Difficult for Whom? A Study of Japanese Lexical Complexity

Adam Nohejl[1] **Akio Hayakawa**[2] **Yusuke Ide**[1] **Taro Watanabe**[1]

[1]Nara Institute of Science and Technology
{nohejl.adam.mt3, ide.yusuke.ja6, taro}@is.naist.jp

[2]Universitat Pompeu Fabra
akio.hayakawa@upf.edu

Abstract

The tasks of lexical complexity prediction (LCP) and complex word identification (CWI) commonly presuppose that difficult to understand words are shared by the target population. Meanwhile, personalization methods have also been proposed to adapt models to individual needs. We verify that a recent Japanese LCP dataset is representative of its target population by partially replicating the annotation. By another reannotation we show that native Chinese speakers perceive the complexity differently due to Sino-Japanese vocabulary. To explore the possibilities of personalization, we compare competitive baselines trained on the group mean ratings and individual ratings in terms of performance for an individual. We show that the model trained on a group mean performs similarly to an individual model in the CWI task, while achieving good LCP performance for an individual is difficult. We also experiment with adapting a finetuned BERT model, which results only in marginal improvements across all settings.

1 Introduction

Complex word identification (CWI) is a task of identifying difficult to understand words in text. CWI systems can be used as components of lexical simplification and readability assessment systems. Lexical complexity prediction (LCP) extends CWI by predicting complexity of words on a continuous scale (Shardlow et al., 2020).

For both tasks, it is necessary to specify for whom we are predicting the complexity. Non-native speakers have very different needs from people with dyslexia (Paetzold and Specia, 2016) or children (Oshika et al., 2024). For non-native speakers, their L1 background (Machida, 2001; Ide et al., 2023) or proficiency level (Lee and Yeung, 2018b) further determines their needs.

A case has recently been made for personalized CWI, which predicts complex words for an individual (Lee and Yeung, 2018b; Gooding and Tragut, 2022), and similar methods were earlier proposed for personalized reading assistance (Ehara et al., 2013). While most research has been done on English as a second language, a personalized CWI system for Chinese as a second language has also been proposed (Lee and Yeung, 2018a). A shared element of the previously proposed systems is a binary classifier based on a small number of features, such as word frequency or a level from a pedagogical word list. This fits the hypothetical scenario of deployment to user devices and training them using very little labeled data.

Meanwhile, models of increasing size have been applied to lexical complexity prediction targeting relatively wide target populations. In a recent multi-lingual shared task (Shardlow et al., 2024b), systems based on large language models (GPT-4) or encoder models (BERT) performed well, especially on relatively high-resource languages such as English or Japanese. The systems were, however, evaluated only on the basis of complexity averaged across all annotators.

We will attempt to answer the following questions for the specific case of the Japanese data employed by the shared task (Shardlow et al., 2024a,b), MultiLS-Japanese:

1. Is the data representative of the intended target population?
2. Can complexity predictions for individuals be improved by training personalized models?
3. How does a simple frequency-based model using a suitable corpus compare to the recent computationally intensive models?

2 Analysis

The MultiLS-Japanese dataset is designed as an evaluation dataset consisting of 30 trial instances and 570 test instances. Annotation instructions, annotator profiles, and separate complexity data for each annotator were released online as well.[1] Each instance of the dataset is a target word in a sentence

[1]https://github.com/naist-nlp/multils-japanese

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 69–81
November 15, 2024 ©2024 Association for Computational Linguistics

	Original Data	Non-CK L1 Replication	Chinese L1 Reannotation
Native languages	English (5), Swedish (1), Portuguese & English (1), French & English (1), Basque & Spanish (1), French (1)	Czech (7), English & Czech (1), Czech & Ukrainian (1), Slovak (1)	Chinese (9), Chinese & Cantonese (1)
JLTP level	1 (3), N1 (3), N2 (3), 2 (1)	N2 (7), N1 (3)	N1 (5), N2 (4), 1 (1)
Studied Japanese at university	7 of 10	10 of 10	2 of 10
Currently lives in Japan	10 of 10	0 of 10	10 of 10
Lived in Japan (total yrs)	16.7 (8.3)	0.7 (0.4)	4.6 (2.5)
Reading in Japanese (hrs/week)	5.7 (7.6)	2.6 (2.3)	9.5 (8.7)
Age (yrs)	40.8 (9.1)	23.6 (2.7)	28.2 (2.5)
Education (total yrs)	18.4 (3.7)	17.2 (2.4)	19.5 (2.9)
Non-native languages	1.7 (0.5)	3.1 (1.1)	2.6 (0.8)

