General Paper Peer-Reviewed

A Japanese Dataset and Efficient Multilingual LLM-Based Methods for Lexical Simplification and Lexical Complexity Prediction

Adam Nohejl[a], Akio Hayakawa[b], Yusuke Ide[a] and Taro Watanabe[a]

Lexical simplification (LS) is the task of making text easier to understand by replacing complex words with simpler equivalents. LS involves the subtask of lexical complexity prediction (LCP). We present MultiLS-Japanese, the first unified LS and LCP dataset targeting non-native Japanese speakers, and one of the ten language-specific MultiLS datasets. We propose methods for LS and LCP based on large language models (LLMs) that outperform existing LLM-based methods on 7 and 8 of the 10 MultiLS languages, respectively, while using only a fraction of their computational cost. Our methods rely on a single prompt across languages and introduce a novel calibrated token-probability scoring technique, G-Scale, for LCP. Our ablations confirmed the benefits of G-Scale and of concrete wording in the LLM prompt. We made the MultiLS-Japanese dataset available online under a CC-BY-SA license, including detailed metadata.

Key Words: *Automated Text Simplification, Lexical Simplification, Lexical Complexity Prediction, Large Language Models*

1 Introduction

Lexical simplification (LS) is the task of making text easier to understand by replacing complex words with simpler equivalents. Therefore, it is a specific type of automatic text simplification (ATS) that operates at the word level without changing the overall sentence or text structure.

An LS system can serve as a reading assistance tool for end users, a tool for human editors, or a component of a larger text simplification system. It has been studied in the context of various

[a] Nara Institute of Science and Technology
[b] Universitat Pompeu Fabra
Sections 3.1 and 3.2 of this article partially overlap in content with previous publications (Shardlow et al. 2024b; Nohejl et al. 2024, Sec. 4.1.6 and Sec. 2, respectively); part of the MultiLS-Japanese dataset was briefly described in these publications. As part of this work, the complete dataset is made available at `https://huggingface.co/datasets/naist-nlp/multils-japanese`

target populations, such as non-native speakers (Hading et al. 2016; Paetzold and Specia 2016c; Lee and Yeung 2018, *inter alia*), or people with dyslexia, aphasia, or low literacy levels (Paetzold and Specia 2017).

The LS process is often broken down into two subtasks: (1) lexical complexity prediction (LCP) and (2) LS of the identified complex words, which can be further conceptualized as a multi-step pipeline (Paetzold and Specia 2017). LCP can also be used for other purposes, such as readability assessment.

LS datasets are scarce for languages other than English. For Japanese, only two LS datasets with an unspecified target audience and one LCP dataset for non-native speakers are available. Because each target audience has very different needs, we constructed a dataset targeting non-native Japanese speakers, MultiLS-Japanese, which provides gold labels for both LS and LCP. We carefully designed the dataset considering the target audience and the specifics of the Japanese language: we sampled sentences from varied genres read by non-native speakers, included target words from categories that pose difficulties to learners (e.g., compound verbs), and substituted the appropriate units given the Japanese morphology. For lexical complexity annotation, we excluded Chinese and Korean L1 speakers, whose native languages share a large part of their vocabulary with Japanese. The LS annotators were native speakers with experience teaching non-native Japanese learners.

MultiLS-Japanese is part of a wider family of ten language-specific MultiLS datasets. We therefore designed and evaluated LS and LCP methods that are applicable to any of these languages. Similar to current state-of-the-art methods, our methods are based on large language models (LLMs); in contrast, our methods use a single LLM inference per data instance and a novel calibrated token-probability-based scoring technique, G-Scale, for LCP. Thus, we reduced the computational cost by an order of magnitude compared to other LLM-based methods while providing superior performance.

Our LS and LCP methods outperform the existing methods on 7 and 8 of the 10 languages with available MultiLS datasets, respectively. G-Scale has the potential to be applied to other tasks of rating a text input on a continuous scale using LLMs. We evaluated ablations of both of our methods.

1.1 Lexical Simplification and Automatic Text Simplification

LS is one of the many approaches to ATS. Other approaches operate at the sentence, paragraph, or text level, allowing deletions or rewriting of larger text spans, with the common goal of making the text easier to understand. Datasets that enable the evaluation, and to some extent,

the fine-tuning of sentence simplification models for Japanese have recently become available (see Section 2.2). Furthermore, recent LLMs can, in principle, be employed to simplify text even without fine-tuning (Brown et al. 2020), and thus without extensive parallel data. This development raises a question: Should we abandon LS as an obsolete task and limit our focus to more general text simplification systems?

Because LS is both an indispensable operation for any ATS system and a practically useful task by itself, we believe that datasets and methods for LCP and LS are still an important research area complementary to higher-level ATS methods:

Practical utility. Vocabulary knowledge has been recognized as one of the strongest factors, and possibly the strongest, in reading comprehension (Grabe and Yamashita 2022, Ch. 11). LS can be used as a tool to assist language learners while still displaying the original text. Even without the LCP component, LS can be used for on-demand simplification or glossing of words using a simple user interface. In conjunction with LCP, words to be simplified can be selected automatically based on language proficiency or the particular learner's needs (personalized LCP). Both of these modes of use, showing simplifications as glosses and adjusting the simplification level using lexical complexity thresholds, allow balancing comprehension with opportunities to learn.[1]

Detailed data. Both LS and LCP datasets provide detailed word-level data, consisting of multiple simplifications and complexity ratings, which are not available in sentence simplification or readability datasets. Such data has been used for analyses (Gooding et al. 2021; Nohejl et al. 2024; Saggion et al. 2024), leading to the recognition of the needs of various target populations or problems in human-created simplifications. The insights from analyses of LS and LCP data are useful for any kind of ATS systems, as they inevitably perform simplifications at the lexical level as well.

Meaning preservation. Meaning preservation is a general challenge for all ATS systems, but it particularly affects systems operating at the sentence level or higher, which employ deletion or summarization (Agrawal and Carpuat 2024). LS is well-positioned to preserve the meaning of the original text. See Section 5.1.1 for an analysis of our method in terms of errors in meaning preservation and other areas.

[1] Commercially available Amazon Kindle e-readers have a feature called Word Wise, offering "simple definitions and synonyms displayed inline above more difficult words while you read" (`https://www.amazon.com/gp/help/customer/display.html?nodeId=201645250`) based on the desired simplification level, which demonstrates the desirability of this function. Word Wise, however, does not employ LS and requires publishers to supply the glosses and levels.

2 Related Work

2.1 Japanese LS and LCP Datasets

Lexical simplification. Two LS datasets are available for Japanese: The SNOW E4 dataset (Kajiwara and Yamamoto 2015) is based on newspaper articles and restricts target words to lemma form, although verbs and adjectives are inflected in Japanese. The Controlled and Balanced Dataset for Japanese Lexical Simplification (CBDJLS) (Kodaira et al. 2016) is based on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al. 2014). CBDJLS follows a more realistic setting, allowing not only target words in any word form but also substitutions spanning the adjacent *fuzokugo* (dependent words).[2] Both datasets consist of approximately 1,500 simplifiable instances (target words in sentences), covering around 200 unique target words, and provide a list of simplifications for each instance but no complexity ratings.

Neither of the two LS datasets specifies what population (e.g., non-natives or children) it targets, although it has been shown that non-native speakers have different needs from natives, further differentiated by their proficiency level and native language (Paetzold and Specia 2016c). Both datasets employed five crowdsourced annotators per instance, with SNOW E4 reporting no details and CBDJLS specifying only that annotators have completed 95% of their previous assignments correctly and were native speakers. The SNOW E4 dataset is available in full text, although redistribution is prohibited, and the CBDJLS dataset is distributed in a stand-off format that requires paid access to BCCWJ to reconstruct the complete data.

In contrast to these two available datasets, our dataset, MultiLS-Japanese, targets non-native Japanese speakers and is available under a permissive license, namely, the Creative Commons Attribution-ShareAlike (CC-BY-SA).

Lexical complexity. The only lexical complexity dataset for Japanese, Dataset of Japanese Lexical Complexity for Non-Native Readers (JaLeCoN) (Ide et al. 2023), provides complexity ratings by non-native speakers for single words and multiword expressions (MWEs). It consists of separate complexity ratings from annotators with Chinese or Korean L1 background and annotators with other L1 backgrounds, addressing the considerable advantage of the former in Japanese reading comprehension. The study (Ide et al. 2023) showed that words of Chinese origin or containing Chinese characters, which form a large part of the Japanese vocabulary, result in increased perceived complexity for annotators other than those with a Chinese or Korean L1

[2] See Appendix A for an English–Japanese glossary of linguistic terms. The authors of CBDJLS (Kodaira et al. 2016) mention only "particles"; however, in the actual data, other parts of speech are included in the substitutable context as well, all of which may be categorized as *fuzokugo*, e.g., the whole sequence だったら**撤退**しなければならないので including copulas, auxiliaries, and light verbs can be substituted.

background. The dataset is based on texts from news and governmental press conferences and is available in full text for non-commercial purposes. It comprises around 18,000 words of running text, all of which were annotated with complexity ratings. As a natural result of this annotation scheme, the vast majority of the words have zero complexity, and the rest are heavily skewed to the low complexity levels (Ide et al. 2023).

Our dataset, MultiLS-Japanese, also provides lexical complexity annotation for non-native speakers but aims to provide higher-quality ratings for a smaller number of words by explicitly annotating selected target words, as opposed to asking annotators to annotate complex words in a running text, skipping simple words at their discretion. Moreover, in contrast to JaLeCoN, which provides data only for LCP, MultiLS-Japanese has gold labels for both LCP and LS based on the MultiLS framework (North et al. 2024), enabling consistent evaluation of the whole LS pipeline.

2.2 Other Japanese ATS Datasets

LS is a specific task within the broader area of ATS. Another common ATS task is sentence simplification, and more recently, paragraph- and document-level simplification tasks have been studied as well.

In addition to LS datasets, there are several Japanese datasets for sentence simplification: An early news-based dataset (Goto et al. 2015) is not publicly available. The SNOW T15 (Maruyama and Yamamoto 2018) and SNOW T23 (Katsuta and Yamamoto 2018) datasets provide simplifications of sentences from the Tanaka corpus, a collection of short example sentences from language textbooks. JADES (Hayakawa et al. 2022) consists of simplifications of 3,907 sentences from news articles targeting non-native speakers. MATCHA (Miyata et al. 2024), which is based on articles from a website for international visitors to Japan,[3] provides 16,000 sentence simplifications targeting non-native speakers. JASMINE (Horiguchi et al. 2024) is a sentence-level medical domain evaluation dataset. JASMINE and the SNOW datasets focus on word- or phrase-level substitutions, forming an intermediate step between LS and sentence simplification, whereas JADES and MATCHA employ larger-scale rewriting operations.

All these datasets share the common features of being sentence-level and providing a single gold simplification. JADES and MATCHA target non-native speakers at the proficiency level corresponding to the Japanese Language Proficiency Test (JLPT) level N4, equivalent to level A2 (Japan Educational Exchanges and Services 2025) in the Common European Framework of Ref-

[3] `https://matcha-jp.com`

erence for Languages (CEFR). While MATCHA is larger, JADES has been manually annotated for operations (e.g., REPLACE, DELETE, and SPLIT), enabling easier analysis.

Compared with JADES and MATCHA (Hayakawa et al. 2022; Miyata et al. 2024, also see Section 2.2), our dataset targets more proficient speakers. JADES and MATCHA target relative beginners (JLPT N4), who need more aggressive simplification, involving deletions, splitting, or rewriting. At JLPT levels N1 and N2, which we target, we hypothesize that LS is sufficient to achieve a very high level of comprehension. Although our dataset is smaller than JADES or MATCHA, it covers a variety of genres and topics, whereas JADES and MATCHA are each based on a single source. The sentence simplification datasets and our dataset complement each other by providing data for different tasks (as discussed in Section 1.1) and targeting different proficiency levels.

2.3 Lexical Simplification

The early approaches to LS were rule-based (Devlin and Tait 1998; De Belder and Moens 2010). Data-driven approaches were soon applied (Yatskar et al. 2010; Biran et al. 2011; Horn et al. 2014) to English, where a large simplified corpus, Simple English Wikipedia,[4] is available. Later methods that did not require rules, databases, or simplified corpora were based on static word embeddings (Glavaš and Štajner 2015) and, more recently, on masked language models (MLMs) (Qiang et al. 2020a, 2020b).

The current state-of-the-art methods for LS leverage LLMs such as OpenAI's GPT-3 (Brown et al. 2020) and GPT-4 (OpenAI 2024a). An ensembled prompting method using GPT-3 (Aumiller and Gertz 2022) achieved a top result for English, Spanish, and Portuguese in a shared task. A derived method (Enomoto et al. 2024) using GPT-4 achieved the best overall result for ten languages in the Multilingual Lexical Simplification Pipeline (MLSP) shared task (Shardlow et al. 2024b), where Japanese was evaluated on MultiLS-Japanese (our dataset introduced in Section 3). This method is based on aggregating results from multiple LLM inferences, thus using significant resources to simplify each word.

Our LS method is also based on LLM prompting but uses a single LLM inference while improving accuracy.

2.4 Lexical Complexity Prediction

Word frequency has long been used as the variable most predictive of lexical complexity, particularly the frequency in film or YouTube subtitles (Shardlow 2013; Nohejl et al. 2025), which

[4] https://simple.wikipedia.org

approximates the spoken language. Lexical dispersion measured by the number of YouTube channels or videos containing a word was demonstrated to be an even stronger predictor (Nohejl and Watanabe 2025). Various other word-level features have been used as well; however, their applicability (e.g., for stress patterns) and availability (e.g., for psycholinguistic data) vary based on language (Shardlow et al. 2022, Sec. 3 lists 20 features possibly applicable to English). Shardlow et al. also emphasized that context is an important factor of lexical complexity because it determines in which sense a polysemous word is used and can add to complexity if a word is used in an unexpected context.

In feature-based models, attempts were made to model the interaction with context using n-gram probability of the immediate context of the target word or MLM-based surprisal (to model the unexpectedness of the word) and features based on dependency parsing. The results were mixed. Contextual surprisal was found to have a minimal effect (Tack 2021). Two shared tasks (Paetzold and Specia 2016a; Yimam et al. 2018) were dominated by feature-based systems focused on word-level features. Features that depend on context were limited to either only n-grams (Paetzold and Specia 2016b), without further analysis of the feature's contribution, or only dependency-based features (Gooding and Kochmar 2018), but in that case, dependencies were not among the top ten contributing features.

Later, using fine-tuned MLMs became the dominant approach in cases where enough training data was available (Shardlow et al. 2021).

In the MLSP shared task (Shardlow et al. 2024b), where lexical complexity was predicted in ten languages, the top-performing systems across all languages were, depending on the metric used, TMU-HIT (Enomoto et al. 2024) and Archaeology (Cristea and Nisioi 2024). TMU-HIT was based entirely on prompting LLMs, whereas Archaeology was based entirely on word-level features. A study (Smădu et al. 2024) comparing LCP using feature-based models, fine-tuned MLMs, and LLMs concluded that the computationally intensive methods based on LLMs barely outperform the earlier, more lightweight methods.

The method we propose for LCP is also LLM-based, but it reduces the computational cost compared to TMU-HIT while providing more accurate predictions. Our method can also combine LLM predictions with features, such as frequency, leading to further improvements in accuracy.

3 Dataset: MultiLS-Japanese

Japanese is well known to be difficult to learn as a second language (Grainger 2005; Foreign Service Institute 2025). Attaining reading or writing proficiency is particularly demanding be-

cause it involves mastering its complex writing system. Yet, no Japanese LS dataset targeting non-native speakers is available. To fill this gap, we constructed a unified LS and LCP dataset targeting non-native Japanese speakers.

