685	0.1395	0.0916
GMU	Filipino	1	0.0561	0.0684	0.1392	0.0914
TMU-HIT	Filipino	A3	0.0404	0.0878	0.1592	0.1061
Archaeology	Filipino	1	0.0175	0.0298	0.0313	0.0215
GMU	Filipino	2	0.0157	0.0245	0.0449	0.0338
RETUYT-INCO	Filipino	A1	0.0087	0.0087	0.0154	0.0094
Archaeology	Filipino	2	0.007	0.0122	0.0141	0.0095
RETUYT-INCO	Filipino	A2	0.007	0.0087	0.0082	0.0051
Baseline	Filipino	1	0.007	0.007	0.0225	0.014

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
TMU-HIT	French	A1	0.426	0.6197	0.6977	0.5466
TMU-HIT	French	A2	0.4242	0.6197	0.694	0.5443
ISEP	French	1	0.3743	0.5711	0.6484	0.4996
GMU	French	A1	0.3661	0.514	0.5148	0.3946
GMU	French	1	0.3655	0.5131	0.5141	0.394
TMU-HIT	French	A3	0.3257	0.6197	0.6815	0.5368
RETUYT-INCO	French	A1	0.301	0.4559	0.3974	0.2754
Baseline	French	1	0.2952	0.3760	0.3674	0.2626
RETUYT-INCO	French	A2	0.2764	0.4278	0.3776	0.2662
GMU	French	A2	0.0845	0.2394	0.1725	0.149
Archaeology	French	2	0.072	0.1704	0.1447	0.121
Archaeology	French	1	0.065	0.1265	0.1044	0.0819

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
TMU-HIT	German	A1	0.4885	0.6695	0.4882	0.3548
TMU-HIT	German	A2	0.4411	0.6695	0.481	0.3504
GMU	German	A1	0.42	0.5817	0.4002	0.2874
GMU	German	1	0.4192	0.5824	0.4004	0.2874
TMU-HIT	German	A3	0.355	0.6695	0.4633	0.3398
RETUYT-INCO	German	A1	0.3022	0.434	0.2699	0.1787
RETUYT-INCO	German	A2	0.2671	0.4165	0.2626	0.1765
ISEP	German	1	0.2187	0.25	0.1984	0.1344
Baseline	German	1	0.1719	0.2192	0.1562	0.1054
GMU	German	2	0.1192	0.3	0.1852	0.1463
Archaeology	German	1	0.0614	0.114	0.0626	0.0484
Archaeology	German	2	0.028	0.0771	0.0388	0.0294

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
TMU-HIT	Italian	A1	0.4762	0.7188	0.5661	0.4126
TMU-HIT	Italian	A2	0.4657	0.7188	0.558	0.4078
ISEP	Italian	1	0.4245	0.6614	0.5064	0.3788
GMU	Italian	A1	0.4042	0.6309	0.4615	0.3328
GMU	Italian	1	0.4035	0.6315	0.4616	0.3328
TMU-HIT	Italian	A3	0.3708	0.7188	0.5454	0.4002
RETUYT-INCO	Italian	A1	0.3163	0.4973	0.3511	0.2434
RETUYT-INCO	Italian	A2	0.3022	0.485	0.3305	0.2253
Baseline	Italian	1	0.2964	0.4684	0.3310	0.2254
GMU	Italian	A2	0.1546	0.3567	0.246	0.1965
Archaeology	Italian	2	0.0947	0.1929	0.1145	0.092
Archaeology	Italian	1	0.0491	0.1508	0.0975	0.0755

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
TMU-HIT	Japanese	1	0.4	0.5771	0.4883	0.3588
TMU-HIT	Japanese	A1	0.3989	0.5764	0.4881	0.3586
TMU-HIT	Japanese	2	0.3824	0.5771	0.4779	0.3526
GMU	Japanese	A1	0.2583	0.4393	0.3618	0.2599
GMU	Japanese	1	0.2578	0.4385	0.3612	0.2595
Baseline	Japanese	1	0.1561	0.2421	0.1735	0.1173
GMU	Japanese	A2	0.1195	0.2847	0.2144	0.171
RETUYT-INCO	Japanese	A1	0.0949	0.137	0.1026	0.0665
RETUYT-INCO	Japanese	A2	0.0878	0.1405	0.0949	0.0607
Archaeology	Japanese	2	0.0368	0.0929	0.0592	0.0441
Archaeology	Japanese	1	0.0263	0.0824	0.0516	0.0391

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
ISEP	Portuguese	1	0.485	0.6684	0.3538	0.2421
TMU-HIT	Portuguese	A1	0.4432	0.6595	0.3451	0.2285
TMU-HIT	Portuguese	A2	0.4095	0.6595	0.3341	0.2219
TMU-HIT	Portuguese	A3	0.3776	0.6595	0.3297	0.2193
Baseline	Portuguese	1	0.3509	0.4973	0.2330	0.1516
RETUYT-INCO	Portuguese	2	0.2768	0.4514	0.2094	0.136
RETUYT-INCO	Portuguese	A1	0.2748	0.4503	0.2088	0.1356
RETUYT-INCO	Portuguese	A2	0.2606	0.4202	0.207	0.1341
Archaeology	Portuguese	2	0.097	0.2539	0.092	0.0704
Archaeology	Portuguese	1	0.0864	0.2116	0.079	0.0574

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
GMU	Sinhala	A1	0.2284	0.3163	0.1387	0.0894
GMU	Sinhala	1	0.2283	0.32	0.14	0.0902
TMU-HIT	Sinhala	A2	0.2214	0.3585	0.1673	0.108
TMU-HIT	Sinhala	A1	0.2144	0.3585	0.1709	0.1101
GMU	Sinhala	A2	0.13	0.3057	0.1147	0.0759
TMU-HIT	Sinhala	A3	0.1195	0.3585	0.1469	0.0957
Archaeology	Sinhala	1	0.0466	0.0783	0.0359	0.0242
ANU	Sinhala	1	0.0133	0.0166	0.0074	0.0045
RETUYT-INCO	Sinhala	A1	0.0017	0.0017	0.0041	0.0024
Archaeology	Sinhala	2	0	0	0	0
RETUYT-INCO	Sinhala	A2	0	0	0.0032	0.0019
Baseline	Sinhala	1	0.0000	0.0033	0.0028	0.0017

Team Name	Language	ID	Acc@1@Top1	Acc@3@Top1	MAP@3	MAP@5
TMU-HIT	Spanish	A1	0.4536	0.6526	0.6763	0.5276
TMU-HIT	Spanish	A2	0.4502	0.6526	0.6721	0.5251
GMU	Spanish	1	0.4182	0.6087	0.5987	0.4653
GMU	Spanish	A1	0.4165	0.6053	0.5948	0.4627
TMU-HIT	Spanish	A3	0.3642	0.6526	0.6592	0.5174
RETUYT-INCO	Spanish	3	0.3288	0.4839	0.4124	0.298
Baseline	Spanish	1	0.3254	0.4519	0.4157	0.3019
RETUYT-INCO	Spanish	A1	0.3187	0.4957	0.4075	0.2879
RETUYT-INCO	Spanish	A2	0.3069	0.4688	0.399	0.2789
GMU	Spanish	A2	0.236	0.4704	0.4371	0.3542
Archaeology	Spanish	2	0.0674	0.1736	0.1565	0.1292
Archaeology	Spanish	1	0.0455	0.1112	0.0951	0.0756