Table 1: Comparison of the annotator groups of the original data, our replication (same conditions), and our reannotation by Chinese L1 speakers. In the last five rows, we report means followed by standard deviations in parentheses.

context, for which lexical complexity values and simpler substitutions are provided. In this study, we ignore the substitutions as well as the context.

Each instance of both trial and test data was rated by the same set of annotators, which allows us to use the individual ratings in a personalized setting.

2.1 Target Population

The annotators were holders of Japanese Language Proficiency Test (JLPT) levels N1 or N2 (or their older equivalents 1 and 2). These levels of JLPT are often required by employers and universities (JASSO, 2024) and have been compared to CEFR levels B2 and C1 (Sophia University, 2024). The native language of the annotators was purposely not Chinese or Korean (non-CK), as both languages share a large part of their vocabulary with Japanese. Maekawa et al. (2014) estimates the proportion of words of Chinese origin[2] in Japanese text as 17% to 47% based on register. Heo (2010) estimates the proportion of words of Chinese origin in Korean text as 66%.

As shown in Figure 1, the distribution of complexity values in the trial set closely mimics the test set. The distributions of word origins and parts of speech are comparable as well (see Appendix A). We therefore used the trial set to evaluate how representative the dataset is of its target population. For this purpose we had the trial set reannotated by two groups of annotators: one is from the same target population, while the other has Chinese as their native language. Demographics of each group are summarized in Table 1.

[2]The traditional terminology for Japanese vocabulary distinguishes between *wago*, indigenous Japanese words; *kango* Sino-Japanese words; and *gairaigo*, foreign words from other languages (e.g. English). For simplicity we will call them words of Japanese, Chinese, and other origin, respectively.

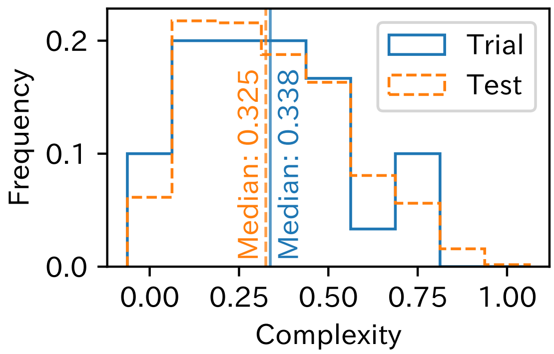


Figure 1: Complexity histogram of the trial and test sets.

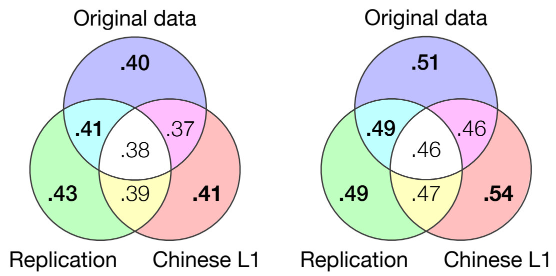
For the **non-CK L1 replication** we recruited annotators fulfilling the conditions of the original data. Notably, their native languages are neither Chinese nor Korean, but have almost no overlap with native languages of the original annotators. Additionally, while the original annotators have been living in Japan for an average 16.7 years, for the replication we have recruited undergraduate students or recent graduates of Japanese studies from Charles University in Prague, most of whom have been learning Japanese for 3 to 4 years, out of which no more than 1 year was spent in Japan.

The **Chinese L1 reannotation** group consists entirely of native Chinese speakers, students or recent graduates of Nara Institute of Science of Technology. The distribution of their proficiency levels is the same as that of the original annotators (six hold JLPT level N1/1 and four hold N2/2). Their mean age and time spent in Japan falls between the means of the original annotators and replication annotators.