3.1 Overview

MultiLS-Japanese is the first Japanese LS dataset targeting non-native speakers. More specifically, it targets non-native Japanese speakers whose first language is neither Chinese nor Korean. The data consists of 600 instances, where each instance is a target word in a sentence context, for which a lexical complexity value and simpler substitutions are provided. Table 1 compares MultiLS-Japanese with the available Japanese datasets for LS and LCP. While MultiLS-Japanese consists of a relatively small number of instances, it employs more annotators per instance and

Dataset	SNOW E4 (Kajiwara and Yamamoto 2015)	CBDJLS (Kodaira et al. 2016)	JaLeCoN (Ide et al. 2023)	MultiLS-Japanese (Ours)
Tasks	LS	LS	LCP	LS, LCP
Target Population	*Unspecified*	*Unspecified*	Non-native	Non-native
– L1	—	—	CK, non-CK	Non-CK
– L2 Proficiency	—	—	B1–C2[a]	B2–C1[b]
Annotators	5 *unspecified*	5 native	7 CK + 7 non-CK, all non-native	LCP: 10 non-native, LS: 10 native
– Profiles	—	—	L1; L2 proficiency	Detailed (→Sec. 3.7)
– Instructions	—	—	Summary	Full (→Appendix D)
Target Words	Single uninflected	Single + surrounding *fuzokugo*	Single or MWE (auto. + manual)	Single or MWE (manual)
# LS Instances	1,586	1,457	—	600
– Unique Words	205	201	—	600
# LCP Instances	—	—	18,220	600
– Non-Zero Complexity	—	—	657 CK, 3011 non-CK	572
– Unique Words	—	—	433 CK, 1658 non-CK	572
Individual Data	—	—	—	LS and LCP
Source Texts	News	Varied (BCCWJ)	News, press conferences	Varied (→Sec. 3.4)
License	Redistribution prohibited[c]	Annotation: MIT; text: BCCWJ[d]	CC-BY-NC-SA	CC-BY-SA

Table 1 Comparison of Japanese LS and LCP datasets.For all datasets, we count only simplifiable words as "LS Instances." "CK" stands for Chinese or Korean L1. [a] Self-reported. [b] Holders of JLPT (N)1, N(2). [c] Details in Japanese: `https://www.jnlp.org/GengoHouse/snow/e4` [d] Annotation in stand-off format, requires paid copyrighted BCCWJ data.

provides lexical simplifications for more unique words (types) than the previous datasets. For LCP, 572 of the 600 annotated instances in MultiLS-Japanese have non-zero mean complexity rating, in contrast with JaLeCoN, which is larger, but only a small fraction of the annotated words have non-zero complexity. This is a result of a different annotation scheme: JaLeCoN provides ratings for all words in a running text, whereas MultiLS-Japanese provides ratings only for selected target words.

For lexical complexity, a mean value and individual ratings by ten annotators are provided. The annotators were holders of JLPT levels N1 or N2, or their older equivalents 1 and 2. These two highest levels of JLPT are often the minimum requirement for employment or university acceptance (Japan Student Services Organization 2024) and correspond approximately to CEFR levels B2 and C1 (Japan Educational Exchanges and Services 2025). JLPT has been criticized for not testing productive skills (Nishizawa et al. 2022); however, that actually aligns with our task.

For LS, we provide an aggregated list of substitutions as well as individual substitutions from ten annotators. The annotators for LS were native Japanese speakers, with at least one year of experience teaching Japanese as a foreign language, instructed to provide simplifications for language learners without assuming a particular native language.

For lexical complexity, we target non-native speakers whose native language is neither Chinese nor Korean. As the two languages largely overlap in lexicon with Japanese, Japanese LS targeting Chinese or Korean (CK) L1 speakers should be evaluated separately. This was reflected in differences between CK and non-CK complexity ratings in JaLeCoN (Ide et al. 2023). The systematic difference in Japanese lexical complexity perception by Chinese L1 speakers was further confirmed by a reannotation and analysis of a subset of the present dataset, while no such difference was observed for a group of Czech and Slovak L1 speakers (Nohejl et al. 2024).

We split the dataset into a trial set (30 instances) and test set (570 instances), which share a similar word frequency, lexical complexity, and part of speech (POS) distribution (Nohejl et al. 2024) to make the trial set representative and suitable for training simple models or for in-context learning. As we explain in Section 3.6, the targets co-occur in the same sentence by groups of three. Therefore, the trial set comprises 10 sentences, while the test set comprises another 190 sentences.

The gold LS and LCP labels for the trial set were published on the open web for the MLSP shared task (Shardlow et al. 2024b). We therefore recommend using only the test set to evaluate models whose training data could have been contaminated. We made the full MultiLS-Japanese

dataset[5] and accompanying files (annotator profiles and annotation instructions)[6] available online. We request users not to disseminate the gold labels on the open web. The format of the dataset is described in Appendix B.

3.2 MultiLS Framework

The MultiLS framework (North et al. 2024) specifies that lexical simplifications and lexical complexity ratings are given for the same target words in the same contexts; hence, it is **multi-task**, whereas previous datasets handled only one of these tasks. Furthermore, MultiLS datasets for ten diverse languages are currently provided, all available in a common format via Hugging Face Hub as MLSP2024,[7] making MultiLS **multilingual**. The ten datasets are of comparable size and use the same annotation protocol (Shardlow et al. 2024a). Articles detailing the Portuguese data (North et al. 2024), Spanish and Catalan data (Saggion et al. 2024), and Italian data (Occhipinti 2024) have been published. In the following sections, we focus on the Japanese dataset, including its previously unpublished parts.

3.3 Further Goals and Additional Data

In addition to providing annotation for both LCP and LS, MultiLS-Japanese fulfills the following goals that are orthogonal to the MultiLS framework:

(1) varied sources representative of texts read by non-native speakers (Section 3.4),
(2) appropriate handling of Japanese morphology and MWEs (Section 3.5),
(3) sampling of target words representative of multiple factors of complexity (Section 3.6),
(4) minimization of external influences on perception of lexical complexity (Section 3.6.2),
(5) a single set of annotators who annotate each instance, enabling individual data analysis,
(6) a permissive license for the full dataset to facilitate further analysis or extension.

In addition to the data in the common MultiLS format, we also release individual annotator profiles, individual annotations, complete annotation instructions (see Section 3.7), detailed automatic and manual tags, and metadata for each dataset instance (see Appendix B).

3.4 Source Texts

A common approach to constructing LS and LCP datasets has been to use texts from a single source—e.g., Wikipedia (Horn et al. 2014, LexMTurk dataset) or newspaper (Kajiwara and

[5] https://huggingface.co/datasets/naist-nlp/multils-japanese
[6] https://github.com/naist-nlp/multils-japanese
[7] https://huggingface.co/datasets/MLSP2024/MLSP2024

Yamamoto 2015, SNOW E4 dataset)—or only a few genres, e.g., Wikipedia and news (Yimam et al. 2017, CWIG3G2 dataset). This may poorly represent texts read by the target population and severely bias the language towards certain registers or topics.

While Wikipedia is at least relatively diverse in terms of topics and offers favorable licensing terms, relying solely on Wikipedia would be problematic for Japanese, where a large part of the content is created through translation from English (Hautasaari and Ishida 2011), potentially leading to underrepresentation of local topics or culturally dependent vocabulary.

Instead, we strove to sample sentences from diverse genres and topics, emphasizing sources that Japanese non-native speakers may read. The sources constituting the final dataset are shown in Table 2. We carefully selected the sources such that the complete dataset could be distributed under the CC-BY-SA 4.0 International License,[8] which was the main limiting factor for source selection. All sources had to be released under a compatible Creative Commons license or to be in the public domain. Although we prioritized recent texts, we felt that the inclusion of literary fiction is important for representativeness and justifies a limited amount of slightly dated language. We have therefore included several works with expired copyrights published by well-known authors in the years 1945–1965.[9]

Category	Proportion	Source	License	Web Address	# Sentences
Wikipedia	50.0%	Wikipedia	CC	ja.wikipedia.org	100
Practical Information	21.0%	Government of Japan	CC	gov-online.go.jp	27
		Local Symbiotic Society	CC	mhlw.go.jp/kyouseisyakaiportal	3
		Consumer Affairs Agency	CC	caa.go.jp	3
		Hikikomori Voice Station	CC	hikikomori-voice-station.mhlw.go.jp	3
		Hakodate City	CC	city.hakodate.hokkaido.jp	3
		Hokkaido Prefecture	CC	pref.hokkaido.lg.jp	1
		Wikivoyage[a]	CC	ja.wikivoyage.org	2
Fiction	19.5%	Aozora Bunko[b]	PD	aozora.gr.jp	39
News	5.5%	Wikinews	CC	ja.wikinews.org	9
		Prime Minister's Office	CC	kantei.go.jp	2
Culture and History	4.0%	Japan Heritage	CC	japan-heritage.bunka.go.jp	8

Table 2 Sources of MultiLS-Japanese by category. The dataset is composed of a total of 200 sentences. Licenses: CC are Creative Commons licenses compatible with CC-BY-SA, PD is public domain. [a] Only pages about locations in Japan to avoid translated content. [b] Selected works of Dazai Osamu (†1948), Umezaki Haruo (†1965), Tanizaki Jun'ichirō (†1961), Yoshikawa Eiji (†1955), Sakaguchi Ango (†1955), Edogawa Ranpo (†1965), Yamamoto Shūgorō (†1967), and Yanagita Kunio (†1962) published 1945–1965.

[8] https://creativecommons.org/licenses/by-sa/4.0/

[9] Until 2019, the Japanese copyright law protected works for only 50 years after the author's death.

We considered but decided not to use Tatoeba,[10] which provides examples for language learners as opposed to naturally occurring sentences, and Japanese Wikiversity,[11] which was available only as a beta version when we were collecting data. We also could not use Japanese WikiHow[12] because its CC-BY-NC-SA license is incompatible with the CC-BY-SA license.

3.5 Target Words and Japanese Morphology

LS occurs at the level of words. However, this does not mean that we are merely substituting one lexeme for another: we are substituting word forms so that the sentence remains grammatical after the substitution. Some datasets make this task artificially simpler. For instance, the Portuguese LS dataset SIMPLEX-PB (Hartmann et al. 2018) contains substitutions only in lemma form (e.g., *coletaram*, past, 3rd pers. pl. → *recolher*, inf.), and the Japanese SNOW E4 (Kajiwara and Yamamoto 2015) contains only target words in uninflected form. The former approach results in ungrammatical simplified sentences, while the latter restricts the dataset to a small subset of language; therefore, we avoided both.

In the case of the Japanese language, we need to deal not only with inflections but also with segmentation. Multiple units can be considered a word for the purposes of LS. Japanese script does not separate words in writing, and in Japanese NLP and linguistics, sentences are commonly segmented into short unit words (SUWs), long unit words (LUWs), or *bunsetsu* (Omura et al. 2021). Table 3 illustrates that targets consisting of a single SUW or LUW would not be sufficient for LS, as verbs generally cannot be simplified by substituting a single SUW or LUW. While

Original	SUW	彼女 She	は TOP	読書 book reading		し do	て GER	いる be-PRS	。
	LUW	彼女	は	読書し			ている		。
	Bunsetsu	彼女は		読書して				いる	。
Simplified (substituted units in bold)	SUW	彼女 She	は TOP	**本** book	**を** ACC	**読ん** read	**で** GER	いる be-PRS	。
	LUW	彼女	は	**本**	**を**	**読ん**	**でいる**		。
	Bunsetsu	彼女は		**本を**		**読んで**		いる	。
Translation	She is reading a book.								

Table 3 Three common units of segmentation, SUW, LUW, and *bunsetsu*, when simplifying the verb in the sentence 彼女は読書している (She is reading a book). Dashed line marks an optional (traditional) *bunsetsu* division. Abbreviations in glosses: TOPic, ACCusative, GERund, PReSent.

[10] https://tatoeba.org
[11] https://ja.wikiversity.org
[12] https://www.wikihow.jp

a *bunsetsu* is long enough to support verb substitution, it would also include many dependent words that are unnecessary for simplification, e.g., particles with nouns. Therefore, we decided to target such sequences of SUWs that are long enough to substitute the given word. The resulting target sequences span at most one *bunsetsu*, except in cases where a longer MWE needs to be simplified as a single unit (e.g., 人目を引く "to draw attention"). For SUW segmentation, we used MeCab (Kudo et al. 2004), a Japanese morphological analyzer, and the UniDic 2.1.2 dictionary (Den et al. 2007), which is distributed as `unidic-lite`. The precise guidelines for delimiting the boundaries of target words and an overview of MWEs in the dataset are given in Appendix C.

Our approach slightly differs from both JaLeCoN and SNOW E4. The approach used for JaLeCoN (Ide et al. 2023) is mostly based on SUWs and manually selected MWEs, thus typically preferring smaller units. This does not pose practical problems for LCP, but as we have shown in Table 3, it is insufficient for substituting verb forms. CBDJLS (Kodaira et al. 2016), on the other hand, uses larger spans including all dependent words surrounding the target word on both sides (as discussed in Section 2.1), which generally extend over two *bunsetsu*. When inspecting the dataset, we found that dependent words on the left side of the target words were rarely altered by the substitutions. This was the expected result because dependent words in the Japanese language follow the corresponding independent words rather than precede them.

3.6 Selection of Target Words and Sentences

To better control and partially automate the sampling of suitable sentence contexts and target words, we adopted a scheme in which sentences with a single target word were first sampled automatically from the source texts. From this sample, we have manually selected 200 sentences to constitute the final dataset, manually picking two additional target words in each sentence for a total of 600 target words. We refer to the automatically tagged words as primary and to the manually selected ones as secondary.

Simplifying or predicting complexity for three words per sentence has the following advantages compared to the common format of LS datasets with a single target word per sentence:

(1) It reduces the amount of text to be read, enabling annotators to concentrate better.
(2) It reflects that simplifying a single word in a sentence is often insufficient.
(3) It allows us to analyze how lexical complexity is affected by other words in the sentence.

3.6.1 Automated pre-selection of words

With a relatively small dataset and many factors affecting lexical complexity, we faced the risk of not representing some of the factors if we sampled the words randomly. Therefore, the primary

target words were selected to represent diverse word frequencies, character compositions, and parts of speech. Additionally, the primary targets include specific categories of words known to be difficult for learners: compound verbs (Himeno 1999; National Institute for Japanese Language and Linguistics 2025), complex auxiliaries, complex particles (together covering most relatively advanced function words), and mimetics (Yoshioka 2016).

We partitioned candidate word properties in a discrete space with three axes: word frequency, character composition, and POS, as shown in Figure 1. For each element of the space (e.g., nouns written in *katakana* with $\log_{10}$ frequency from -4.5 to -4.0), we pre-selected candidate words and sentences containing them, and determined a number to be selected manually from them to achieve an approximately even distribution of target words in the space, given the number of words with each property.

The three groups of parts of speech are composed as follows:

- Nouns (90): nouns except proper nouns and verbal nouns (VN)+*suru* (i.e., *sahen* verbs).
- Verbs (50): VN+*suru* and independent verbs except verb+verb compound verbs.
- Other (60): verb+verb compound verbs (10); prefixes (5); suffixes (5); adjectives (10); adjectival nouns (10); non-mimetic adverbs (7); mimetic adverbs (3); particles, including complex (5); and auxiliaries, including complex ones (5).

The division into the three groups is only a technical means to ensure approximately even distribution of the two most important groups (Nouns and Verbs) across all combinations of word frequencies and character compositions. At the automatic pre-selection stage, the word candidates consist of a single SUW—except for VN+*suru* (e.g., 活躍する), complex particles (e.g., に

Word $\log_{10}$ freq.	#
[-7.5, -7.0)	29
[-7.0, -6.5)	28
[-6.5, -6.0)	29
[-6.0, -6.5)	28
[-5.5, -5.0)	29
[-5.0, -4.5)	28
[-4.5, -4.0)	29
Total	200

×

Characters	#
Katakana only	10
Hiragana, optionally *katakana*, no *kanji*	40
Least frequent *kanji* in:	
– top 101– 500 *kanji*	25
– top 501– 900 *kanji*	25
– top 901–1300 *kanji*	25
– top 1301–1700 *kanji*	25
– top 1701–2100 *kanji*	25
– top 2101–2500 *kanji*	25
Total	200

×

POS group	#
Nouns	90
Verbs	50
Other	60
Total	200

=

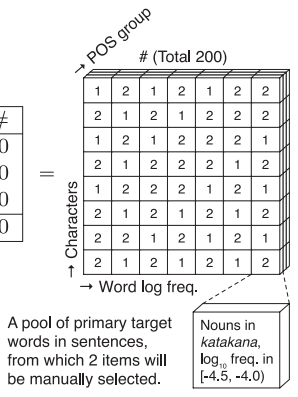


Figure 1 Selection of primary target words from a Cartesian product of word frequencies, character compositions, and parts of speech. # is the number of words with the given properties to be selected.

もかかわらず), and complex auxiliaries (e.g., ねばならぬ).[13] These compound word categories are an important part of Japanese vocabulary and include words difficult for learners. Without this special handling, it would be impossible to automatically select them based only on the SUW POS. On the other hand, the prefixes and suffixes are single SUWs, but we manually extended the selected targets to the whole words (e.g., 尽くめ → 黒尽くめ), except for cases of very productive, easy-to-substitute affixes (e.g., the temporal prefix 翌 or the honorific suffix 氏; see Appendix C.1). We used a custom version of the Japanese TUBELEX corpus (Nohejl et al. 2025) to determine word frequencies, including those of the aforementioned multi-SUW words.

