An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset using the MultiLS Framework

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[10,14], Adam Nohejl[9], Kai North[11], Laura Occhipinti[12], Nelson Peréz Rojas[5], Nishat Raihan[11], Tharindu Ranasinghe[13], Martin Solis Salazar[5], 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]CATALPA, 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]University of Bath
m.shardlow@mmu.ac.uk

Abstract

We present preliminary findings on the MultiLS dataset, developed in support of the 2024 Multilingual Lexical Simplification Pipeline (MLSP) Shared Task. This dataset currently comprises of 300 instances of lexical complexity prediction and lexical simplification across 10 languages. In this paper, we (1) describe the annotation protocol in support of the contribution of future datasets and (2) present summary statistics on the existing data that we have gathered. Multilingual lexical simplification can be used to support low-ability readers to engage with otherwise difficult texts in their native, often low-resourced, languages.

Keywords: lexical simplification, lexical complexity prediction, MultiLS

1. Introduction

The lexical simplification pipeline is a family of systems concerned with the task of automatically identifying and replacing 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) which directly rewrites entire sentences. The two core operations included in the lexical simplification pipeline are (1) lexical complexity prediction and (2) the replacement of complex words with simple synonyms. Other varied operations exist in the text simplification ecosystem (Cardon and Bibal, 2023) which may be handled within lexical simplification depending on the specific implementation of the pipeline.

The task of Lexical Complexity Prediction (LCP) (Shardlow et al., 2020, 2022; North et al., 2023b), a form of Complex Word Identification (CWI) (Shardlow, 2013), involves assigning continuous values in the range 0-1 to given tokens in context, representing the difficulty that an intended reader population may associate with that target word. LCP was previously explored through a shared task (Shardlow et al., 2021) at SemEval 2021.

The second task, often referred to just as lexical simplification (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 (Pavlick and Callison-Burch, 2016; Maddela et al., 2021).

In addition to these two tasks, lexical simplification pipeline systems often take into account word sense disambiguation (Saggion et al., 2016), independent substitution generation / selection (Qiang et al., 2020) and grammaticality filtering (Gooding and Kochmar, 2019) steps — which are not explicitly explored in our dataset.

We identify two shortcomings of current work on the lexical simplification pipeline as follows:

1. Current datasets only explore one pipeline operation, but no dataset exists 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 some recent efforts exist to provide open source data in languages other than English, there is no guarantee that these datasets are created using similar protocols.

We introduce the MultiLS dataset[1] to address these two issues, based on the MultiLS framework (North et al., 2024). MultiLS is a new dataset that unites the related tasks of LCP and lexical simplification. Each instance in MultiLS contains a single target within an authentic context, which has been annotated for both the difficulty of the target (0-1) and relevant simplifying substitutions for the target. MultiLS is available in 10 languages and each language has the same amount of data, providing equality in provision between language sources.

2. Related Work

Current systems adopting the lexical simplification pipeline make use of transformer technology as described in detail in a recent survey by North et al. (2023b). In this section we particularly focus on the multilingual resources available for (1) Full-pipeline lexical simplification systems (2) LCP datasets and (3) Lexical simplification datasets.

Whilst several recent works exist implementing the lexical simplification pipeline in English making use of transformer-based technology (Qiang et al., 2021a; Baez and Saggion, 2023), there have also been significant efforts in Spanish to implement lexical simplification systems both for European Spanish (Alarcon et al., 2021; Stajner et al., 2023) and for Latin American variants such as Ecuadorian Spanish (Ortiz-Zambrano et al., 2023). The full pipeline has also been implemented in Swedish (Graichen and Jonsson, 2023), French (Rolin et al., 2021) and Chinese (Qiang et al., 2021b), making use of language-specific monolingual transformer based models. The lexical simplification pipeline is typically implemented as a monolingual task. However, there are also efforts to implement multilingual systems for simplification (Sheang and Saggion, 2023; Liu et al., 2023), which rely on multilingual language models trained on the TSAR-2022 shared task data for English, Spanish and Portuguese (Štajner et al., 2022).