We measured inter-annotator agreement (IAA), using Krippendorff's (1970) α for interval values, as well as the mean pairwise correlation between

Word Origin	#Words	TUBELEX $\log_{10}$ frequency	Original Data Complexity	Difference in Complexity From Original Data: Non-CK L1 Replication	Difference in Complexity From Original Data: Chinese L1 Reannotation
Japanese	12	−5.423 (1.427)	0.327 (0.231)	+0.040 (0.129)	+0.079 (0.102)
Chinese	13	−5.247 (0.913)	0.342 (0.180)	+0.029 (0.119)	−0.131 (0.093)
p-value		0.714	0.866	0.843	< **0.0001**

Table 2: Difference in complexity perceived by two groups of annotators and by the original annotators according to word origin. Mean values are followed by standard deviations in parentheses. We also show log-frequency and original complexity. The *p*-values were obtained from the two-sided unpaired exact permutation test (Good, 2004). Bold font denotes a statistical significant difference in the means between words of Japanese and Chinese origin.



(a) IAA (Krippendorff's α) (b) Mean pairwise PCC

Figure 2: Inter-annotator agreement and mean pairwise correlation in the three annotator groups, the unions of their pairs, and union of all three. Light text denotes that union decreases agreement (correlation).

annotators. As we can see in Figure 2, the values achieved in both the replication and the Chinese L1 reannotation are similar to the original data. When we merge the original data with the replication, however, the inter-annotator agreement does not drop below the agreements in the two groups. In other words, the annotators agree across these two groups as much as within them. This contrasts with the Chinese L1 reannotation, which lowers the agreement when combined with the original data, the replication, or their union. The same applies to the mean pairwise correlation. A similar tendency for IAA of CK and non-CK L1 annotators was reported by Ide et al. (2023) for an earlier non-native Japanese LCP dataset, but their native Chinese and Korean annotators also tended to have higher proficiency levels, which complicates the interpretation. In this study, they have the same distribution of proficiency levels.

The underlying cause is a different perception of complexity of words of Japanese and Chinese origin between native Chinese speakers and others. The words perceived as less complex by the native Chinese annotators are almost exclusively words of Chinese origin and vice versa (details provided in Appendix C). As we can see in Table 2, the gap in complexity perceived by the two groups differs significantly between words of Chinese and Japanese origin. The words of Chinese and Japanese origin do not, however, differ significantly in their frequency, complexity perceived by the original annotators, or the gap between complexity perceived by the original and the Chinese L1 annotators.

The statistical similarity with annotations by a group with very different demographics supports the hypothesis that the dataset is representative of the target population of non-native Japanese speakers with JLPT proficiency level N2 and higher, whose L1 is not Chinese or Korean.

While the difference between native Chinese speakers and others was to be expected, the similarity with the replication is remarkable. Within the boundaries of target population, we tried to find a homogeneous group of annotators with much less exposure to Japanese language than the original annotators. The students who replicated the trial annotation usually reach level N2 or N1 around their graduation after three to four years of study with limited exposure to Japanese outside their classes. The original annotators not only have lived on average 16.7 years in Japan, but in four cases also acquired their JLPT certificates before year 2010 (as evidenced by old JLPT levels 1 and 2 as opposed to N1 and N2), having ample opportunity to widen their vocabulary beyond the certified level. It is rather surprising how well the two groups agree.

We would like to emphasize that similarly low levels of IAA are common for LCP (e.g. $\alpha = 0.32$ or 0.31 reported by Ide et al., 2023), which reflects the subjectivity of the task and shows that there is a room for improvement by personalization.

2.2 Correlation Analysis

Word frequency has long been used as a feature for modeling lexical complexity (Devlin and Tait, 1998, is an early example). Furthermore, Nohejl et al. (2024) demonstrated for multiple languages including Japanese that frequency in TUBELEX, a YouTube subtitle corpus, has a stronger correlation with lexical complexity than frequency in other corpora. We examine correlation with several other variables, not considered by Nohejl et al. (2024).