We determined the order of the most frequent 2,500 *kanji* based on Japanese Wikipedia. We opted for Wikipedia and a limit (2,500) slightly higher than the requirements of JLPT N1 or the number of the *jōyō kanji*,[14] to enable the selection of relatively specialized or less frequent vocabulary if it satisfies the frequency criteria of our selection. See Limitations for further discussion of *kanji* and complexity.

3.6.2 Automated pre-selection of sentences

Other than containing the primary target word, the sentences automatically pre-selected from the source texts had to fulfill the following criteria to minimize external influences on the perception of lexical complexity: (1) The sentence is 16–32 SUWs and 24–60 characters long. (2) The sentence does not contain any *kanji* outside the top 2,500 most frequent *kanji* in Japanese Wikipedia. (3) The primary target word is the lowest frequency word in the sentence.

We restricted the sentence length to reduce the effect that a long, complicated sentence could have on the judgments about lexical complexity, and we used a range in which we expected both to find two more simplifiable words and to cover most of the high-complexity words for our target population by the three target words. The relatively short sentences also enabled the annotators to concentrate better on their task. We also avoided rare *kanji* or words rarer than the target words in order not to affect the lexical complexity ratings.

At the end of this stage, we had pools of candidate primary target words in sentences, each corresponding to a combination of word properties (frequency, character composition,[15] and the POS group) and coupled with the required number of items to be manually selected.

[13] The verb+verb compound verbs generally consist of a single SUW.

[14] As of 2025, there are 2,136 *jōyō kanji* "*kanji* for regular use", promulgated by the Japanese Ministry of Education, Culture, Sports, Science and Technology.

[15] The main classes of characters of the Japanese script are: *hiragana*, a basic syllabary; *katakana*, a syllabary used mostly for foreign-origin words from Western languages and mimetic words, and *kanji*, Chinese characters used mostly for content words. In particular, *kanji* present a cognitive and memorization challenge for learners.

3.6.3 Manual selection of sentences and words

The authors of the paper manually selected the required number of sentences from each pool, choosing two secondary target words from each and adjusting the extent of all target words to the appropriate morphological boundary as outlined in Section 3.5. Each selection was initially conducted by a native Japanese speaker and reviewed by a non-native Japanese speaker, or vice versa, and corrected if necessary until we completed the selection of 200 sentences with 1 primary and 2 secondary targets each, fulfilling the following criteria:

- A missing context does not complicate the comprehension of individual words or the sentence.
- Each target word is simplifiable and spans the appropriate SUW segments.
- Each target word in the dataset is unique.
- No single source contributes more than 50% to the entire dataset.

As the least frequent word in the sentence, we expect the primary target to be among the most difficult words in the sentence. Among the remaining words, we selected as secondary targets the two that we expected to be most difficult for our target population. While our judgment may not perfectly match our target population's aggregate perception, due to the relatively short length of the sentences, we expect the mismatches to be relatively rare and to affect sentences where multiple candidates of similar complexity are available for the third target word. Table 4 shows examples of the selected sentences and target words along with two instances of words of similar complexity that were not selected, i.e., cases where we may have not selected the most complex words as targets. The examples also illustrate target words of various parts of speech and character compositions.

3.7 Annotation and Review

Annotating lexical complexity is a relatively straightforward task, which relies on the subjective judgments of the annotators from the target population. On the other hand, the annotation of appropriate lexical substitutions poses a conundrum: annotators from the target group cannot annotate appropriate substitutions when the text is difficult for them to understand in the first place. We addressed this problem by working with two separate annotator groups.

We hired two groups of annotators through an intermediary company, which allowed us to recruit annotators fulfilling our requirements and to ensure quality control. The ten annotators performing lexical complexity annotation were non-native Japanese speakers who had JLPT levels N2 or N1 (or the older equivalents 2 and 1) and whose first language was neither Chinese nor Korean. The ten annotators performing LS annotation were native Japanese speakers with at

Source	Primary Target [POS, Characters]	Sentence with **Targets** and Other Candidates	Approximate Translation
W Wikipedia	ブッとばす (knock down) [VV, *kata.+hira.*]	「気に入らねえ奴はブッとばす」というのが口癖。	"I'll **knock down** the **fellas** I don't like" is his **motto**.
W Wikipedia	肯定して (affirmed) [VN+*suru*, *kanji*]	彼らがそれらのアプリを使うのは、友人から肯定してもらうことで自分の存在の正当性を確認しようとする試みなのである。	They use those apps in an **attempt** to prove the **validity** of their existence by getting it **affirmed** from their friends.
Consumer Affairs Agency	直ちに (immediately) [ADV, *kanji*]	旧基準の製品を使用する場合は、収納部分の扉のロックが壊れていたら、直ちに使用を中止してください。	When using a product of the **outdated** standard, stop using it **immediately** if the **clasp** on the storage door is broken.
Wikinews	パトカー (patrol car) [N, *katakana*]	パトカーは 600 メートルほど離れた路上でワゴン車を見失ったものの、その後タクシーと衝突したワゴン車を発見した。	Although the **patrol car** lost sight of the van on the road about 600 meters away, it later **discovered** the van after it had **collided** with a taxi.

Table 4 Examples of sentences and target words (bold) selected for the dataset. Other words of complexity comparable to the selected target words are underlined. Rather than being literal, the translations illustrate the complexity of the original words in English. "*Kata.+hira.*" stands for *katakana+hiragana*.

least one year of professional experience in teaching Japanese as a foreign language and were instructed to provide simplifications for language learners without assuming a particular native language. Therefore, in the case of LS, we were relying on the professional experience of the annotators and their ability to follow these instructions.

Each annotator received instructions for their task and a profile questionnaire. Non-native speakers were given instructions and questionnaires in Japanese appropriate for their proficiency level and in English. Appendix D provides the full instructions and questionnaires along with English translations for cases in which the original texts were only in Japanese. The corresponding files are in the project's online repository.

We asked both groups to annotate the first ten sentences, which constitute the trial data of the final dataset. We evaluated the initial annotation both manually and by computing agreement measures between individuals and the rest of the group. We did not find any outliers or errors that would prevent us from using the annotation or would require us to replace any annotators.

In the case of lexical simplifications, we gave each annotator individual feedback on their annotation, generally encouraging them to maintain the good quality of their work and pointing out areas for improvement in the annotation. We also supplied more examples to all annotators, illustrating how we expected the instructions to apply to the items they had already annotated

	Lexical Complexity (non-native speakers)	Lexical Simplification (native language teachers)
Native languages	English (5), Swedish (1), Portuguese & English (1), French & English (1), Basque & Spanish (1), French (1)	Japanese (10)
JLPT level	1 (3), N1 (3), N2 (3), 2 (1)	*(native)*
Studied Japanese at university	7 of 10	—
Currently lives in Japan	10 of 10	10 of 10
Lived in Japan (total years)	16.7 ± 8.3	—
Reading in Japanese (hours/week)	5.7 ± 7.6	15.9 ± 14.5
Age (years)	40.8 ± 9.1	54.1 ± 5.5
Education (total years)	18.4 ± 3.7	16.8 ± 2.8
Non-native languages	1.7 ± 0.5	1.4 ± 0.8
Teaching Japanese (hours/week)	—	4.4 ± 9.3
Experience teaching (total years)	—	2.6 ± 2.3

Table 5 Summary of the annotator profiles. The last seven rows report means and standard deviations separated by ±. Full questions in Appendices D.2 and D.4.

(provided in Figure 5 in Appendix D.3). In the case of one LS annotator, in whose annotation issues with simplicity or meaning preservation occurred repeatedly, we reviewed their work once more after they completed another ten sentences, to ensure that they understood the feedback.

We reviewed the annotation of the entire dataset automatically for formal errors (e.g., duplicate simplifications or spurious characters) and agreement with other annotators (we did not detect any outliers or suspiciously similar annotations). We also performed character normalization and manual corrections based on automated screening to correct unambiguous errors such as ungrammatical word forms, as detailed in Appendix E. Finally, we reviewed the dataset for offensive language (see Ethical Considerations for details).

Each annotator was given a unique identifier so that their individual annotations could be matched to their profiles. We did not collect any personally identifying information. In addition to aggregated data, we released all of the individual annotations for both tasks and the annotator profile data. Table 5 summarizes the annotator profiles.

4 Proposed Methods for LS and LCP

The MultiLS-Japanese dataset, presented in the previous section, was used for an evaluation

of a shared task (Shardlow et al. 2024b), in which systems representative of the state-of-the-art approaches participated. We therefore proposed new methods with the goal of improving both accuracy and efficiency over the shared task results, not only on MultiLS-Japanese but across all languages with available MultiLS datasets.

The current state-of-the-art methods for LS and LCP require multiple LLM inferences to simplify a single word or predict its complexity. In contrast, we propose methods for both subtasks that use only a single LLM inference per data instance. Our methods use the same prompt across all languages, and for LCP we employ a novel calibrated token-probability-based scoring technique, G-Scale.

4.1 Lexical Simplification

In LS, multiple substitutions are generated for target words in a sentence context. The ordered set of substitutions is then evaluated against an ordered set of gold substitutions. In the general case, the goal is to match both the gold substitutions and their order.

The system UniHD (Aumiller and Gertz 2022), which won the TSAR 2022 shared task, ensembled six combinations of prompts and temperatures, combining their outputs using rank average. In all cases, the temperature was non-zero. As motivations for this approach, the authors cite safeguarding against malformed outputs (the chance of multiple outputs being discarded due to automatic filtering is lower) and a higher diversity of substitutions. The system TMU-HIT (Enomoto et al. 2024) used an approach derived from UniHD but ensembled ten instances of the same prompt and temperature.

We simplified the approach by using a single prompt and greedy decoding (by setting temperature to zero), as we hypothesized that the safeguarding against malformed outputs (Aumiller and Gertz 2022) was no longer necessary on more recent LLMs and we questioned whether sampling multiple responses to the same prompt (Enomoto et al. 2024) provides any advantage over greedy decoding. Our prompt (see Appendix F.1) is based on the three conditions that lexical simplifications must fulfill (i.e., grammaticality, meaning preservation, and simplicity). Additionally, we specifically mention simplicity for "language learners," which are targeted by many of the MultiLS datasets.

Note that our method does not involve any reranking: it simply outputs simplifications in the order in which they were generated by the LLM. The effectiveness of this approach was previously demonstrated in the TMU-HIT system (Enomoto et al. 2024), where it outperformed two reranking approaches. While we did not experiment with reranking, we compared ranking based on a single LLM output (our method) with ordering by mean rank from multiple outputs

(TMU-HIT) and ordering by median rank, as discussed in Section 5.1.

We conducted a few preliminary experiments with different prompt wording on the English and Japanese trial sets: We observed that using complex words such as "synonym" or "substitute" in the prompt led to more complex words appearing near the top of the generated simplification lists, and therefore, we attempted to limit them. Additionally, we saw an improvement in the quality and consistency of formatting on Japanese after including the identifiers S, X, and Y for the sentence, target word, and simplifications, respectively.

4.2 Lexical Complexity Prediction

In LCP, a complexity score is predicted for target words in a sentence context. The gold complexity score is an average of ratings by annotators from the target population. Therefore, although the annotators rate complexity on a discrete scale, the prediction methods must predict values on a continuous scale.

We propose a new method, G-Scale, applicable to any tasks of rating a text input on a continuous scale. The method is motivated by the following observations about existing LCP methods based on LLM prompting (Enomoto et al. 2024; Smădu et al. 2024):

(1) **Low efficiency:** The current state-of-the-art LLM-based LCP methods are extremely computationally intensive, using 20 inferences (Enomoto et al. 2024; Smădu et al. 2024) per instance to obtain results competitive with simpler models.

(2) **Lack of calibration:** As the LLM-based LCP methods are based on weighting the scores by token probabilities, they depend on two conditions: (1) that the probabilities provide valuable information and (2) that they have the correct magnitude. Current methods do not perform any calibration, taking the probabilities at face value.

(3) **Bad fit:** The LLM-based TMU-HIT (Enomoto et al. 2024) achieved a better correlation but a worse fit, i.e., a larger error, than a feature-based model.

Although we were motivated by the weaknesses of existing LCP methods, we formulated our method so that it is potentially applicable to other tasks where a text input is rated on a scale, and the model is expected to accurately predict human ratings, not only to correlate with them, e.g., automated essay scoring. We devised a scoring function and a training algorithm for G-scale by addressing the problems mentioned above:

Efficient use of token probabilities. The existing LLM-based LCP methods (Enomoto et al. 2024; Smădu et al. 2024) were derived from G-Eval (Liu et al. 2023), which was proposed as a method for evaluating natural language generation (NLG) on a numerical scale via LLM prompting. G-Eval's authors proposed weighting discrete scores by their token probabilities

or, when the probabilities are not available, estimating the probabilities by sampling multiple responses. Only the latter option was used in LCP (Enomoto et al. 2024; Smădu et al. 2024). As token probabilities are now available even for the latest OpenAI models as `logprobs` in the API,[16] we used probabilities, reducing the computation cost by a factor of 20. Furthermore, we predict only a single token, avoiding the generation of superfluous explanations.

Calibrated temperature scaling. G-EVAL and the two LCP methods use the same weighting scheme, where token probabilities (or probability estimates) are used as weights for the scores:

$$\text{score}_{\text{G-EVAL}} = \sum_{i=1}^{n} p(s_i) \cdot s_i \tag{1}$$

where $s_i \in S$ are discrete scores from a scale S (e.g., 1, 2, 3, 4, and 5), which the model is instructed to use in the prompt. Consequently, the distribution of $\text{score}_{\text{G-EVAL}}$ depends on the token probability distribution. The distribution may be too spiky, i.e., assigning almost all probability to the top token, or too noisy, i.e., assigning the remainder of the probability randomly. The two LCP methods represent scores by one to two tokens (one token: 0, 1; two tokens: 0.25, 0.5, 0.75), which may further complicate the interpretation of the probabilities, and use the temperature values 0.7 (Enomoto et al. 2024) and 0.8 (Smădu et al. 2024) to sample tokens and estimate probabilities from them without explaining how the temperature values were determined.

We use single-token scores as in the original G-EVAL and replace $p(s_i)$ in its formula with temperature-scaled softmax:

$$\text{score}^{(T)}_{\text{softmax}} = \sum_{i=1}^{n} \frac{\exp\left(\frac{\log p(s_i)}{T}\right)}{\sum_{j=1}^{n} \exp\left(\frac{\log p(s_j)}{T}\right)} \cdot s_i \tag{2}$$

Note that in contrast with the previous methods, we do not use the temperature T for decoding an output consisting of multiple tokens but for computing a score from a single token distribution using the formula above. We calibrate the value of T by a search in a fixed set of values using a training set: First, we perform LLM inference using a prompt once for each instance of the training set, recording the token probability distributions. Then, we select a value of T that maximizes the linear correlation between the scores predicted by (2) and the gold scores of the training set.

[16] `https://platform.openai.com/docs/api-reference/completions/create`

Fitting a linear regression. To address the problem of a relatively large error despite a good correlation, we fit a linear regression model with $x_1 = \text{score}^{(T)}_{\text{softmax}}$ and optionally other features $(x_i)_2^k$ as independent variables. Because this can result in predicted values being out of the scale, we clip the result to the interval $[s_1, s_n]$, assuming that s_1 and s_n are the lowest and highest discrete scores, respectively, of the scale S. This yields the final formula of the G-SCALE's scoring function:

$$\text{score}^{(T,\boldsymbol{\beta})}_{\text{G-SCALE}} = \max\left(s_1, \min\left(s_n, \beta_0 + \beta_1 \cdot \text{score}^{(T)}_{\text{softmax}} + \sum_{i=2}^{k} \beta_i \cdot x_i\right)\right) \tag{3}$$

Linear regression also makes it possible to ensemble predictions from multiple prompts or models or to add features, e.g., word frequency in the case of LCP, for a hybrid approach combining feature-based modeling and LLM prompting. Depending on the additional features and the training set size, a more powerful model (e.g., support vector machine or random forests) can be used instead of linear regression.

Training steps. The parameters T and $\boldsymbol{\beta}$ are trained using the following steps:

(1) Perform LLM inference using a prompt once for each instance of the training set. Record the token probabilities for each instance of the training set.

(2) Select a temperature T that maximizes the linear correlation of $\text{score}^{(T)}_{\text{softmax}}$ (computed from the token probabilities using (2)) with the gold scores of the training set.

(3) Fit a linear regression model $\boldsymbol{\beta}^\mathsf{T}\mathbf{x}$ on the training set. (The clipping in (3) is applied only at inference time.)

The temperature scaling and linear regression, which differentiate our method from G-EVAL, are not specific to LCP and can be applied to any tasks where the fit of continuous values is important. G-EVAL (Liu et al. 2023) was originally proposed for ranking alternative LLM outputs in NLG tasks, where linear correlation and fit are not important, as long as the score has a good rank correlation with human judgment.