An LCP dataset comprises of target words in context with a continuous value representing the difficulty of that target. LCP datasets were released for *English* through previous shared tasks (Yimam et al., 2018; Shardlow et al., 2021). There are 70K instances of LCP judgements available for English across these three shared task datasets, with additional data released through these efforts for *Spanish* and *German*. Recent research addressed the prediction of lexical difficulty for foreign language readers of French (Tack, 2021). Additionally, other research for French has implemented LCP annotations in the medical context (Sheang et al., 2022; Koptient and Grabar, 2022). For *Japanese*, the recent JaLeCoN dataset (Ide et al., 2023) provides 10K LCP annotations for news text and 8K LCP annotations for governmental texts. Published work to develop LCP annotations for languages other than English has also taken place for *Russian* (Abramov and Ivanov, 2022; Abramov et al., 2023), *Turkish* (Ilgen and Biemann, 2023), *Chinese* (Yang et al., 2023) and *Malay* (Omar et al., 2022).

Lexical simplification datasets comprise of a context, with a marked target word and a list of potential simplifying substitutions for that target word. The TSAR-2022 shared task data (Štajner et al., 2022) provides instances of lexical simplifications for *English*, *Spanish* (Ferrés and Saggion, 2022) and *Portuguese* (North et al., 2022, 2023a). Additionally, for Spanish, the EASIER Corpus (Alarcon et al., 2023) provides further simplification data. We also identified suitable simplification resources for *French* (Billami et al., 2018), *Japanese* (Kajiwara and Yamamoto, 2015; Kodaira et al., 2016), *Chinese* (Qiang et al., 2021b) and an additional resource for Portuguese (Hartmann et al., 2018).

3. MultiLS Dataset

We introduce the MultiLS dataset and describe the Trial data, comprising of 30 instances per language for 10 available languages. MultiLS provides LCP and lexical simplification annotations on common targets and contexts for each available language, significantly extending the availability of multilingual lexical simplification pipeline data. The full MultiLS dataset including a further 5,600 test instances across all 10 languages will be released as part of the 2024 MLSP shared task (Shardlow et al., 2024).

3.1. Annotation Protocol

We gathered an international team of 21 researchers representing 14 institutions and based across 8 countries. Each researcher was tasked with coordinating the annotations for one or more of the languages in our dataset. To guide the varied teams of dataset providers, we produced a comprehensive set of annotation guidelines. The key points from these guidelines are described below. The full guidelines are available with the dataset to encourage future contributions of additional languages.

3.1.1. Data Preparation

Dataset providers selected appropriate instances in their target language, focusing on contexts or single words (i.e., not multi-word expressions).

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

The definition of *word* may change from one language to another, especially when handling languages with non-Alphabetic scripts. The words were selected in each language to ensure sufficient difficulty to warrant lexical complexity annotation, and particularly to ensure that annotators will be able to find some simpler substitutions for the word in context. We provided a sample list of 200 words in English with the aim of encouraging common words across languages. However, due to language-specific constraints, not all providers used this list to make their selection. Whilst this was not enforced, there are some common targets across language pairs which can be used for future investigations.

Once the words had been selected, dataset providers identified 200 contexts in their target language, where each context contained one of the target words. Data providers were also free to select 200 contexts and then choose appropriate target words within those contexts. The contexts were selected from a readily available source in each language, specifically one that is related to an educational setting and released under a license that allows further redistribution of the text. For each context, an additional 2 words were selected for annotation. The requirement to select 200 contexts, with 3 words per context gave rise to 600 instances in total per language. An example is given below in English, with the selected target words highlighted in bold text:

> **Folly** is set in **great dignity**.

Note that the highlighted words: 'Folly', 'great' and 'dignity' all bear semantic content. The remaining words (the copula 'is' and the preposition 'in') are short words that do not have much influence on the overall meaning of the text. Particularly, it would be hard to find substitutions for these.

3.1.2. Annotator Selection

We requested that data providers solicited a minimum of 10 annotations per instance. Data providers were instructed to select annotators according to a 'Target group', which was also recorded as metadata indicating that the annotations received were reflective of the needs of the target group. For each annotator, the following additional elements of metadata were collected: (1) The number of years the annotator has spent in education; (2) Whether or not they are a native speaker of the language that is being annotated; (3) Age; (4) Typical number of hours they spend reading per week; (5) First Language; and (6) Number of languages they speak.

Dataset providers were able to either choose the same annotators to perform both lexical simplification and LCP, or to choose different groups for each task. For example if the target group was language learners, the data provider may have chosen to ask the learners to provide LCP annotations and their teachers to provide lexical simplification annotations.