Variable	Data Source	PCC	Potential PCC
Word Log-Frequency	TUBELEX	**−0.66**	**−0.66**
	Lang-8-non-CK	−0.64	−0.64
	Lang-8-CK	−0.61	−0.61
	CSJ	−0.57	−0.56
	BCCWJ	−0.55	−0.57
L2 Level	JEV	0.43	0.63
Character Log-Frequency	BCCWJ	−0.35	−0.37
	Lang-8-non-CK	−0.35	−0.36
	Lang-8-CK	−0.33	−0.34
L1 Familiarity	WLSP (Asahara)	−0.23	−0.55

Table 3: Correlation (PCC) of MultiLS-Japanese test set complexity with log-frequencies, learner levels and native familiarity. For BCCWJ, CSJ, JEV, and WLSP, values were looked up by lemma. Potential PCC only considers words present in each data source. Rows are ordered by PCC strength (absolute value): naturally, *high* complexity is associated with *low* frequency and familiarity, hence the negative values.

For a fair comparison, we measure Pearson's correlation coefficients (PCC) on the full MultiLS-Japanese data. The handling of words missing in data sources is detailed in Appendix B. We also report "Potential PCC" measured only on words present in the individual data sources, thus effectively evaluating each data source on a different subsets of MultiLS-Japanese.

As shown in Table 3, TUBELEX achieves the strongest correlation, followed by the subset of the learner corpus Lang-8 (Mizumoto et al., 2011), where the learners' L1 is not Chinese or Korean (Lang-8-non-CK). The difference between the two is not statistically significant, whereas the difference between Lang-8-non-CK and Lang-8-CK (L1 is Chinese or Korean) is significant.[3]

Among word frequencies, the weakest correlations are achieved by the corpora CSJ (NINJAL, 2016) and BCCWJ (Maekawa et al., 2014). Character frequencies further underperform word frequencies. The Japanese Educational Vocabulary (JEV)[4] by Sunakawa et al. (2012), targeting L2 learners, and the WLSP-Familiarity database (Asahara, 2019), rated by native speakers, have strong potential correlations, but their practical usefulness for LCP is limited by their low coverage, reflected by low actual PCC.

3 Experiments

Following the design of MultiLS-Japanese, we use the 30 trial instances for training, and the 570 test instances for evaluation. We only use the datasets original data, not the replication or reannotation, for the experiments.

We evaluate models in four settings determined by training and test data, e.g. the "Group-Individual" denotes training on group data (mean for LCP or majority class for CWI) and evaluation on individual data. With the exception of the Group-Group setting, where a single model is evaluated on a single test set, we therefore report the results as means and standard deviations. In the case of Individual-Individual, we evaluate each model trained on individual data only on the corresponding individual test data.

We also evaluate models in the CWI task by considering complexity values ≥ 0.375 (the midpoint between the *easy* and *neutral* ratings in MultiLS-Japanese) to be complex. The results in CWI are easier to interpret, and can be compared with previous personalized CWI research. In addition to CWI models (binary classifiers), we also evaluate LCP models in CWI (henceforth LCP-CWI) by interpreting their values as the positive class if they exceed the threshold.

For LCP, we measure R^2, the coefficient of determination. For CWI, we measure performance using macro-averaged F1 score, i.e. the average of F1 scores for the positive and negative class, in line with previous research (Yimam et al., 2018; Gooding and Tragut, 2022).

Detailed information about the experimental models is provided in Appendix D.

3.1 Frequency Baseline

As a baseline for LCP, we fit a linear regression using log-frequency in TUBELEX to the trial data. As shown in Table 4, the model performs well in the Group-Group setting (0.41), on par with the best R^2 result for Japanese in the shared task (0.413) obtained using a GPT-4-based model (Enomoto et al., 2024).

If we, however, train and evaluate the same baseline on individual data, the performance drops drastically (0.13). This may be counter-intuitive, as we are training and evaluating on the data annotated by the same individual, but it shows that that the strong correlation with log-frequency, and consequently the good performance of the baseline on group data, is mostly a result of individual idiosyncrasies being smoothed out by the group average. For LCP, the personalized Individual-Individual frequency baseline did not fare well. Results in the other settings were even worse with mean R^2 below zero.

[3] Based on Steiger's (1980) test for dependent correlations with significance level $\alpha = 0.01$.

[4] http://jhlee.sakura.ne.jp/JEV/

Test Train	Group	Individual
Group	**0.41**	−0.10 (0.41)
Individual	−0.06 (0.64)	**0.13 (0.15)**

Table 4: LCP results (R^2) using TUBELEX log-frequency as a single feature.