Prompt for LCP. We employ a few-shot prompt with simple instructions (see Appendix F.3) and the complexity scale description used for the MultiLS English dataset (Shardlow et al. 2024a). Similar to our LS prompt, we try to avoid difficult words in the prompt itself (e.g., "complexity"), and explicitly target language learners. We instruct the model to rate the difficulty for an "intermediate language learner" to obtain consistent complexity ratings. The prompt is not based on G-EVAL prompt methodology (Liu et al. 2023) and does not elicit chain-of-thought output. We did not perform preliminary experiments with multiple prompts; instead, we performed an ablation of the prompt, as detailed in Section 5.1.

G-Eval involves a specific methodology for creating a prompt. The fixed step-by-step instructions in the prompt are referred to as "chain-of-thoughts" (CoT) by G-Eval's authors (Liu et al. 2023). This is an unusual usage of the term, as "CoT prompting" has been used to refer to the model *generating* a description of thoughts leading to a conclusion in response to the prompt (Wei et al. 2022, also cited by Liu et al.). In the context of LCP methods based on G-Eval, "CoT" was implemented both as G-Eval-style fixed instructions (Enomoto et al. 2024) and as per-instance thought process generation (Smădu et al. 2024). Both result in an increased computational cost: the former involves a longer prompt (see Appendix F.4) compared to ours and the latter results only in a small improvement (Smădu et al. 2024) while incurring a longer generated output compared to a single token in our case.

5 Experiments

For the experiments, we used GPT-4o[17] (OpenAI 2024b), a paid closed-weight model, with the following parameters: `frequency_penalty=0.5` and `presence_penalty=0.3`. For our methods, we used `temperature=0`, and when performing stochastic inference, we used `temperature=0.7`. For LCP, we additionally used `logprobs=True` and `top_logprobs=20` to obtain token log-probabilities, which are necessary for G-Scale and G-Eval. For a fair comparison, the choice of penalties and temperature for stochastic inference followed previous methods (Aumiller and Gertz 2022; Enomoto et al. 2024), which also used OpenAI's GPT models.

For LCP using G-Scale, we selected the temperature $T = 2.0$, which achieved the highest mean correlation on the trial sets of all languages from values (0.1, 0.2, ..., 1.9, 2.0). The parameters $\boldsymbol{\beta}$ were fitted separately using the trial set for each language using linear regression with L2 regularization ($\alpha = 1$).

Unless specified otherwise, we used the prompt templates described in Appendices F.1 and F.3 for LS and LCP, respectively. For LCP, we used 3-shot examples sampled from the trial data for each language such that their scores are as close as possible to 1, 5, and 3 (the lowest point, highest point, and midpoint of the scale, respectively), and we rounded their scores to the nearest integer. The same three examples for each language were used at all times.

We performed evaluation on the MultiLS datasets for ten languages (including MultiLS-Japanese). We reported results in each language and the mean of the results computed across languages. The best results for each language and for the mean are printed in **bold**.

[17] `gpt-4o-2024-08-06`

5.1 Lexical Simplification

We compared the performance of our system with the top MLSP shared task submissions (Shardlow et al. 2024b), namely ISEP (Dutilleul et al. 2024), GMU (Goswami et al. 2024), and TMU-HIT (Enomoto et al. 2024) using accuracy@1@top1 and accuracy in Tables 6 and 7, respectively. Accuracy@1@top1, which was the main evaluation metric of the shared task, strictly evaluates the top generated simplification against the top simplification given by human annotators, whereas accuracy evaluates the top generated simplification against all simplifications given by annotators, counting less frequent simplifications as correct too. To enable comparison with the top shared task submission, we used the shared task's evaluation script to compute the metrics. As a result, the maximum achievable accuracy and accuracy@1@top1 for languages with unsimplifiable instances was lower than 1 (Shardlow et al. 2024b, Sec. 4.2, App. A). For Filipino, the maximum accuracy was 0.772; for other languages, it was in the interval [0.997, 1].

Our system outperforms the top shared task submissions on most languages, achieving the best performance on 7 out of 10 languages in both accuracy@1@top1 and accuracy, thereby setting a new state-of-the-art result on the MultiLS datasets.

The top shared task submission, TMU-HIT used GPT-4, whereas we used GPT-4o, a newer model by OpenAI, which may be more effective in this task, although its API usage is cheaper.

System	Catalan	English	Filipino	French	German	Italian	Japanese	Portug.	Sinhala	Spanish	Mean
ISEP	0.272	0.468	—	0.374	0.219	0.424	—	**0.485**	—	—	—
GMU	0.225	0.516	0.056	0.366	0.420	0.404	0.258	—	0.228	0.418	—
TMU-HIT	0.258	**0.524**	0.065	0.426	0.488	0.476	0.400	0.443	0.221	**0.454**	0.376
Ours	**0.283**	0.493	**0.084**	**0.445**	**0.544**	**0.554**	**0.416**	0.467	**0.257**	0.421	**0.396**

Table 6 Accuracy@1@top1 of LS on the MultiLS datasets compared with the top MLSP shared task submission results (ISEP, GMU, TMU-HIT) reported in the shared task findings (Shardlow et al. 2024b).

System	Catalan	English	Filipino	French	German	Italian	Japanese	Portug.	Sinhala	Spanish	Mean
ISEP	0.679	0.765	—	0.831	0.302	0.744	—	0.633	—	—	—
GMU	0.535	0.749	0.293	0.741	0.641	0.735	0.541	—	0.311	0.801	—
TMU-HIT	0.661	**0.798**	0.365	0.856	0.742	0.803	**0.723**	0.601	0.320	**0.847**	0.671
Ours	**0.681**	0.770	**0.414**	**0.880**	**0.809**	**0.842**	0.721	**0.642**	**0.412**	0.821	**0.699**

Table 7 Accuracy (equal to potential@1 and MAP@1) of LS on the MultiLS datasets compared with the top MLSP shared task submission results (ISEP, GMU, TMU-HIT) reported in the shared task findings (Shardlow et al. 2024b).

Because our system and TMU-HIT use different underlying models, prompts, and inference methods, we ran another set of experiments, where we used the same model (GPT-4o) and the following prompt/inference method combinations to perform further analysis:

- **Ours / Median Rank:** Using our prompt, we repeat the inference 10 times with temperature 0.3, similarly to TMU-HIT, but instead of ensembling the simplifications using the mean rank, we ensemble them using the median rank.
- **Ours / TMU-HIT:** Using our prompt, we repeat inference 10 times with temperature 0.3, and ensemble the simplifications using the mean rank, exactly as TMU-HIT did.
- **TMU-HIT / Ours:** Using the TMU-HIT prompt,[18] we greedily decode output, as outlined in Section 4.1.
- **Ours / Ours:** Using our prompt, we greedily decode output.

To evaluate the results, as shown in Table 8, we used the same 10 stochastic inferences for the first two rows of the table, and a single greedy inference for each of the last two rows to ensure comparability. We computed p-values to indicate whether the difference from the best performing system in each column is statistically significant. To this end, we used the paired permutation test with 20,000 resamples of the dataset instances.

We can observe that using the TMU-HIT prompt instead of ours consistently resulted in poorer performance. The system using the TMU-HIT prompt achieved significantly lower accuracy@1@top1 in Japanese and significantly lower accuracy in Japanese and Spanish.

None of the other results differ significantly. However, an important difference lies in the number of inferences used: Ensembling 10 stochastic inferences with the same prompt using mean

Prompt / Infer. Method	#Inf.	Accuracy@1@top1				Accuracy			
		English	Japanese	Spanish	Mean	English	Japanese	Spanish	Mean
Ours / Median Rank	10	0.526	**0.428**	**0.491**	0.482	0.765	**0.726**	**0.835**	**0.775**
Ours / TMU-HIT	10	**0.537**	0.424	0.486	**0.482**	**0.775**	0.719	0.831	**0.775**
TMU-HIT / Ours	1	0.531	0.360***	0.462	0.451	0.770	0.665**	0.792*	0.743
Ours / Ours	1	0.535	0.417	0.477	0.477	0.770	0.721	0.821	0.771

Table 8 Number of inferences per instance (#Inf.), accuracy@1@top1 and accuracy of LS on the MultiLS English, Japanese, and Spanish datasets by prompt and inference method. Significance levels: $p < 0.001$ (***), $p < 0.01$ (**), $p < 0.05$ (*).

[18] We use the TMU-HIT prompt for languages other than Japanese (see Appendix F.2). We do not evaluate TMU-HIT's method for the Japanese language, which requires specific postprocessing.

rank (TMU-HIT) or median rank yields results without any significant difference ($p \geqslant 0.05$) from performing a single greedy inference (our method).

In a manual analysis of the outputs, we observed a non-intuitive behavior of averaging ranks using the arithmetic mean: an item that has the top rank in the majority of cases may not obtain the top mean rank if it is weighed down by low ranks (or omission of the item) in the other cases. If ranks are averaged using the median instead, an item that has the top rank in the majority of cases will always get the top median rank. The effect on accuracy metrics, however, was mixed—while median rank performed better for Japanese and Spanish, it performed worse for English—and in none of the cases was the effect significant.

We can conclude that our prompt outperformed the prompt used by TMU-HIT (Enomoto et al. 2024), and more importantly, our inference method achieved essentially the same results at one tenth of the computational cost of TMU-HIT's inference method.

5.1.1 Error analysis on MultiLS-Japanese

On the MultiLS-Japanese dataset, our method achieved an accuracy of 0.721. In other words, in 72.1% of the test set instances, the top predicted simplification was in the gold data. We manually analyzed the errors in the remaining 27.9% (159 instances), categorizing them by issue in Table 9 (examples in Appendix I). In 21% of the errors, we found the simplifications valid, although they did not match the gold data. For the actually invalid simplifications, most errors stemmed from the meaning of the sentence not being preserved.

Validity	Proportion	Issue	Proportion	# Instances
Invalid	79%	Not preserving meaning	21%	81
		Not a simplification	13%	21
		Ungrammatical	5%	8
		Not preserving meaning & not a simplification	5%	8
		Not preserving meaning & ungrammatical	4%	6
		Ungrammatical & not a simplification	1%	1
Valid	21%	Simplification not listed by annotators	13%	20
		Orthography not listed by annotators	4%	7
		Superfluous reading in parentheses attached	4%	7
Total Errors in Terms of Accuracy			100%	159

Table 9 Analysis of LS errors in terms of accuracy (instances where the top system simplification is not among gold simplifications by human annotators) of our system on MultiLS-Japanese.

While the amount of meaning lost or changed varies, the inability to consistently preserve meaning would be a serious problem for a real-world application: a system making grammatical errors or occasionally producing relatively complex outputs may still be usable, but a system that does not preserve meaning would not be fit for deployment. We therefore believe that meaning preservation should be prioritized in future LS research. The general formulation of LS as a task, which does not involve deletion or summarization, makes it well-suited for settings where meaning preservation is crucial.

5.2 Lexical Complexity Prediction

In Table 10, we compare the performance of our system with the top MLSP shared task (Shardlow et al. 2024b) submissions Archaeology (Cristea and Nisioi 2024), GMU (Goswami et al. 2024), TMU-HIT (Enomoto et al. 2024), and the shared task's frequency baseline using the coefficient of determination R^2. We use R^2, which is computed as a linear function of the mean squared error,[19] as it is the strictest evaluation metric. A more lenient evaluation using Pearson's and Spearman's correlation coefficients can be found in Appendix G. Our method, G-Scale, outperformed all shared task submissions in 9 of 10 languages, and in the mean over languages, it outperformed them by a wide margin. It also outperformed all ablated versions of our scoring

System	Catalan	English	Filipino	German	French	Italian	Japanese	Portug.	Sinhala	Spanish	Mean
Shared Task Top Results:											
Freq. Baseline	−0.370	0.547	0.004	0.073	0.146	0.227	0.340	**0.489**	−0.287	0.256	0.142
Archaeology	**0.011**	0.439	−0.076	0.069	0.214	−0.417	−0.098	0.242	−0.459	0.251	0.017
GMU	−0.338	0.525	−0.046	−0.528	0.048	0.077	0.024	—	−0.037	−0.073	−0.039
TMU-HIT	−0.161	0.515	−0.354	−0.558	0.270	0.242	0.413	0.153	−1.603	**0.494**	−0.059
Ablation of the Scoring Function:											
Direct	−1.237	0.417	−2.125	−1.003	0.493	−0.669	0.370	−0.343	−3.280	0.032	−0.734
Direct + LR	−0.056	0.532	0.040	0.127	0.445	0.273	0.457	0.405	−0.024	0.393	0.259
G-Eval	−0.894	0.536	−1.790	−0.726	0.558	−0.411	0.475	−0.136	−2.743	0.141	−0.499
G-Eval + LR	−0.056	0.561	0.070	0.136	0.458	0.298	0.483	0.437	0.037	0.417	0.284
G-Scale − LR	−0.656	**0.591**	−1.545	−0.616	**0.586**	−0.236	**0.536**	0.023	−2.395	0.205	−0.351
Ours: G-Scale	−0.060	0.566	**0.078**	**0.146**	0.460	**0.312**	0.500	0.461	**0.043**	0.429	**0.293**

Table 10 R^2 of lexical complexity prediction on the MultiLS datasets, compared with the top MLSP shared task submission and frequency baseline results reported in the shared task findings (Shardlow et al. 2024b), and ablated versions of the G-Scale scoring function. In addition to printing the best result in **bold**, we underlined the second best result.

[19] This is consistent with the shared task evaluation (Shardlow et al. 2024b) and the `r2_score` implementation in scikit-learn (`https://scikit-learn.org/`) but different from R^2 defined merely as a square of the Pearson's correlation coefficient.

function (also shown in Table 10). When evaluated using correlation coefficients (in Appendix G), the differences between methods were smaller, because errors in mean and variance (and linear correlation, in the case of the Spearman's rank coefficient) were not penalized. Yet, our method still achieved better mean correlation than the top shared task systems and ablations.

The ablations of the scoring function were:

- **Direct:** We predicted the score represented by the highest probability token (greedy decoding of a single token)
- **Direct + LR:** We used the value represented by the highest probability token as an independent variable for linear regression.
- **G-Eval**: We used the token probabilities to compute $\text{score}_{\text{G-Eval}}$ ((1)).
- **G-Eval + LR:** We used the value of $\text{score}_{\text{G-Eval}}$ as an independent variable for linear regression.
- **G-Scale − LR:** We calibrated T and use the token probabilities to compute $\text{score}^{(T)}_{\text{softmax}}$ ((2)).

For fairness of comparison, we always clipped the values predicted by linear regression to the range [0, 1], in the same way it is done for G-Scale.

In Table 11, we ablated the prompt and compared it with a prompt adapted to our method from the top MLSP shared task submission, TMU-HIT. To save computational resources, we evaluate only on three languages (English, Japanese, and Spanish). Our prompt achieved the best mean performance, the best performance in English and Japanese, and the second best performance in Spanish. This also demonstrates that assigning a concrete role in the prompt, i.e., "You are an intermediate learner of {*language*}", was beneficial even if it did not match

Prompt	English	Japanese	Spanish	Mean
Ours	**0.566**	**0.500**	0.429	**0.498**
– No language	0.558	0.482	**0.452**	0.497
– No language, generic role	0.557	0.486	0.424	0.489
– No language, no role	0.418	0.335	0.333	0.362
– No role	0.395	0.373	0.311	0.360
TMU-HIT-like	0.458	0.378	0.316	0.384

Table 11 R^2 of lexical complexity prediction on the MultiLS English, Japanese, and Spanish datasets, compared with the prompt ablations (described in Appendix F.1) and the prompt used by TMU-HIT with a minor modification (see Appendix F.1). Our scoring function (G-Scale) is used in all cases.

the actual target population of the datasets: Japanese was annotated by rather advanced non-native speakers, English by 10 natives and 11 non-native university students, and Spanish was annotated mostly by native speakers. With G-SCALE, it was not important whether the scale of scores output by the LLM was shifted, as long as it was consistent. The concrete role seems to have helped the LLM achieve this consistency. Additionally, specifying the language at the end of the prompt, i.e., "assign a difficulty rating to the {*language*} word" was beneficial too, except for Spanish, where it slightly hurt the performance. Note that we used the same 3-shot examples in all the compared prompts.