3.1.3. Data Annotation for Lexical Complexity

Annotations for lexical complexity were performed using a 5-point Likert scale, with the following points translated into each language:

1. Very Easy - Words that are very familiar to you
2. Easy - Words that are mostly familiar to you
3. Neutral - Words that are neither difficult nor easy to you
4. Difficult - Words whose meaning is unclear, but that you may be able to infer from context
5. Very Difficult - Words that you have never seen before, or whose meaning is very unclear

Each instance was presented to the annotator with the full context and the annotators were asked to provide an independent judgement for each of the three highlighted words per context. Dataset providers additionally performed manual quality control on the resulting annotations, such as checking that the annotators had used the full range of annotations and that the complexity judgements were in line with those of other annotators'. The 1-5 annotations were converted to 0-1 following the Complex 2.0 format (Shardlow et al., 2022).

We did not typically enforce annotator agreements, but instead relied on manual evaluation of the outputs of annotators by the dataset providers. All provided data was quality checked and adjusted to ensure consistency where needed.

3.1.4. Data Annotation for Lexical Simplification

For each target word, annotators provided a minimum of 1 and a maximum of 3 words that could be used to simplify the target in the given context. The substitutions were selected to ensure (a) that the meaning of the original word and the overall context was preserved, and (b) that the substitution was easier to understand than the original target. For some of the target words, it was not easy to find appropriate simplifications in the contexts that they are presented in. For instance, a word may already be sufficiently simple, or despite being complex there may be no simpler alternatives. In these cases, the annotators were instructed to write the original word, or to leave the field blank and indicate that the original word is the simplest word that could fit in this context.

Data providers performed quality control through manual verification of the submissions of each annotator by checking (a) the suitability of the substitutions within the context, and (2) the frequency with which annotators were unable to find a simplification.

For some languages, a substitution may cause issues regarding the agreement with surrounding words (e.g. a masculine noun replaced by a feminine substitution will require to revise the gender of its related adjectives or determiners). We decided to treat morphological adaptation as a separate task that is left aside. Annotators were informed that they may propose substitutions that do not strictly fit in the grammatical context regarding gender/number agreement.

3.2. MultiLS Trial Data

Presently, we have released 30 instances per language for the 10 languages in Table 1. We report the aggregated metadata for each language, as well as summary statistics on the overall dataset.

4. Discussion

In this work we have presented a data annotation effort for LCP and lexical simplification which is intended to be extensible to a wide variety of languages. Currently, 8 out of the 10 languages represented are Indo-European, with 5 Romance Languages (French, Italian, Spanish, Portuguese and Catalan), 2 West-Germanic languages (English and German) and Sinhala which is of the Indo-Aryan family. Additionally we have Filipino and Japanese which are of the Austronesian and Japonic families respectively. Eight of the languages make use of the Latin script, with Sinhala and Japanese being exceptions to this. The Latin script languages are alphabetic, whereas Sinhala is an abugida language (characters represent a combination of vowel and consonants) and Japanese script features kanji (logographic characters loaned from Chinese) and kana (syllabic characters). The available languages are a result of the collaborative team that we were able to gather. We hope to extend the language families, scripts and script types represented in future iterations of the dataset.

The target groups and text genres represented in the language subsets of our dataset are varied as shown in the second and third column of Table 1. This reflects the availability of target texts in each language as well as the available pools of annotators that we were able to access. We expect that this will result in some variations between datasets reflecting the interests of the target groups. We have exposed this information to help those working with our dataset to adjust systems according to each target group. We will also make summary metadata regarding our annotators for each subset available alongside the dataset.

The average complexity in our dataset varies across subsets. The average complexity of each subset is below 0.5 (0 = Very Easy, 1 = Very Difficult). All datasets contain examples of complex language (> 0.5), but the low average complexities represents the fact that the majority of identified tokens were assigned easier complexity values (< 0.5) by annotators. This is representative of Zipfian language distributions, where most frequent words are familiar, with few rare complex words.

The context length also varies between subsets of our data. Japanese has a particularly short context length as each kanji character is a logographic unit, leading to fewer characters per sentence. Considering the other languages, the texts selected for Filipino have a typically short context length (64.066) whereas those selected for Catalan have a generally long context length (239.533). We also note significant variations in the number of unique substitutions per language with an average of 3.967 substitutions per instance for Filipino and 15.8 for Japanese. Each language is unique and the variations arise from the target groups, text genres, annotator pool and language specific factors. We deliberately present the MultiLS dataset as a composite of sub-language datasets to allow and encourage the development of language-specific and multilingual simplification interventions and technologies.