Model	Test Train	Group	Individual
LCP-CWI	Group	0.71	0.65 (0.05)
	Individual	0.56 (0.18)	0.56 (0.11)
CWI	Group	**0.78**	0.67 (0.06)
	Individual	0.77 (0.02)	**0.67 (0.04)**

Table 5: CWI results (F1) using TUBELEX log-frequency as a single feature.

For CWI, we fit a logistic regression model using the same single feature, and compare it with the LCP model, evaluated as LCP-CWI. As in the previous case, the results in Table 5 show that both kinds of models perform worse in the Individual-Individual setting than in the Group-Group setting, although the difference is smaller in CWI. Surprisingly, however, the CWI model in the Group-Individual setting reaches almost the same F1 score as personalized Individual-Individual CWI models. Additionally, the LCP model in the Group-Individual setting is very competitive when evaluated as LCP-CWI (0.65), outperforming the personalized LCP model (0.56) and nearing the performance of personalized CWI models (0.67). While it is difficult to predict the exact complexity in LCP, models trained on the group perform relatively well in the CWI task, even for individuals.

3.2 BERT-Based Model

The target population of MultiLS-Japanese is similar to that of non-CK L1 data of the Japanese Lexical Complexity for Non-Native Readers (JaLeCoN) dataset (Ide et al., 2023). We finetuned the BERT model described by Ide et al. for CK and non-CK data of the whole JaLeCoN dataset. To adapt it to MultiLS-Japanese, we used its output (predicted complexity) as a feature for linear and logistic regression either alone or together with the TUBELEX log-frequency. Appendix E provides results of all variants.

The best results, shown in Tables 6 and 7, were achieved by combining frequency with the model finetuned on JaLeCon-non-CK. All settings achieved only a marginal improvement over the frequency baseline.

Test Train	Group	Individual
Group	**0.43**	−0.08 (0.41)
Individual	−0.04 (0.65)	**0.15 (0.15)**

Table 6: LCP results (R^2) using TUBELEX log-frequency and output of the BERT model trained on JaLeCoN-non-CK.

Model	Test Train	Group	Individual
LCP-CWI	Group	0.72	0.66 (0.05)
	Individual	0.57 (0.19)	0.57 (0.12)
CWI	Group	**0.79**	**0.67 (0.06)**
	Individual	0.77 (0.02)	0.67 (0.04)

Table 7: CWI results (F1) using TUBELEX log-frequency and output of the BERT model trained on JaLeCoN-non-CK.

4 Conclusion

We demonstrated that the MultiLS-Japanese dataset is representative of its intended target population by comparing its IAA and correlation with an annotation replicated by a group with different demographics but fulfilling the conditions of proficiency and not having a Chinese or Korean L1 background.

Additionally, we demonstrated a clear difference in complexity perception of Japanese words, based on word origin, between this population and native Chinese speakers of the same proficiency levels in Japanese. To which extent this applies to native Korean speakers is a question for future research.

We found that achieving good performance in individual LCP is more difficult than in individual CWI. In individual LCP, personalization resulted in a small improvement over training on group data, but in individual CWI, personalization and training on group data performed similarly well.

The TUBELEX frequency baseline performed on par with the GPT-4-based model that achieved the best result in a recent shared task. Combining the frequency feature with a fine-tuned BERT model resulted only in marginal improvements in both the group and the individual setting.

In future work, we would like to investigate the effect of larger training data paired with additional features (e.g. register of a word) and the performance of different methods of sampling training data, such as uncertainty sampling.

Lay Summary

To make text easier to understand using an automated system, it is necessary to identify difficult words, which depends on the text's reader. The automated systems, therefore, need to focus on a specific target population, such as non-native speakers, or be personalized for the reader. The difficulty of words is called "lexical complexity" and can be rated on a scale.

The performance of systems for estimating lexical complexity can be scored using specialized datasets in which complexity is rated by people from the target population. The systems for estimating lexical complexity are usually scored based on the average rating by a group from the target population, not individuals. Additionally, some of the best performing systems use large language models such as GPT-4, which are costly to run.