We aimed to further improve the performance on English, Japanese, and Spanish by adding corpus-based features to G-SCALE, namely log-frequency and log-range. As shown in Table 12, the best mean performance is achieved by adding log-range from the TUBELEX corpus (Nohejl et al. 2025). Range is a lexical dispersion measure, whose logarithm correlates particularly well with lexical complexity in the MultiLS datasets and psycholinguistic variables (Nohejl and Watanabe 2025). The logarithm of word frequency is the most common feature used for lexical complexity prediction. Both log-frequency and log-range alone (upper half of the table) provided very competitive predictions compared to G-SCALE based on LLM prompting, but the best performance was achieved by their combination. This shows that the LLM predictions and log-frequency (or log-range) were relatively uncorrelated, complementing each other in LCP. In Appendix H we evaluated the addition of log-frequency to all ten languages. For three of them, it hurt the performance, which we assume reflects the low quality of the respective corpora.

Features	English	Japanese	Spanish	Mean
Linear Regression without LLM Prompting:				
wordfreq Log-Frequency	0.547	0.359	0.256	0.387
TUBELEX Log-Frequency	*0.567*	0.409	0.327	0.434
TUBELEX Log-Range	*0.601*	0.406	0.362	0.456
G-SCALE (with LLM Prompting):				
LLM	0.566	0.500	0.429	0.498
LLM + wordfreq Log-Frequency	0.695	0.563	0.468	0.575
LLM + TUBELEX Log-Frequency	0.681	**0.580**	0.489	0.583
LLM + TUBELEX Log-Range	**0.696**	0.576	**0.502**	**0.591**

Table 12 Corpus-based features: R^2 of lexical complexity prediction on the MultiLS English, Japanese, and Spanish datasets. We use frequency from wordfreq and TUBELEX, and range (number of YouTube channels in which a word occurs) from TUBELEX. Features that outperform LLM-only prediction are *italicized*.

6 Conclusion

We built the first LS dataset targeting non-native Japanese speakers, MultiLS-Japanese. Among Japanese LS datasets, the proposed dataset has the unique feature of providing gold labels not only for LS but also for the subtask of LCP. MultiLS-Japanese is based on diverse source texts, has a permissive license (CC-BY-SA), and is part of the larger family of ten language-specific MultiLS datasets.

We proposed and evaluated methods for LS and LCP that are applicable to any language supported by an LLM. In the case of LCP, our method can be calibrated on a small number of examples (we used the trial set of 30 examples from each language-specific MultiLS dataset) and it can leverage corpus-based features in addition to LLM prompting. The computational cost of both methods is an order of magnitude smaller than that of the previous state-of-the-art LLM-based approaches.

For LS, we used a simple prompting approach, which nonetheless outperformed the state-of-the-art results for most languages. Our approach consists of a prompt that led to this performance improvement and an inference method that reduces the computational cost to one tenth of the previous state of the art.

We proposed G-Scale, a novel method for rating text input on a continuous scale using LLMs. G-Scale outperformed the previous state-of-the-art LLM-based LCP method (Enomoto et al. 2024), which is based on G-Eval (Liu et al. 2023), as well as a previous feature-based approach (Cristea and Nisioi 2024). G-Scale can efficiently combine other features with LLM prompting and leverage training examples while keeping the computational cost at a single LLM inference per data instance. Through ablation, we demonstrated that both differences of our method from G-Eval result in improved performance on LCP. We plan to apply G-Scale to other rating tasks in future work.

We believe that together with the Japanese LS and LCP dataset, our LS and LCP methods, which are widely applicable across languages, form a solid basis for future research into language-specific or personalized methods.

Limitations

MultiLS-Japanese was carefully constructed as an evaluation dataset. Its main limitation, shared with the other currently available MultiLS datasets, is that it provides only 30 training examples in the trial set. We attempted to provide resources for the evaluation of personalization,

with lexical complexity ratings and demographic profiles for ten individuals. The lack of data for personalization is a weak point shared by all LS and LCP datasets and an important area for future research.

MultiLS-Japanese targets non-native speakers of Japanese, holders of JLPT levels N2 or N1 (or the older equivalents 2 and 1) whose first language is not Chinese or Korean. While lexical complexity annotation was performed by individuals from this population, lexical substitutions were provided by professionals: as explained in Section 3.7, the individuals from the target population would not be able to perform the simplification of words that are difficult to understand for them. We strove to find professionals capable of performing the task and to provide them with appropriate instructions. Nonetheless, our approach does not guarantee optimal substitutions for the target population. We suggest exploring annotation methods that would involve feedback from individuals in the target population for future research.

Uncommon *kanji*, namely characters outside the top 2,500 *kanji* in Japanese Wikipedia, do not occur in the MultiLS-Japanese dataset. In Japanese Wikipedia, which has a relatively high proportion of both proper names and specialized vocabulary, such characters are used to spell only 0.32% of the words in running text.[20] Similarly, words with frequency less than $10^{-7.5}$ in TUBELEX, corresponding to 0.24% tokens in TUBELEX, were excluded.[21] Given the limited size of our dataset (600 target words), we preferred to represent words that are difficult for the target population, yet relatively frequent. This may be limiting in the future as LS systems improve: it will become desirable for evaluation datasets to focus more on lower-frequency words, which are potentially also more difficult to simplify. As for lower frequency *kanji*, a distinction can be made between lexical complexity proper and complexity stemming from rare *kanji*, which is commonly handled in Japanese by adding *furigana* (a phonetic guide in the form of ruby characters, e.g., the small characters かいらい in 傀儡(かいらい)政権, *a puppet regime*). MultiLS-Japanese does not contain *furigana*; we have not specifically tested annotators for *kanji* proficiency or experimented with the effect of adding *furigana* on perceived complexity or comprehension.

In both tasks, we used the same method across datasets for all ten languages, avoiding prompt or feature engineering for individual datasets. We did this to demonstrate the general applicability of our methods; however, at the same time, it is a limitation of our approach. We expect that

[20] The top 50 of the excluded words by frequency include Japanese and Chinese proper names, e.g., 齋藤 (*Saitō*), 馮 (*Féng*); words that have a more common alternative spelling, e.g., 蛋白質 (*protein*), commonly spelled タンパク質; and words regularly spelled using the uncommon characters, e.g., 嫉妬 (*jealousy*), 弾劾 (*impeachment*).

[21] Many of such words are proper nouns or English words in *katakana* and the alphabet. Simplifiable Japanese words in the top 50 excluded words by frequency include 多党 (*multiparty*) or 綬章 (*a medal of honor with a ribbon*).

further performance gains can be achieved by considering language specifics, target population, or the text genre, and adjusting the prompts and features used by our methods accordingly.

We evaluated G-SCALE, our method for rating text input on a continuous scale using LLMs, only on LCP. Its performance on other tasks is yet to be investigated. Both of our methods were devised to be applicable to closed-weight LLMs available only via API, such as OpenAI's GPT model family, relying only on prompting and access to token probabilities. In principle, our methods are applicable to any LLM, but we have not evaluated them on other models.

Ethical Considerations

We manually reviewed the content and language of the MultiLS-Japanese dataset, both original sentences and their simplifications, and found no ethically problematic content. Nevertheless, some instances of colloquial language may be considered inappropriate depending on the context. Discretion is therefore advised when using the dataset for purposes beyond its intended use (evaluating LCP and LS systems), for instance, in a classroom setting.

The following target words occurring in the dataset are informal and may be inappropriate in certain contexts, though none of them is particularly offensive: 奴 (*fella*, see Table 4 for context), ブッとばす (*send flying*, *knock down*, also in Table 4), おまえ (*you*, informal), 愚かな (*silly, stupid*), ケチる (*to be stingy, to skimp*, informal), キレさせる (*make* sb. *lose their temper*, *make* sb. *snap*, informal). Other informal words appear in the sentence contexts and sometimes in simplifications. Such words are also listed in existing pedagogical vocabulary resources: やつ, おまえ, 愚か, けち (Sunakawa et al. 2012), and ぶっ飛ばす (National Institute for Japanese Language and Linguistics 2025) are variants of the words above. While not typical examples of textbook vocabulary, we believe that such words are an important part of the complex lexicon that LS for non-native speakers should be tested against.

Text simplification, including LS, involves a trade-off between simplicity and meaning preservation (Agrawal and Carpuat 2024). We have highlighted meaning preservation as a priority for future research in Section 5.1.1. However, several issues beyond the preservation of factual meaning are worth noting. Compared to translation, it is more difficult for simplification to preserve the register or negative and positive nuances. For example, in our dataset, the simplifications of the formal and neutral word 犯人 (*criminal, culprit*) include 悪い人 (*bad person*) and 悪い奴 (*bad guy*, informal), which could be perceived as more negative, and simplifications of ケチる (*to skimp*, informal), which carries a negative connotation, include 節約する (*to save*, *to economize*), introducing a more positive tone. Similar issues were identified in the Catalan MultiLS dataset

(Saggion et al. 2024). We therefore recommend that text simplification systems, particularly when intended to assist vulnerable individuals, are deployed only with human oversight.

Acknowledgement

The MultiLS-Japanese dataset was created as part of a joint research project between Nara Institute of Science and Technology and Nikkei, Inc. We are grateful to the anonymous reviewers for their insightful comments and suggestions.

References

Agrawal, S. and Carpuat, M. (2024). "Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension." *Transactions of the Association for Computational Linguistics*, **12**, pp. 432–448.

Aumiller, D. and Gertz, M. (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)*, pp. 251–258, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Biran, O., Brody, S., and Elhadad, N. (2011). "Putting It Simply: A Context-Aware Approach to Lexical Simplification." In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 496–501, Portland, Oregon, USA. Association for Computational Linguistics.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). "Language Models Are Few-Shot Learners." In *Advances in Neural Information Processing Systems*, Vol. 33, pp. 1877–1901. Curran Associates, Inc.

Cristea, P. and Nisioi, S. (2024). "Archaeology at MLSP 2024: Machine Translation for Lexical Complexity Prediction and Lexical Simplification." In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z. (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 610–617, Mexico City, Mexico. Association for Computational Linguistics.

De Belder, J. and Moens, M.-F. (2010). "Text Simplification for Children." In *Proceedings of the*

2010 SIGIR Workshop on Accessible Search Systems, pp. 19–26.

Den, Y., Ogiso, T., Ogura, H., Yamada, A., Menematsu, N., Uchimoto, K., and Koiso, H. (2007). "The Development of an Electronic Dictionary for Morphological Analysis and Its Application to Japanese Corpus Linguistics [Kōpasu Nihongogaku No Tame No Gengo Shigen : Keitaiso Kaiseki Yō Denshika Jisho No Kaihatsu to Sono Ōyō] (in Japanese)." *Japanese Linguistics [Nihongo Kagaku]*, **22** (5), pp. 101–123.

Devlin, S. and Tait, J. (1998). "The Use of a Psycholinguistic Database in the Simplification of Text for Aphasic Readers." In Nerbonne, J. A. (Ed.), *Linguistic Databases*, Lecture Notes, pp. 161–173. CSLI Publications, Stanford, USA.

Dutilleul, B., Debaillon, M., and Mathias, S. (2024). "ISEP_Presidency_University at MLSP 2024 Shared Task: Using GPT-3.5 to Generate Substitutes for Lexical Simplification." In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z. (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 605–609, Mexico City, Mexico. Association for Computational Linguistics.

Enomoto, T., Kim, H., Hirasawa, T., Nagai, Y., Sato, A., Nakajima, K., and Komachi, M. (2024). "TMU-HIT at MLSP 2024: How Well Can GPT-4 Tackle Multilingual Lexical Simplification?" In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z. (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Fernando, A. and Dias, G. (2021). "Building a Linguistic Resource : A Word Frequency List for Sinhala." In Bandyopadhyay, S., Devi, S. L., and Bhattacharyya, P. (Eds.), *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pp. 606–610, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

Foreign Service Institute (2025). "Foreign Language Training." `https://www.state.gov/foreign-service-institute/foreign-language-training/`.

Glavaš, G. and Štajner, S. (2015). "Simplifying Lexical Simplification: Do We Need Simplified Corpora?" In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 63–68, Beijing, China. Association for Computational Linguistics.

Gooding, S. and Kochmar, E. (2018). "CAMB at CWI Shared Task 2018: Complex Word Identification with Ensemble-Based Voting." In Tetreault, J., Burstein, J., Kochmar, E., Leacock, C., and Yannakoudakis, H. (Eds.), *Proceedings of the 13th Workshop on Innovative*

Use of NLP for Building Educational Applications, pp. 184–194, New Orleans, Louisiana. Association for Computational Linguistics.

Gooding, S., Kochmar, E., Yimam, S. M., and Biemann, C. (2021). "Word Complexity Is in the Eye of the Beholder." In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4439–4449, Online. Association for Computational Linguistics.

Goswami, D., North, K., and Zampieri, M. (2024). "GMU at MLSP 2024: Multilingual Lexical Simplification with Transformer Models." In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z. (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 627–634, Mexico City, Mexico. Association for Computational Linguistics.

Goto, I., Tanaka, H., and Kumano, T. (2015). "Japanese News Simplification: Tak Design, Data Set Construction, and Analysis of Simplified Text." In *Proceedings of Machine Translation Summit XV: Papers*, pp. 17–31, Miami, USA.

Grabe, W. and Yamashita, J. (2022). *Reading in a Second Language: Moving from Theory to Practice* (2nd edition). Cambridge Applied Linguistics. Cambridge University Press, Cambridge.

Grainger, P. (2005). "Second Language Learning Strategies and Japanese: Does Orthography Make a Difference?" *System*, **33** (2), pp. 327–339.

Hading, M., Matsumoto, Y., and Sakamoto, M. (2016). "Japanese Lexical Simplification for Non-Native Speakers." In *Proceedings of the 3rd Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA2016)*, pp. 92–96, Osaka, Japan. The COLING 2016 Organizing Committee.

Hartmann, N. S., Paetzold, G. H., and Aluísio, S. M. (2018). "SIMPLEX-PB: A Lexical Simplification Database and Benchmark for Portuguese." In Villavicencio, A., Moreira, V., Abad, A., Caseli, H., Gamallo, P., Ramisch, C., Gonçalo Oliveira, H., and Paetzold, G. H. (Eds.), *Computational Processing of the Portuguese Language*, Lecture Notes in Computer Science, pp. 272–283, Cham. Springer International Publishing.

Hautasaari, A. and Ishida, T. (2011). "Discussion about Translation in Wikipedia." In *2011 2nd International Conference on Culture and Computing*, pp. 127–128.

Hayakawa, A., Kajiwara, T., Ouchi, H., and Watanabe, T. (2022). "JADES: New Text Simplification Dataset in Japanese Targeted at Non-Native Speakers." In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)*, pp. 179–187, Abu Dhabi, United Arab Emirates (Virtual). Association for Computational Linguistics.

Himeno, M. (1999). *Fukugō Dōshi No Kōzō to Imi Yōho [in Japanese]*. Hitsuji Shobō, Tōkyō.

Horiguchi, K., Kajiwara, T., Arase, Y., and Ninomiya, T. (2024). "Evaluation Dataset for Japanese Medical Text Simplification." In Cao, Y. T., Papadimitriou, I., Ovalle, A., Zampieri, M., Ferraro, F., and Swayamdipta, S. (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pp. 219–225, Mexico City, Mexico. Association for Computational Linguistics.

Horn, C., Manduca, C., and Kauchak, D. (2014). "Learning a Lexical Simplifier Using Wikipedia." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 458–463, Baltimore, Maryland. Association for Computational Linguistics.

Ide, Y., Mita, M., Nohejl, A., Ouchi, H., and Watanabe, T. (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)*, pp. 477–487, Toronto, Canada. Association for Computational Linguistics.

Japan Educational Exchanges and Services (2025). "Indication of the CEFR Level for Reference | JLPT Japanese-Language Proficiency Test." `https://www.jlpt.jp/e/about/cefr_reference.html`.

Japan Student Services Organization (2024). "Universities (Undergraduate) and Junior Colleges | Study in Japan Official Website." `https://www.studyinjapan.go.jp/en/planning/learn-about-schools/universities/`.

Kajiwara, T. and Yamamoto, K. (2015). "Evaluation Dataset and System for Japanese Lexical Simplification." In *Proceedings of the ACL-IJCNLP 2015 Student Research Workshop*, pp. 35–40, Beijing, China. Association for Computational Linguistics.

Katsuta, A. and Yamamoto, K. (2018). "Crowdsourced Corpus of Sentence Simplification with Core Vocabulary." In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T. (Eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kodaira, T., Kajiwara, T., and Komachi, M. (2016). "Controlled and Balanced Dataset for Japanese Lexical Simplification." In *Proceedings of the ACL 2016 Student Research Workshop*, pp. 1–7, Berlin, Germany. Association for Computational Linguistics.

Kudo, T., Yamamoto, K., and Matsumoto, Y. (2004). "Applying Conditional Random Fields to

Japanese Morphological Analysis." In Lin, D. and Wu, D. (Eds.), *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, Barcelona, Spain. Association for Computational Linguistics.