We initially aimed for a high degree of uniformity in our dataset across language subsets. However, to prioritise the inclusion of more languages we chose to relax the inclusion criteria to incorporate existing efforts to annotate LCP/LS for interesting and diverse text types and genres. Additionally, our approach gave significant agency to the native-speaking dataset providers in each language-setting to make linguistically appropriate decisions for their bespoke context. The result is a dataset with lower intra-lingual conformity, but ultimately a larger, more diverse and easily extensible dataset.

Recent efforts in lexical simplification within English have focussed on personalised approaches to (a) complexity detection (Gooding and Tragut, 2022) and (b) simplification (Sukiman et al., 2024), which seeks to model the individual reader, as opposed to building a single model for all readers. Our proposed dataset only provides a single aggregated judgement per instance, meaning that it is not useful for personalised lexical simplification pipelines in its current form. The authors will ex-

Language	Target Group	Text Genre	Mean Complexity	Mean Context Length	Mean # Unique Subs
Catalan	Varied	News	0.487 (0.125)	239.533 (70.128)	14.167 (3.354)
English	University Students	Wikibooks	0.200 (0.201)	111 (36.992)	6.167 (1.859)
Filipino	University Staff	Educational Books	0.171 (0.126)	64.066 (22.137)	3.967 (1.098)
French	Language Learners	Varied	0.371 (0.229)	129.1 (45.564)	10.067 (3.463)
German	High-School Students	Wiki / Literary	0.413 (0.191)	195.733 (59.604)	8.067 (2.791)
Italian	Native Speakers	Wikibooks/Wikiquote	0.248 (0.168)	168.4 (67.614)	7.800 (2.952)
Japanese	Language Learners	Varied	0.259 (0.173)	37.8 (7.303)	15.800 (4.634)
Portuguese	MTurk Workers	Varied	0.273 (0.165)	165.9 (74.062)	5.367 (1.217)
Sinhala	University Staff	News / Religious	0.243 (0.214)	163.4 (52.554)	4.333 (0.606)
Spanish	Varied	Educational Books	0.449 (0.233)	178.7 (48.075)	10.867 (3.785)

Table 1: Dataset metadata and statistics for the MultiLS trial data organised alphabetically by the English name of the language. All values given as mean average with standard deviation in brackets. Context length is reported as character length for cross-lingual comparison.

plore the use of the unaggregated annotator-level complexity predictions to better understand how we can use this data for personalised judgements.

5. Conclusion

We present the MultiLS dataset, comprising of LCP and lexical simplification data for 10 languages. MultiLS is an extensible framework and is open to contributions of additional languages and to additional data for the existing languages (North et al., 2024). MultiLS will allow future researchers to develop truly multilingual lexical simplification pipeline systems. We include one instance per language in Table 2 in the Appendix.

6. Acknowledgments

Andrea Horbach 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). 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. Horacio Saggion, Stefan Bott and Akio Hayakawa acknowledge funding from the European Union´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).

7. Bibliographical References

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

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

Rodrigo Alarcon, Lourdes Moreno, and Paloma Martínez. 2021. Lexical simplification system to improve web accessibility. *IEEE Access*, 9:58755–58767.

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

Anthony Baez and Horacio Saggion. 2023. LSLlama: Fine-tuned LLaMA for lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 102–108.

Mokhtar B. Billami, Thomas François, and Núria Gala. 2018. ReSyf: a French lexicon with ranked synonyms. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2570–2581, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rémi Cardon and Adrien Bibal. 2023. On operations in automatic text simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 116–

130, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Daniel Ferrés and Horacio Saggion. 2022. ALEXSIS: A dataset for lexical simplification in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3582–3594, Marseille, France. European Language Resources Association.

Sian Gooding and Ekaterina Kochmar. 2019. Recursive context-aware lexical simplification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4853–4863, Hong Kong, China. Association for Computational Linguistics.

Sian Gooding and Manuel Tragut. 2022. One size does not fit all: The case for personalised word complexity models. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 353–365, Seattle, United States. Association for Computational Linguistics.