Lee, J. and Yeung, C. Y. (2018). "Personalizing Lexical Simplification." In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., and Zhu, C. (2023). "G-Eval: NLG Evaluation Using Gpt-4 with Better Human Alignment." In Bouamor, H., Pino, J., and Bali, K. (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522, Singapore. Association for Computational Linguistics.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). "Balanced Corpus of Contemporary Written Japanese." *Language Resources and Evaluation*, **48** (2), pp. 345–371.

Maruyama, T. and Yamamoto, K. (2018). "Simplified Corpus with Core Vocabulary." In Calzolari, N., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Hasida, K., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., and Tokunaga, T. (Eds.), *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Miyata, R., Koretaka, H., Yamauchi, H., Yanamoto, D., Kajiwara, T., Ninomiya, T., and Nishiwaki, Y. (2024). "MATCHA: Parallel Corpus for Japanese Text Simplification Based on Professionally Simplified Articles [in Japanese]." *Journal of Natural Language Processing*, **31** (2), pp. 590–609.

National Institute for Japanese Language and Linguistics (2025). "Compound Verb Lexicon." `https://www2.ninjal.ac.jp/vvlexicon/`.

Nishizawa, H., Isbell, D. R., and Suzuki, Y. (2022). "Review of the Japanese-Language Proficiency Test." *Language Testing*, **39** (3), pp. 494–503.

Nohejl, A., Hayakawa, A., Ide, Y., and Watanabe, T. (2024). "Difficult for Whom? A Study of Japanese Lexical Complexity." In Shardlow, M., Saggion, H., Alva-Manchego, F., Zampieri, M., North, K., Štajner, S., and Stodden, R. (Eds.), *Proceedings of the 3rd Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pp. 69–81, Miami, Florida, USA. Association for Computational Linguistics.

Nohejl, A., Hudi, F., Kardinata, E. A., Ozaki, S., Riera Machin, M. A., Sun, H., Vasselli, J., and Watanabe, T. (2025). "Beyond Film Subtitles: Is YouTube the Best Approximation of Spoken Vocabulary?" In Rambow, O., Wanner, L., Apidianaki, M., Al-Khalifa, H., Eugenio,

B. D., and Schockaert, S. (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics*, pp. 9566–9585, Abu Dhabi, UAE. Association for Computational Linguistics.

Nohejl, A. and Watanabe, T. (2025). "Dispersion Measures as Predictors of Lexical Decision Time, Word Familiarity, and Lexical Complexity." *arXiv preprint arXiv:2501.06536*.

North, K., Ranasinghe, T., Shardlow, M., and Zampieri, M. (2024). "MultiLS: An End-to-End Lexical Simplification Framework." In Shardlow, M., Saggion, H., Alva-Manchego, F., Zampieri, M., North, K., Štajner, S., and Stodden, R. (Eds.), *Proceedings of the 3rd Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pp. 1–11, Miami, Florida, USA. Association for Computational Linguistics.

Occhipinti, L. (2024). "Introducing MultiLS-IT: A Dataset for Lexical Simplification in Italian." In *Proceedings of the 10th Italian Conference on Computational Linguistics*, Pisa, Italy.

Omura, M., Wakasa, A., and Asahara, M. (2021). "Word Delimitation Issues in UD Japanese." In de Lhoneux, M. and Tsarfaty, R. (Eds.), *Proceedings of the 5th Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pp. 142–150, Sofia, Bulgaria. Association for Computational Linguistics.

OpenAI (2024a). "GPT-4 Technical Report." *arXiv preprint arXiv:2303.08774*.

OpenAI (2024b). "GPT-4o System Card." *arXiv preprint arXiv:2410.21276*.

Paetzold, G. and Specia, L. (2016a). "SemEval 2016 Task 11: Complex Word Identification." In Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., and Zesch, T. (Eds.), *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 560–569, San Diego, California. Association for Computational Linguistics.

Paetzold, G. and Specia, L. (2016b). "SV000gg at SemEval-2016 Task 11: Heavy Gauge Complex Word Identification with System Voting." In Bethard, S., Carpuat, M., Cer, D., Jurgens, D., Nakov, P., and Zesch, T. (Eds.), *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 969–974, San Diego, California. Association for Computational Linguistics.

Paetzold, G. and Specia, L. (2016c). "Understanding the Lexical Simplification Needs of Non-Native Speakers of English." In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 717–727, Osaka, Japan. The COLING 2016 Organizing Committee.

Paetzold, G. H. and Specia, L. (2017). "A Survey on Lexical Simplification." *Journal of Artificial Intelligence Research*, **60**, pp. 549–593.

Qiang, J., Li, Y., Zhu, Y., Yuan, Y., and Wu, X. (2020a). "Lexical Simplification with Pretrained

Encoders." *34th AAAI Conference on Artificial Intelligence*, pp. 8649–8656.

Qiang, J., Li, Y., Zhu, Y., Yuan, Y., and Wu, X. (2020b). "LSBert: A Simple Framework for Lexical Simplification." *arXiv preprint arXiv:2006.14939v1*.

Saggion, H., Bott, S., Szasz, S., Pérez, N., Calderón, S., and Solís, M. (2024). "Lexical Complexity Prediction and Lexical Simplification for Catalan and Spanish: Resource Creation, Quality Assessment, and Ethical Considerations." In Shardlow, M., Saggion, H., Alva-Manchego, F., Zampieri, M., North, K., Štajner, S., and Stodden, R. (Eds.), *Proceedings of the 3rd Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pp. 82–94, Miami, Florida, USA. Association for Computational Linguistics.

Shardlow, M. (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*, pp. 103–109, Sofia, Bulgaria. Association for Computational Linguistics.

Shardlow, M., Alva-Manchego, F., Batista-Navarro, R., Bott, S., Calderon Ramirez, S., Cardon, R., François, T., Hayakawa, A., Horbach, A., Huelsing, A., Ide, Y., Imperial, J. M., Nohejl, A., North, K., Occhipinti, L., Peréz Rojas, N., Raihan, N., Ranasinghe, T., Solis Salazar, M., Zampieri, M., and Saggion, H. (2024a). "An Extensible Massively Multilingual Lexical Simplification Pipeline Dataset Using the MultiLS Framework." In Wilkens, R., Cardon, R., Todirascu, A., and Gala, N. (Eds.), *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pp. 38–46, Torino, Italia. ELRA and ICCL.

Shardlow, M., Alva-Manchego, F., Batista-Navarro, R., Bott, S., Calderon Ramirez, S., Cardon, R., François, T., Hayakawa, A., Horbach, A., Hülsing, A., Ide, Y., Imperial, J. M., Nohejl, A., North, K., Occhipinti, L., Rojas, N. P., Raihan, N., Ranasinghe, T., Salazar, M. S., Štajner, S., Zampieri, M., and Saggion, H. (2024b). "The BEA 2024 Shared Task on the Multilingual Lexical Simplification Pipeline." In Kochmar, E., Bexte, M., Burstein, J., Horbach, A., Laarmann-Quante, R., Tack, A., Yaneva, V., and Yuan, Z. (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pp. 571–589, Mexico City, Mexico. Association for Computational Linguistics.

Shardlow, M., Evans, R., Paetzold, G. H., and Zampieri, M. (2021). "SemEval-2021 Task 1: Lexical Complexity Prediction." In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pp. 1–16, Online. Association for Computational Linguistics.

Shardlow, M., Evans, R., and Zampieri, M. (2022). "Predicting Lexical Complexity in English Texts: The Complex 2.0 Dataset." *Language Resources and Evaluation. arXiv preprint*

arXiv:2102.08773, pp. 1153–1194.

Smădu, R.-A., Ion, D.-G., Cercel, D.-C., Pop, F., and Cercel, M.-C. (2024). "Investigating Large Language Models for Complex Word Identification in Multilingual and Multidomain Setups." In Al-Onaizan, Y., Bansal, M., and Chen, Y.-N. (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 16764–16800, Miami, Florida, USA. Association for Computational Linguistics.

Speer, R. (2022). "Rspeer/Wordfreq: V3.0." Zenodo. `https://zenodo.org/record/7199437`.

Sunakawa, Y., Lee, J.-h., and Takahara, M. (2012). "The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries." *Acta Linguistica Asiatica*, **2** (2).

Tack, A. (2021). *Mark My Words! On the Automated Prediction of Lexical Difficulty for Foreign Language Readers*. Ph.D. thesis, KU Leuven.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V., and Zhou, D. (2022). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pp. 24824–24837, Red Hook, NY, USA. Curran Associates Inc.

Yatskar, M., Pang, B., Danescu-Niculescu-Mizil, C., and Lee, L. (2010). "For the Sake of Simplicity: Unsupervised Extraction of Lexical Simplifications from Wikipedia." In Kaplan, R., Burstein, J., Harper, M., and Penn, G. (Eds.), *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 365–368, Los Angeles, California. Association for Computational Linguistics.

Yimam, S. M., Biemann, C., Malmasi, S., Paetzold, G., Specia, L., Štajner, S., Tack, A., and Zampieri, M. (2018). "A Report on the Complex Word Identification Shared Task 2018." In *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 66–78, New Orleans, Louisiana. Association for Computational Linguistics.

Yimam, S. M., Štajner, S., Riedl, M., and Biemann, C. (2017). "CWIG3G2 - Complex Word Identification Task across Three Text Genres and Two User Groups." In *Proceedings of the 8th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pp. 401–407, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Yoshioka, K. (2016). "Acquisition of Mimetics and the Development of Proficiency in L2 Japanese: A Longitudinal Case Study of an L1 Dutch Speaker's Speech and Gesture." In *The Grammar of Japanese Mimetics*, pp. 172–192. Routledge.

Appendix

A English–Japanese Glossary and Manual Dataset Tags

Term	Abbreviation (🏷 Tag)	Japanese Term
auxiliary (convenience category[a])	🏷 AUX	
auxiliary verb		助動詞
bunsetsu		文節
adjective	🏷 ADJ	形容詞
adjectival noun	🏷 AN	形容動詞（形状詞）
adverb	🏷 ADV	副詞
classifier → numeral classifier		
conjunction	🏷 CONJ	接続詞
complex particle	🏷 CPART	複合助詞
complex auxiliary		複合助動詞
dependent word		付属語
formal noun	🏷 FN	形式名詞
fuzokugo → dependent word		
honorific prefix	🏷 HON	敬語接頭辞
independent word		自立語
interjection	🏷 INTERJEC	感動詞
long unit word	LUW	長単位
noun phrase	🏷 NP	名詞句
numeral classifier	🏷 CLF	助数詞
particle[b]	🏷 PART	助詞
prefix	🏷 PREF	接頭辞
prenominal modifier	🏷 PREN	連体詞
pronoun	🏷 PRON	代名詞
mimetic word	🏷 MIM	擬声語・擬音語・擬態語・擬情語
sahen verb		サ変動詞 = サ行変格活用動詞
short unit word	SUW	長単位
suffix	SUF	接尾辞
verb	🏷 V	動詞
verb+verb compound verb	🏷 VV	動詞＋動詞型の複合動詞
verbal noun	🏷 VN	動名詞

🏷 marks abbreviations that are also used as manual tags in the dataset (see Appendix B).

[a] For convenience, we understand as auxiliaries (AUX) all of the following: auxiliary verbs, including copulas (e.g., だ), attached typically to AN; the so-called "conjunctive particles" (接続助詞) that attach to verbs (e.g., て, ば); the light verb する used in honorific verb templates or with attributive forms AN, ADV (e.g., はっきりした); the dummy verb いる (after て); and the dummy copula ある (in である).

[b] In the dataset, we categorize the so-called "conjunctive particles" (接続助詞) that attach to verbs (e.g., て, ば) in the convenience auxiliary category (AUX).

B The MultiLS-Japanese Dataset Format

The dataset is available in a TSV (tab-separated values) format, where the fields below are columns and data instances are rows. See Appendix A for POS tags used for manual tagging.

Field	Also in MLSP2024:	Description
`id`	●	Unique identifier in the MLSP2024 dataset.
`context`	●	Sentence context.
`target`	●	Target (a substring of `context`).
`position`		Target span in `context` as "*start*:*stop*" 0-based Unicode code point indices.
—— *Metadata:* ——		

`primary_pos`
`primary_wf`
`primary_cf`
		For primary targets, the triple (POS, word frequency band, character composition band) identifies them in the space of candidate word properties (see Section 3.6.1). For other targets, the fields are empty.

`source_category`		Category of the source (identifies Category in Table 2).
`source_id`		Identifier of the source (identifies Source in Table 2).
`source_name`		Full website title.
`source_doc_url`		Full document URL (e.g., Wikipedia page or Aozora card).
`source_doc_title`		Full document title (e.g., Wikipedia page, literary work name and author).
—— *Annotation:* ——		
`lcp_annotator_`{1 . . . 10}		Individual lexical complexity ratings ∈ {1 . . . 5}.
`ls_annotator_`{1 . . . 10}		Individual LS annotations, up to 3 separated by comma. Equal to `target` if none given.
`complexity`	●	Mean lexical complexity normalized to [0, 1].
`substitution_`{1 . . . 28}	●	Aggregated LS annotations, repeated as many times as given by the annotators. Right-padded with empty fields.
—— *Automatic segmentation and tagging:* ——		
SUW in following field names represents `tokenized`, `base`, `lemma`, `type`, and `pos`, fields resulting from MeCab/Unidic SUW segmentation, namely: surface (表層形), orthBase (書字形基本形), lemma+subLemma (語彙素+語彙素細分類), goshu/wType (語種), and pos (品詞). The SUWs are separated by vertical bars.		
`token_position`		Target span in `context_tokenized` as "*start*:*stop*" 0-based SUWs indices.
`context_`*SUW*		SUW segmentation and tags of `context`.
`target_`*SUW*		SUW segmentation and tags of `target`.
—— *Manual segmentation and tagging:* ——		
Fields named *`_pos` use tags ADJ, ADV, AN, AUX, CONJ, INTERJEC, N, PART, PREF, PREN, PRON, SUF, V. Fields named *`_morphology` additionally use CLF, FN, HON (subtypes of SUF, N, PREF, respectively).		

`main`
`aux`
`dep`
		`target` manually split into the independent word `main`, auxiliaries that we consider advanced (likely to contribute to complexity) `aux`, and (other) dependent words `dep`. Any of `aux` and `dep` may be empty.

`main_pos`		A POS manually assigned to `main`, or "MWE" if `main` is an MWE.
`main_morphology`		POS manually assigned to morphemes of `main`, where words of an MWE are separated by hyphens, morphemes of a word by vertical bars.
`main_special`		A special subtype of `main`. May be ADV:MIM, N:FN, PART:CPART, SUF:CLF, V:VN, V:VV, or empty.
`main_mwe_type`		If `main_pos` is "MWE", one of N-V (noun-verb), NP (noun phrase), ADV-V (adverb-verb), else empty.
`aux_morphology`		POS manually assigned to morphemes of `aux`, separated by vertical bars.
`dep_pos`		A POS manually assigned to `dep`.
—— If `main` is part of a longer productively formed word (LUW), e.g. "翌" in "翌 1988 年 8 月 29 日": ——		
`whole`		A productively formed word of which `main` is a part.
`whole_pos`		A POS manually assigned to `whole`.
`whole_part`		May be "start" or "end" based on whether `whole` starts or ends with `main`.

●: fields common with other languages in MLSP2024: `https://huggingface.co/datasets/MLSP2024/MLSP2024`

C Guidelines for Determining Target Word Spans

The **Guidelines** we used to determine the boundaries of each of the 600 target words in the MultiLS-Japanese dataset are divided into Appendices C.1, C.2, and C.3, going bottom–up in terms of linguistic structure. In each section's **Results**, we report how the respective guidelines are reflected in targets' manual tags, and how many targets they were applied to. See Section 3.5 for our motivation and an overview of basic segmentation units (SUW, LUW, and *bunsetsu*).

C.1 Compounding and Derivation

In Japanese, compounding and derivation can be very productive and the resulting words effectively constitute Sino-Japanese phrases (e.g., "962 名収容可能" "able to accommodate 962 persons"). These words would typically constitute a single LUW. In many cases, it is possible to simplify their constituents independently, targeting units smaller than words from the perspective of Japanese morphology.

Note that two-morpheme native Japanese (*wago*) compounds, of which verb+verb compound verbs are a prominent example, and some words that could be considered Sino-Japanese (*kango*) compounds are segmented as a single SUW, and therefore do not require special handling.

Guidelines:

1. If a word is productively formed by derivation or compounding and the target part can be simplified independently, include only the necessary SUWs of the word. e.g., the prefix 翌 in "翌 1988 年 8 月 29 日" or the two-SUW compound 収容可能 in a longer compound "962 名収容可能" can be easily simplified independently.
2. Otherwise, include all constituents (SUWs) of a word, in particular:
   a. prefixes, including honorific, e.g., ほの暗い, お菓子,
   b. suffixes, e.g., 配偶者, 何ら,
   c. compound constituents, e.g., チケット販売委託会社,
   d. all constituents of a complex particle, e.g. について, につきまして,
   e. all constituents of a complex auxiliary, e.g., ざるを得ない, and the verb it attaches to.

Results: We manually tagged the morphology of the target words in the field `main_morphology`. Target words that are part of a larger whole have a non-empty field `whole`. See Appendix B for details. There are 30 such target words: 29 nouns and 1 adjectival noun.

C.2 Dependent Words (*Fuzokugo*)

A sequence of dependent words (SUWs) may be included in a target word, typically spanning at most one *bunsetsu*. In many cases, these dependent words (SUWs) can be thought of as inflections. Note, however, that inflections not analyzed as separate SUWs do not need special handling, e.g., ない → なけれ, and we therefore do not discuss them.

Guidelines:

3. Always include the following dependent words (SUWs):
   a. the light verb する, e.g., はっきりする, がっかりする,
   b. auxiliary verbs attached to verbs and adjectives, e.g., なすべき, 麗しかった,
   c. the conjunctive particle て/で in the *te*-form of verbs and adjectives, e.g., 浮んで, 麗しくて, and the dummy verb いる attached to the *te*-form if expressing a state, e.g., 酔っ払っている,
   d. copula forms (inflections) attached to adjectival nouns, e.g., 巨大な, で, 頻繁に,
   e. particles attached to adverbs, e.g., ぽつんと,
   f. particles and copula forms attached to attributively (adverbially) used nouns, e.g., ぼろぼろの, ぼろぼろだ, 常に,
   g. sequences of the above words, e.g., 取り上げられた(aux. られる + aux. た).
4. Do not include the following dependent words (SUWs), except if necessary for simplification:
   a. particles other than those mentioned above: among others, case particles, e.g., が, から,
   b. anything following the *te*-form, except the dummy verb いる expressing a state,
   c. syntactically compounding (suffix-like) verbs attached to the adverbal form (連用形), e.g., 忘れる, 得る, すぎる, にくい
   d. dependent words preceding the independent word, e.g., を連れて unless they form a complex particle e.g., をめぐって or a similar fixed expressions, e.g., にならい.
5. When in doubt about dependent words, think of the possible simplifications and include them if necessary to allow those simplifications.

Results: You can find the dependent parts of targets in the fields `aux` and `dep`, see Appendix B. 201 of the targets have at least one.

Note that the so-called "conjunctive particle" (接続助詞) て/で, used pervasively in the *te*-form of verbs and adjectives, requires special handling (see 3.c and 4.b above). We generally do include other conjunctive particles in targets (e.g., ば in 手に入れなければ, see 4.a), except when they are part of complex auxiliaries (e.g., ば in なければならない, or つつ in つつも, see 2.d).

C.3 Multiword Expressions

Guidelines:

6. If an expression (collocation) can be easily simplified independently, do not include unnecessary words, e.g., 愚痴 in the collocation 愚痴を言う (to complain, lit. “to say a complaint”) can be simplified independently.
7. Otherwise, include the minimal expression that can be simplified easily, e.g., all words of the idiom 愚痴をこぼす (to complain, lit. “to spill a complaint”) need to be included.

Results: All MWE target words in the dataset have `main_pos` value “MWE”, see Appendix B. The dataset contains the following MWEs:

- 36 noun-verb (N-V): e.g., 責任を持つ (to bear responsibility)
- 2 noun phrase (NP), e.g., 川の氾濫 (river flooding),
- 1 noun-adjective (N-ADJ): 体格のいい (to have a good physique),
- 1 verb-verb (V-V): 打って変わって (completely change to).
- 1 adverb-verb (ADV-V): ぱっとせん (to be inconspicuous),
- 1 particle-verb (PART-V): にならい (following the example of),

D Annotation Instructions and Profile Questionnaires

D.1 Lexical Complexity Annotation Instructions

1. あなた自身の情報を入力する

あなた自身の情報について、「Profile」シートに入力してください。

2. 点数をつける

あなたは、「Annotation」シートを使って作業をします。「Annotation」シートには、日本語の文が200個書かれています。それぞれの文には、下線が引かれた言葉が3つずつあります。これらの言葉は、あなたにとってどれぐらい難しいですか？読み方ではなく、言葉の意味だけを考えてください。読み方ではなく、言葉の意味だけを考えてください。次のリストを参考に、1から5までの点数をつけてください。

1. **とてもやさしい** あなたがとてもよく知っている言葉である。
2. **やさしい** あなたがよく知っている言葉である。
3. **ふつう** 「やさしい」でも「難しい」でもない言葉である。
4. **難しい** この言葉はよく知らないが、漢字や前後の言葉から意味を予想できる。
5. **とても難しい** あなたがまったく知らないか、ほとんど意味がわからない言葉である。

例　潜水艦は隠れることで真価を発揮するため、浮上しないことが望ましい。

真価
浮上しない
望ましい

例のように、文の下に点数のメニューがあります。①、②、③と順番にクリックすることで、1から5までの点数を選べます。下線が引かれた単語それぞれについて、あなたが考える点数を選んでください。

点数をつけるときに、インターネットや本などを使って言葉を調べてはいけません。

Figure 2 Lexical complexity instructions, page 1 of 2: instructions in Japanese.

1. Fill in information about yourself

Please, fill in information about yourself in the “Profile” sheet.

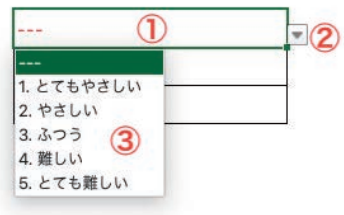
2. Assign points

You will work in the “Annotation” sheet. In the “Annotation” sheet, there are 200 Japanese sentences. In each sentence, there are 3 underlined words. How difficult are these words for you? Please consider only meaning of the words, not their reading (pronunciation). Based on the following list, assign 1 to 5 points to each word.

1. とてもやさしい	**Very easy.** Words you know very well.
2. やさしい	**Easy.** Words you know well.
3. ふつう	**Neutral.** Words that are neither easy nor difficult.
4. 難しい	**Difficult.** Words you do not know very well, but you can guess their meaning from their characters (kanji) or the sentence context.
5. とても難しい	**Very difficult.** Words that you do not know at all, or are very unclear.

例　潜水艦は隠れることで真価を発揮するため、浮上しないことが望ましい。

真価
浮上しない
望ましい



As shown in the example, there are menus with points below the text. By clicking in the order marked by ①, ②, and ③ in the picture, you can choose from 1 to 5 points. Please choose the points for each of the underlined words based on your judgement.

When assigning points, do not search for words on the internet or in books.

Figure 3 Lexical complexity instructions, page 2 of 2: instructions in English.

D.2 Lexical Complexity Profile Questionnaire

(1) 日本語能力試験 (JLPT) レベル。過去に合格したレベルのうち、最も高いものをメニューから選んでください。

Your JLPT level. Please choose from the menu the highest level you hold a JLPT certificate for.

Options: N5, N4, N3, N2, N1, (2009 以前) 4, (2009 以前) 3, (2009 以前) 2, (2009 以前) 1

(2) 年齢。

Your age.

例: 「25」

(3) 小学校から数えて、あなたが学校に行った年数。

How many years did you go to school (spent in formal education) counting from elementary school?

例: 「16」(小学校 6 年、中学校 3 年、高校 3 年、 大学 4 年のとき、6+3+3+4 → 16 / elementary 8y, high school 4y, university 4y: 8+4+4 → 16)

(4) 母語。いくつかある場合は、コンマ（、）で区切り、記入してください。

Your native language. If you have multiple native languages, please separate them with commas.

例: 「タイ語」「Tagalog, English」

(5) 母語を除いて話せる言語の数。

How many languages can you speak, except your native language(s)?

例: 「1」(母語と日本語だけの場合 / if you speak only your native language and Japanese)

(6) 1 週間で日本語を読む時間数。

How many hours a week do you read in Japanese?

例: 「3」「0.5」

(7) 大学や大学の「日本語予備教育」で日本語を勉強したことがありますか。

Have you studied Japanese at university, or in "Japanese language preparatory education" for university?

Options: はい, いいえ

(8) 今、日本に住んでいますか。

Do you currently live in Japan? **Options:** はい, いいえ

(9) 合わせて日本に住んでいた年数。

How many years have you lived in Japan in total?

例: 「0」「0.5」「3」

D.3 Lexical Simplification Annotation Instructions

1. あなた自身の情報を確認する

あなた自身の情報について、「Profile」シートに確認してください。

2. アノテーション作業

アノテーション作業は「Annotation」シートで行います。「Annotation」シートには、アノテーションの対象となる文が200個書かれています。それぞれの文は、下線が引かれた単語やフレーズ（ターゲット）を過不足なく3つ含みます。

作業内容は、これらのターゲットに置換できる、平易な言い換え（シンプラー）を挙げることです。このとき、以下の条件を満たす必要があります。

・シンプラーは、元の文におけるターゲットの意味を概ね保っている。
・シンプラーは、（中国語や韓国語を母語としない）日本語学習者にとって、
　ターゲットより理解しやすい。
・シンプラーでターゲットを置換したとき、文法的に正しく、意味がわかる文となる。

シンプラーは、ターゲットごとに最大で3個挙げてください。文の下に、ターゲットごとにシンプラーを記述する表があります。左詰めでシンプラーを記入してください。もし1つも思いつかなかった場合は、ターゲットをシンプラー (1) にそのまま記入してください。

作業中に辞書などの資料を参照することは構いませんが、ChatGPT をはじめとした自動ツールは使用しないでください。日本語教育についてのあなたの知識や経験を生かして、あなたがふさわしいと思うシンプラーを挙げてください。

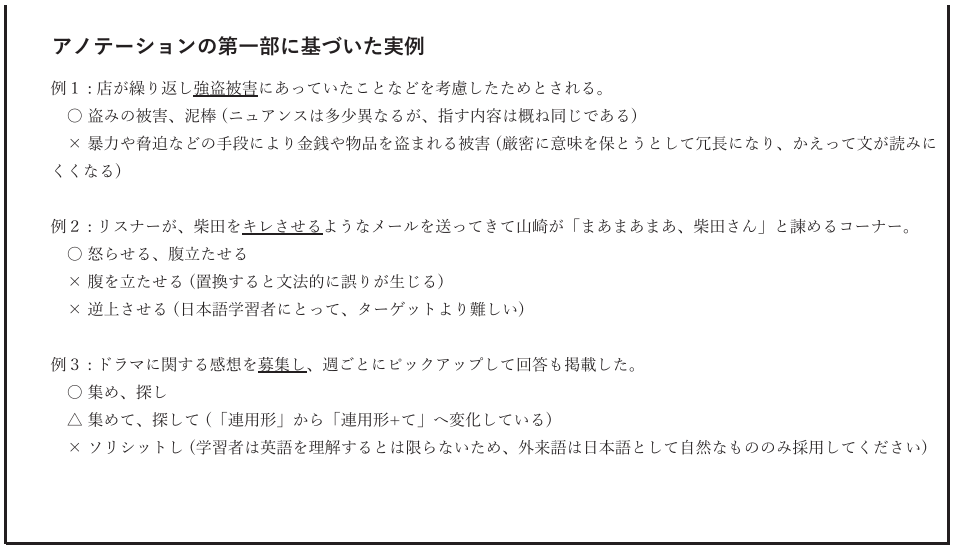
例　潜水艦は隠れることで真価を発揮するため、浮上しないことが望ましい。

	シンプラー (1)	シンプラー (2)	シンプラー (3)
真価	実力	本当の力	真の価値
浮上しない	浮かばない	水の上に出ない	
望ましい	より良い		

シンプラーを挙げる際には、以下の点にも注意してください。

・ターゲットの難しさが語形や機能表現に起因する場合、それらを変えることは問題ない。
　○ せざるを得なかった → しなければならなかった
　○ わからぬ → わからない
・ただし、「終止形(普通形)」と「連用形+ます(マス形)」、また「連用形」と「連用形+て(テ形)」は全て分かりやすい語形であると想定し、これらの語形を互いに変更しないでください。
　× 記入した → 書きました
　× 記入しました → 書いた
　× 記入し → 書き入れて
　× 記入して → 書き入れ
・ターゲットとシンプラーの品詞や語数は、文脈に適応すれば異なってもよい。
　○ 底知れぬ → 無限の、終わりのない
・ターゲットの漢字をひらがなやカタカナに変えただけのものは、シンプラーとみなさない。
　× 望ましい → のぞましい
・シンプラーは、最も一般的な表記であることが望ましい。
　× ほんとう　○ 本当
　× 所為　○ せい
・シンプラーは、読み仮名を含んではいけない。
　× 躊躇する → 躊躇（ちゅうちょ）する

Figure 4 LS instructions, page 1 of 2: initial instructions.


アノテーションの第一部に基づいた実例

例１：店が繰り返し強盗被害にあっていたことなどを考慮したためとされる。
○ 盗みの被害、泥棒（ニュアンスは多少異なるが、指す内容は概ね同じである）
× 暴力や脅迫などの手段により金銭や物品を盗まれる被害（厳密に意味を保とうとして冗長になり、かえって文が読みにくくなる）

例２：リスナーが、柴田をキレさせるようなメールを送ってきて山崎が「まあまあまあ、柴田さん」と諫めるコーナー。
○ 怒らせる、腹立たせる
× 腹を立たせる（置換すると文法的に誤りが生じる）
× 逆上させる（日本語学習者にとって、ターゲットより難しい）

例３：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。
○ 集め、探し
△ 集めて、探して（「連用形」から「連用形+て」へ変化している）
× ソリシットし（学習者は英語を理解するとは限らないため、外来語は日本語として自然なもののみ採用してください）

Figure 5 LS instructions, page 2 of 2: additional examples given after reviewing the trial set annotation (first 10 sentences).

English Translation

The following translation of the text shown in Figures 4 and 5 was not given to the annotators, who were all native Japanese speakers.

1. Fill in information about yourself

Please, fill in information about yourself in the “Profile” sheet.

2. Annotation

Annotation work is performed in the “Annotation” sheet. The “Annotation” sheet contains 200 sentences to be annotated. Each sentence contains exactly three underlined words or phrases (targets). The task is to list simple paraphrases (simplifications) that can be substituted for these targets. The following conditions must be met:

- The simplification preserves most of the meaning of the target in the original sentence.
- The simplification is easier for learners of Japanese (do not assume they are native speakers of Chinese or Korean) to understand than the target.
- When the target is replaced by the simplification, the sentence is grammatically correct and the meaning is understandable.

List at most three simplifications for each target. Below the sentence, there is a table for simplifications of each target. Please fill in the simplification starting from the leftmost column. If you

can't think of any, just enter the target as it is in the field "Simplification (1)".
You may refer to dictionaries and other resources while working, but please do not use ChatGPT or other automated tools. Please list simplifications that you consider appropriate, based on your knowledge and experience in Japanese language teaching.

[Image of the annotation interface, with fields for each target word labeled "Simplification (1–3)"]

When listing simplifications, please pay attention to the following points:

- If the difficulty of the target is due to word forms or functional expressions, feel free to change them:
  ○ ought to do → should do ○ know not → don't know *[English approximations]*
- However, assume that the terminal form (dictionary form) and the adverbal form + *masu* (*masu*-form), or the adverbal form and the adverbal form + *te* (*te*-form) are all easy to understand word forms, and do not change one to the other.
  [Examples of unnecessary changes of verb form]
- The POS and the number of words in the target and the simplification can be different if they fit the context.
  ○ bottomless → without end *[English approximation]*
- Simply changing the target's kanji is to hiragana or katakana is not a simplification.
  × 望ましい → のぞましい *[Change to an easier spelling]*
- Simplifications should use the most common orthography:
  × ほんとう *[Easier but less common]* ○ 本当
  × 所為 *[More difficult and also less common]* ○ せい
- Simplification should not be supplemented with readings.
  × 躊躇する → 躊躇 (ちゅうちょ) する *[Reading (ruby) supplied in parentheses]*

Examples based on the first part of the annotation

[Summary: The Japanese text uses actual good and bad examples of substitutions from the first part of annotation to illustrate the following points:

(1) Prefer rough equivalents over accurate but long explanations, which make the resulting sentence more difficult to comprehend.
(2) Make sure the substitutions are grammatical.
(3) Make sure the substitutions are easier for language learners than the original word.
(4) Do not change the verb form unless necessary for the simplification.
(5) Do not suppose that the learners understand English. Only use loanwords to the extent they are natural in Japanese.*]*

D.4 Lexical Simplification Profile Questionnaire

The questions 1, 2, 3, 4 in the following questionnaire are identical to questions 2, 3, 5, 6 given to lexical complexity annotators (see Appendix D.2). All of the LS annotators were native Japanese speakers living in Japan.