Emil Graichen and Arne Jonsson. 2023. Context-aware Swedish lexical simplification. In *Proceedings of the Second Workshop on Text Simplification, Accessibility and Readability*, pages 11–20, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Nathan S Hartmann, Gustavo H Paetzold, and Sandra M Aluísio. 2018. SIMPLEX-PB: A lexical simplification database and benchmark for Portuguese. In *Computational Processing of the Portuguese Language: 13th International Conference, PROPOR 2018, Canela, Brazil, September 24–26, 2018, Proceedings 13*, pages 272–283. Springer.

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 Eighteenth Workshop on Innovative Use of NLP for Building Educational Applications*. Association for Computational Linguistics.

Tomoyuki Kajiwara and Kazuhide Yamamoto. 2015. Evaluation Dataset and System for Japanese Lexical Simplification. In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pages 35–40, Beijing, China. Association for Computational Linguistics.

Tomonori Kodaira, Tomoyuki Kajiwara, and Mamoru Komachi. 2016. Controlled and Balanced Dataset for Japanese Lexical Simplification. In *Proceedings of the ACL 2016 Student Research Workshop*, pages 1–7.

Anaïs Koptient and Natalia Grabar. 2022. Automatic detection of difficulty of French medical sequences in context. In *Proceedings of the 18th Workshop on Multiword Expressions @LREC2022*, pages 55–66, Marseille, France. European Language Resources Association.

Kang Liu, Jipeng Qiang, Yun Li, Yunhao Yuan, Yi Zhu, and Kaixun Hua. 2023. Multilingual lexical simplification via paraphrase generation. *arXiv preprint arXiv:2307.15286*.

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.

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.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2024. MultiLS: A Multi-task Lexical Simplification Framework.

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.

Jenny A Ortiz-Zambrano, César Espin-Riofrio, and Arturo Montejo-Raéz. 2023. LegalEc: A new corpus for complex word identification research in law studies in Ecuatorian Spanish. *Procesamiento del Lenguaje Natural*, 71:247–259.

Ellie Pavlick and Chris Callison-Burch. 2016. Simple PPDB: A Paraphrase Database for Simplification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 143–148, Berlin, Germany. Association for Computational Linguistics.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, Yang Shi, and Xindong Wu. 2021a. LSBert: Lexical simplification based on BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3064–3076.

Jipeng Qiang, Yun Li, Yi Zhu, Yunhao Yuan, and Xindong Wu. 2020. Lexical simplification with pretrained encoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34-05, pages 8649–8656.

Jipeng Qiang, Xinyu Lu, Yun Li, Yunhao Yuan, and Xindong Wu. 2021b. Chinese lexical simplification. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:1819–1828.

Eva Rolin, Quentin Langlois, Patrick Watrin, and Thomas François. 2021. FrenLyS: A Tool for the Automatic Simplification of French General Language Texts. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021), Held Online, 1-3 September, 2021*, pages 1196–1205. INCOMA Ltd.

Horacio Saggion, Stefan Bott, and Luz Rello. 2016. Simplifying words in context. experiments with two lexical resources in spanish. *Computer Speech and Language*, 35:200–218.

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

Matthew Shardlow. 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, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024. 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, 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, and Marcos Zampieri. 2022. Predicting lexical complexity in english texts: the complex 2.0 dataset. *Language Resources and Evaluation*, 56(4):1153–1194.

Kim Cheng Sheang, Anaïs Koptient, Natalia Grabar, and Horacio Saggion. 2022. Identification of complex words and passages in medical documents in French. In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 116–125, Avignon, France. ATALA.

Kim Cheng Sheang and Horacio Saggion. 2023. Multilingual controllable transformer-based lexical simplification. *Procesamiento del Lenguaje Natural*, 71:109–123.

Sanja Stajner, Daniel Ibanez, and Horacio Saggion. 2023. LeSS: A computationally-light lexical simplifier for Spanish. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1132–1142, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Safura Adeela Sukiman, Nor Azura Husin, Hazlina Hamdan, and Masrah Azrifah Azmi Murad. 2024. A Hybrid Personalized Text Simplification Framework Leveraging the Deep Learning-based Transformer Model for Dyslexic Students. *Journal of Advanced Research in Applied Sciences and Engineering Technology*, 34(1):299–313.

Anaïs Tack. 2021. *Mark my words! On the automated prediction of lexical difficulty for foreign language readers*. Ph.D. thesis, UCL-Université Catholique de Louvain.