(1) 年齢。 例: 「25」
(2) 小学校から数えて、あなたが学校に行った年数。 例: 「16」
(小学校 6 年、中学校 3 年、高校 3 年、 大学 4 年のとき、6+3+3+4 → 16)
(3) 母語を除いて話せる言語の数。 例: 「1」
(4) 1 週間のうち日本語を読む時間数。 例: 「3」「0.5」
(5) 日本語教育の経験年数。 例: 「4」
(6) 1 週間のうち日本語を教える時間数。 例: 「1」

English Translation

The following translation of the text above was not given to the annotators, who were all native Japanese speakers.

(1) Your age.
(2) How many years did you go to school (spent in formal education), counting from elementary school?
(3) How many languages can you speak, except your native language(s)?
(4) How many hours a week do you read in Japanese?
(5) How many years of experience teaching Japanese as a second language do you have?
(6) How many hours a week do you teach Japanese?

E Lexical Simplification Corrections

The authors corrected the LS annotations, making 134 manual edits, among which 45 edits were made in the trial set (0.150 edits per instance and annotator) and 89 in the test set (0.016 edits per instance and annotator), reflecting that the implemented quality control measures (manual check, individual feedback, instruction clarification using additional examples) were effective.

The corrections are outlined below:

(1) Perform a conventional character normalization. We automatically normalize the substitutions to Unicode NFKC, except for the full-width tilde character (～), which is almost exclusively used in the full-width form in Japanese.

(2) Ensure simplifications are different from targets. (a) If the simplification is just an orthographic variant of the target or its reading in hiragana, we delete it, e.g., 強力な: 剛力な, もの: 者. (b) In ambiguous cases, we do not delete the simplification, e.g., 凸凹 (reading でこぼこ or とつおう): でこぼこ. (c) If the simplification differs from the target just by lacking a very common productive prefix or suffix while keeping the stem, we delete it, e.g., 男っぽ過ぎる: 男っぽい, お菓子: 菓子.

(3) Ensure the simplifications preserve the original word form. Simplifications should not unnecessarily change the word form, in particular, the basic level of formality or politeness. In cases they do, we adjust the word form.

(4) Correct grammatical errors. Simplifications should fit the context grammatically. (a) If a simplification can be corrected by a simple change of word form, we correct it, e.g., 死んだ → 死んで. (b) If a simplification repeats a word (typically a particle or a light verb) that is already contained in the context, we remove the repeated word, e.g., 家で → 家 (で already follows the target). (c) If a simplification misses a word (typically a particle or a light verb) that is part of the target, we add that word, e.g., 発表 → 発表されます (されます was in the target). (d) If more substantial changes are necessary (addition of a word not present in the target, replacement of a word, or a POS conversion), we delete the simplification. In particular, we do not make any changes that would require a choice between several possible edits.

(5) Perform an orthographic normalization. Simplifications should use the most common spelling of each word. If the simplification uses a less common spelling, we correct it, e.g., 阿呆 → アホ.

(6) Check simplicity. Simplifications should be easier to understand than the target words. In general, we do not examine simplification for complexity or frequency. If, however, a simplification has any of the issues above, and in addition to that, contains words much rarer or difficult to comprehend than the target, we delete it.

(7) Check semantics. Simplifications should preserve meaning. In general, we do not examine simplification for preservation of meaning. If, however, during the formal checks, we discover a simplification that clearly does not preserve the core meaning, we delete it, e.g., an affirmative instead of a negative, or 美容師 “beautician” simplified as 技術者 “technician”.

F Prompt Templates

In the following prompt templates, only the ↵ symbols represent line breaks. Italic text in curly braces (e.g., {*word*}) denotes placeholders to be replaced when constructing the prompt.

F.1 Our Lexical Simplification Prompt

For English, the placeholder {*language*} and the subsequent space in the following prompt template are omitted; for other languages, it is replaced with the name of the language (e.g., "Japanese"):

```
Answer with a list of ten different, simpler {language} words Y to replace the
difficult word X in the given sentence S. Each of the simpler words Y must follow
these rules:↵
1. When we replace the difficult word X with the simpler word Y, the sentence S is
grammatically correct.↵
2. When we replace the difficult word X with the simpler word Y, the sentence S
has the same meaning as before.↵
3. The simpler word Y is easier to understand than the difficult word X for
language learners.↵
### Sentence S: {context}↵
### Difficult word X: {word}↵
### List of simpler words Y:
```

F.2 TMU-HIT Lexical Simplification Prompt

We run a separate experiment using exactly the same prompt template as that used for languages other than Japanese by the TMU-HIT system (Enomoto et al. 2024). For English, the placeholder {*language*} and the subsequent space in the following prompt template are omitted; for other languages, it is replaced with the name of the language (e.g., "Japanese"):

```
I will give you a {language} sentence and a word in the 'Sentence' and 'Word'
format. List ten alternatives for the Word that are easier to understand,
separated by ','. ↵
You must follow these four rules.↵
1. Take into account the meaning of the Word in the Sentence.↵
2. Alternatives must be easier to understand than the Word.↵
3. Each alternative consists of one word.↵
4. Do not generate an explanation.↵
↵
Sentence: {context}↵
Word: {word}↵
Alternatives:
```

F.3 Our Lexical Complexity Prediction Prompt

The placeholder {*language*} in the following prompt template is replaced with the name of the language (e.g., "Japanese", "English"), and the placeholder {*examples*} is replaced with a concatenation of n few-shot examples in the following format:

```
Sentence: {context_i}↵
↵
Word: {word_i}↵
↵
Difficulty rating: {score_i}↵
↵
```

The prompt template:

```
You are an intermediate learner of {language}.[role] You will be given a sentence and
a word included in the sentence. Your task is to assign a rating to the word based
on how difficult to understand the word is for you.↵
↵
Based on the following list, assign a difficulty rating from 1 to 5 to the word.↵
- 1:  Very Easy - Words which are very familiar to you↵
- 2:  Easy - Words which are mostly familiar to you↵
- 3:  Neutral - When the word is neither difficult or easy↵
- 4:  Difficult - Words which you are unclear of the meaning, but may be able to
infer from the context↵
- 5:  Very Difficult - Words that you have never seen before, or are very unclear↵
↵
Please assign a difficulty rating to the {language} word.[final]↵
↵
{examples}Sentence: {context}↵
↵
Word: {word}↵
↵
Difficulty rating:
```

In addition to this prompt, we experiment with the following ablations:

(1) **No language:** The underlined sentence marked as [final] is replaced with:

```
Please assign a difficulty rating to the word.
```

(2) **No language, generic role:** The combination of (1) and replacing the underlined sentence marked as [role] with:

```
You are an intermediate language learner.
```

(3) **No language, no role:** The combination of (1) and (4).

(4) **No role:** The underlined sentence marked as [role] is removed.

F.4 TMU-HIT-like Lexical Complexity Prediction Prompt

We run a separate experiment using a prompt template based on the prompt P_{role} used by the TMU-HIT system (Enomoto et al. 2024) for English and Portuguese. We modify their prompt only by replacing the original scores represented by multiple tokens (0.0, 0.25, 0.5, 0.75, 1.0) with single-token ones (1, 2, 3, 4, 5), so that it can be used in our G-SCALE method.

The placeholder {*examples*} is replaced with a concatenation of n few-shot examples in the following format:

```
Sentence: {context_i}↵
↵
Word: {word_i}↵
↵
Complexity: {score_i}↵
↵
```

The prompt template:

```
You are an individual without specialized knowledge or expertise in a specific area.↵
↵
You will be given a sentence and a word included in the sentence.↵
↵
Your task is to rate the word on one metric.↵
↵
Please make sure you read and understand these instructions carefully. Please keep this
document open while reviewing, and refer to it as needed.↵
↵
↵
Evaluation Criteria:↵
↵
Complexity (1, 2, 3, 4, 5): the complexity of a word in terms of how difficult the word is
to understand.↵
↵
Evaluation steps:↵
1.  Read the sentence and word carefully to understand the context.↵
2.  Determine the complexity of the word based on the following criteria:↵
   - 1: The word is simple and easily understandable to most people.↵
   - 2: The word may have some complexity or be specific to a certain field, but can still
be understood with some effort.↵
   - 3: The word is moderately complex and may require some background knowledge or
explanation to understand fully.↵
   - 4: The word is quite complex and may be difficult to understand without significant
knowledge or explanation.↵
   - 5: The word is extremely complex and likely only understood by experts or individuals
with specialized knowledge.↵
3.  Assign a complexity rating to the word.↵
↵
Note: Your own familiarity with a word should not impact your rating. This should be based
on an average person 's understanding of the word.↵
↵
{examples}Sentence: '{context}'↵
↵
Word: '{word}'↵
↵

Complexity:
```

G Evaluation of LCP Using Correlation

Table 13 and Table 14 evaluate the same systems as Table 10, but using Pearson's and Spearman's correlation coefficients, respectively, instead of R^2. As both correlation coefficients are invariant to the application of a linear function with a positive slope, ablations of linear regression are omitted.

System	Catalan	English	Filipino	German	French	Italian	Japanese	Portug.	Sinhala	Spanish	Mean
Shared Task Top Results:											
Freq. Baseline	0.301	0.748	0.389	0.591	0.517	0.519	0.642	0.713	−0.196	0.551	0.478
Archaeology	0.274	0.790	0.443	0.551	0.533	0.534	0.485	0.683	0.044	0.527	0.487
GMU	0.155	**0.850**	0.282	0.140	0.319	0.292	0.177	—	0.125	0.244	0.287
TMU-HIT	**0.616**	0.820	**0.569**	**0.718**	0.625	0.601	0.733	**0.786**	0.308	**0.762**	0.654
Ablation of the Scoring Function:											
Direct	0.534	0.794	0.463	0.690	0.738	0.589	0.715	0.729	0.303	0.719	0.627
G-Eval	0.573	0.827	0.494	0.711	0.764	0.629	0.745	0.761	0.328	0.739	0.657
G-Scale	0.583	0.840	0.504	0.714	**0.774**	**0.654**	**0.767**	0.774	**0.330**	0.749	**0.669**

Table 13 Pearson's correlation coefficient of lexical complexity prediction on the MultiLS datasets, compared with the top MLSP shared task submission and frequency baseline results reported in the shared task findings (Shardlow et al. 2024b), and ablated versions of the G-Scale scoring function.

System	Catalan	English	Filipino	German	French	Italian	Japanese	Portug.	Sinhala	Spanish	Mean
Shared Task Top Results:											
Freq. Baseline	0.311	0.745	0.418	0.610	0.522	0.542	0.668	0.743	−0.256	0.530	0.483
Archaeology	0.265	0.755	0.448	0.573	0.531	0.532	0.513	0.692	0.030	0.479	0.482
GMU	0.157	0.798	0.277	0.147	0.321	0.296	0.183	—	0.130	0.198	0.279
TMU-HIT	**0.599**	0.755	**0.582**	**0.737**	0.630	0.622	0.731	**0.799**	**0.334**	**0.746**	0.653
Ablation of the Scoring Function:											
Direct	0.519	0.756	0.476	0.687	0.713	0.583	0.703	0.743	0.295	0.658	0.613
G-Eval	0.575	0.809	0.523	0.711	**0.762**	0.655	0.804	0.789	0.308	0.695	0.663
G-Scale	0.576	**0.811**	0.524	0.710	**0.762**	**0.657**	**0.805**	0.789	0.308	0.696	**0.664**

Table 14 Spearman's rank correlation coefficient of lexical complexity prediction on the MultiLS datasets, compared with the top MLSP shared task submission and frequency baseline results reported in the shared task findings (Shardlow et al. 2024b), and ablated versions of the G-Scale scoring function.

H Wordfreq Log-Frequency as a Feature

In Table 15, we compare the performance of linear regression based on log-frequency only, G-SCALE without additional variables and G-SCALE with log-frequency as an additional variable. For log-frequency, we use `wordfreq`, a Python library (Speer 2022) combining frequency data from multiple corpora, e.g. Wikipedia, Twitter, and film subtitle corpora. For Sinhala, `wordfreq` does not provide any data. Therefore, we substitute a word frequency list representative of diverse domains [22] (Fernando and Dias 2021). The MLSP shared task's baseline was also based mostly on `wordfreq`.

The mean R^2 drops when we add log-frequency as a feature, but this drop is actually driven by only three languages: Catalan, Filipino, and Sinhala. For all other languages, the addition of log-frequency results in an improvement. For Catalan, Filipino, and Sinhala, the log-frequency alone has a negative R^2 and hence the negative effect is not surprising. As all three are low-resource languages, we hypothesize that this reflects the limited quality of the `wordfreq` data for them. Without further analysis, however, we also cannot rule out issues with annotation quality or word distribution of the trial sets of these languages.

System	Catalan	English	Filipino	German	French	Italian	Japanese	Portug.	Sinhala	Spanish	Mean
Linear Regression without LLM Prompting:											
Log-Frequency	−0.385	0.547	−0.165	0.073	0.149	0.227	0.359	0.514	−0.346	0.256	0.123
G-SCALE (with LLM Prompting):											
LLM	**−0.060**	0.566	**0.078**	0.146	0.460	0.312	0.500	0.461	**0.043**	0.429	**0.293**
LLM + Log-F.	−0.081	**0.695**	−0.066	**0.163**	**0.463**	**0.313**	**0.563**	**0.560**	−0.319	**0.468**	0.276

Table 15 Using `wordfreq` log-frequency as an additional variable: R^2 of lexical complexity prediction on the MultiLS datasets.

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

I Examples of Lexical Simplification Errors

Validity	Issue	Example: [Context] Target → Simplification	Gold Data Example
Invalid	Not preserving meaning	メディア → 新聞 media → newspapers	マスコミ mass media
	Not a simplification	想像上の → 架空の imaginary → fictitious	非現実的な unreal
	Ungrammatical	悪化 [について] → 悪くなる [about] deterioration → worsen	低下 decline
	Not preserving meaning & not a simplification	責任を持つ → 責任を果たす have a responsibility → fulfill a responsibility	担当する be in charge (of)
	Not preserving meaning & ungrammatical	[くる] 可能性があります → あります there is a possibility that [it comes] → there is	かもしれません maybe
	Ungrammatical & not a simplification	溺れたり [すること] → 溺れる [cases of] drowning → drown	水に飲まれたり / being swallowed by the water
Valid	Simplification not listed by annotators	匿名の → 名前なしの anonymous → without a name	名前がない not having a name
	Orthography not listed by annotators	詳細に → くわしく minutely → in detail	詳しく in detail
	Superfluous reading in parentheses attached	櫓 → 台（だい） turret → tower (*reading*)	台 tower

Table 16 Examples of LS errors in terms of accuracy (instances where the top system simplification is not among gold simplifications by human annotators) of our system on MultiLS-Japanese. See analysis in Table 9.

Adam Nohejl: Adam Nohejl received his bachelor's and master's degrees in computer science from Charles University in 2009 and 2011, respectively, and a bachelor's degree in Japanese studies from Charles University in 2018. He worked as an independent software developer. He is currently a Ph.D. student at the Nara Institute of Science and Technology. His research interests include applications of natural language processing to text simplification, readability, and language learning.

Akio Hayakawa: Akio Hayakawa received his bachelor's and master's degrees in education from the University of Tokyo in 2015 and 2017, respectively, and obtained a master's degree in engineering from the Nara Institute of Science and Technology in 2023. He is currently a Ph.D. student at the Universitat Pompeu Fabra. His research focuses on applying natural language processing to education and social inclusion.

Yusuke Ide: Yusuke Ide received his B.A. degree from the University of Tokyo in 2017 and a master's degree in engineering from the Nara Institute of Science and Technology in 2023. He is currently a Ph.D. student at the Nara Institute of Science and Technology. His research interests lie in natural language processing, particularly lexical semantics and readability.

Taro Watanabe: Taro Watanabe received his B.E. and M.E. degrees in information science from Kyoto University in 1994 and 1997, respectively, and obtained an M.S. degree in Language and Information Technologies from the School of Computer Science, Carnegie Mellon University in 2000. In 2004, he received a Ph.D. in informatics from Kyoto University. After working as a researcher at ATR, NTT and NICT, and as a software engineer at Google, he is a professor at the Nara Institute of Science and Technology starting in 2020. His research interests include natural language processing, machine learning, language modeling and machine translation.

(Received April 1, 2025)
(Revised June 28, 2025)
(Accepted August 5, 2025)