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.

8. Language Resource References

Aleksei V. Abramov and Vladimir V. Ivanov. 2022. Collection and evaluation of lexical complexity data for russian language using crowdsourcing. *Russian Journal of Linguistics*, 26(2):409–425.

Aleksei V Abramov, Vladimir V Ivanov, and Valery D Solovyev. 2023. Lexical complexity evaluation based on context for russian language. *Computación y Sistemas*, 27(1):127–139.

Bahar Ilgen and Chris Biemann. 2023. CWITR: A corpus for automatic complex word identification in Turkish texts. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 157‑163, New York, NY, USA. Association for Computing Machinery.

Salehah Omar, Juhaida Abu Bakar, Maslinda Mohd Nadzir, Nor Hazlyna Harun, and Nooraini Yusoff. 2022. Malay lexical simplification model for non-native speaker. In *2022 International Conference on Intelligent Systems and Computer Vision (ISCV)*, pages 1–6.

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.

Cheng-Zen Yang, Jin-Jian Li, and Shu-Chang Lin. 2023. Lexical complexity prediction using word embeddings. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 279–287, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

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.

Appendix

Language	Target	Context	Complexity	Substitutions
Catalan	successius	Els manifestants han denunciat que en els darrers anys ”els successius governs d’aquí i d’allà han passat del menyspreu al desmantellament de l’escola pública, començant a Catalunya per la pròpia LEC” i han alertat que s’està ”deteriorant greument la qualitat de l’educació pública i les condicions laborals dels treballadors”.	0.45	diferents, successius, contigus, differents, diversos, continus, consecutius, seguit, seguits, ultims, tots, tot el conjunt, succesius, anteriors, previs
English	external	(If your robot has an external ROM chip, then that is the one that is pulled and replaced.	0.05	outer, outside, exterior, external
Filipino	agaw	Akin na ’yan! biglang agaw ni Karlo sa laruan ni Lara.	0.075	agaw, kuha, kinuwa, nakaw
French	entretien	Cette gratuité n’est que partielle puisque une partie des impôts fonciers payés par les propriétaires est consacrée à l’entretien des réseaux de distribution.	0.675	réparation, maintien, nettoyage, gestion, entretien, tenue, restauration, réfection, revue, maintenance, soins
German	Grausen	Das Grausen überwältigte alle seine Sinne, er stürzte verworren aus dem Zimmer durch die öden widerhallenden Gemächer und Säulengänge hinab.	0.6	Angst, Furcht, Schreck, Panik, Gruseln, Grauen, Entsetzen
Italian	perduta	Aveva l’aria di una perduta nobiltà: la miseria gli si leggeva tutta nel volto e nella camicia sbrindellata sul petto.	0.08	persa, andata, antica, finita, inesistente, passata, smarrita”
Japanese	掲載した	ドラマに関する感想を募集し、週ごとにピックアップして回答も掲載した。	0.32	載せた，書いた，紹介した，発表した，公開した，知らせた，紙面に載せた，記載し情報を伝えた
Portuguese	estradas	as equipes contratadas pelo departamento de estradas de rodagem do paraná (der/pr) estão fazendo a primeira camada de asfalto no trecho da rodovia que passa por baixo da trincheira da jacob macanhan e também nos acessos laterais para a avenida camilo di lellis e para os bairros da região de pinhais	0.08	ruas, caminhos, vias, rodovias
Sinhala	තථාගතයන්	සකල ක්ලේශ ප්‍රහාණයෙන් ලෝකෝත්තර තත්ත්වයෙහි වැඩ විසූ තථාගතයන් වහන්සේ සම්මා සම්බුද්ධ නම් වූ ලොවුතුරා සම්බුදු පදවියට පත්වූයේ උතුම් මිනිස් ජීවිතයක් ලැබීමෙන් බව අප අමතක නොකළ යුතුව ඇත.	0.67	බුදුරජාණන්, භාග්‍යවතුන්, බුදුන්, සම්බුදුරජාණන්, බුදුවරයන්
Spanish	notifique	Notifique a su Banco o institución financiera la pérdida o robo de sus tarjetas, chequeras o si sospecha que alguien está utilizando sus cuentas sin su permiso.	0.45	diga, avise, informe, comunique, avisa, alerten, comuniquen, expliquen, indique

Table 2: One example per language in the dataset