Our study uses a dataset targeting highly proficient non-native Japanese speakers, excluding native Chinese and Korean speakers, who would have the advantage of knowing vocabulary shared among the three languages. We explore the following questions:

1. Is the dataset representative of the target population?
2. Can personalized systems improve estimates over those for the group average?
3. How does a simple word-frequency-based system compare to the costlier models?

By having the data rerated by two new groups, we confirmed that the dataset represents the target group well and that native Chinese speakers perceive Japanese complexity differently.

We compared personalized systems and systems based on the group average in terms of performance for individuals in two scenarios: When estimating lexical complexity rated on a scale, personalized systems performed slightly better. When we only classified the words as difficult or not difficult, the systems based on the group average and the personalized ones performed similarly. Regardless of the system or the scenario, we found it much more challenging to achieve good performance for the individuals than for the group average, which smooths out individual idiosyncrasies.

A simple frequency-based system using word frequency in YouTube subtitles slightly outperformed a recent model based on GPT-4, which is much more expensive to run.

In future work, we would like to investigate the effect of larger training data paired with more complex systems, which would consider other features of the words, such as register (formal vs. informal).

Limitations

We focused on a specific target population of non-native speakers defined by the exclusion of two specific L1s and relatively high proficiency levels. Even the simple personalization methods, which did not perform particularly well in our setting, may provide an advantage for a more diverse population, effectively providing adaptation to large differences in proficiency. We also have not evaluated different methods of training data sampling (e.g. uncertainty sampling in an active learning scenario, which may improve performance while using the same size of training data). We only performed objective metric-based evaluation of the system's performance. An additional human evaluation would also be desirable.

Acknowledgments

The MultiLS-Japanese dataset was created as a part of a joint research project of Nara Institute of Science and Technology and Nikkei, Inc. We would like to express our sincere gratitude to the volunteer annotators for their cooperation on partial replication and reannotation. The replication was performed by students and graduates of Japanese Studies at Charles University. The reannotation by native Chinese speakers was performed by students and graduates of Nara Institute of Science and Technology. We would like to thank Petra Kanasugi and Huayang Li for helping us recruit the annotators. We are grateful to the anonymous reviewers for their insightful comments and helpful suggestions.

References

Masayuki Asahara. 2019. Word familiarity rate estimation using a Bayesian linear mixed model. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 6–14, Hong Kong. Association for Computational Linguistics.

Marc Brysbaert and Kevin Diependaele. 2013. Dealing with zero word frequencies: A review of the existing rules of thumb and a suggestion for an evidence-based choice. *Behavior Research Methods*, 45(2):422–430.

Siobhan Devlin and John Tait. 1998. The use of a psycholinguistic database in the simplification of text for aphasic readers. In John A. Nerbonne, editor, *Linguistic Databases*, Lecture Notes, pages 161–173. CSLI Publications, Stanford, USA.

Yo Ehara, Nobuyuki Shimizu, Takashi Ninomiya, and Hiroshi Nakagawa. 2013. Personalized reading support for second-language web documents. *ACM*

Transactions on Intelligent Systems and Technology, 4(2):1–19.

Taisei Enomoto, Hwichan Kim, Tosho Hirasawa, Yoshinari Nagai, Ayako Sato, Kyotaro Nakajima, and Mamoru Komachi. 2024. TMU-HIT at MLSP 2024: How well can GPT-4 tackle multilingual lexical simplification? In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 590–598, Mexico City, Mexico. Association for Computational Linguistics.

Phillip I. Good. 2004. *Permutation, Parametric, and Bootstrap Tests of Hypotheses (Springer Series in Statistics)*. Springer-Verlag, Berlin, Heidelberg.

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

Chul Heo. 2010. Examination how many using compound of chinese character words and investigate the frequency of use by using analysis of Modern Korean words 1.2 (in Korean). *Journal of Chinese Characters Education in Korea*, (34):221–244.

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

Klaus Krippendorff. 1970. Bivariate Agreement Coefficients for Reliability of Data. *Sociological Methodology*, 2:139–150.

John Lee and Chak Yan Yeung. 2018a. Automatic prediction of vocabulary knowledge for learners of Chinese as a foreign language. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–4.

John Lee and Chak Yan Yeung. 2018b. Personalizing lexical simplification. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 224–232, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Sayuki Machida. 2001. Japanese text comprehension by Chinese and non-Chinese background learners. *System*, 29(1):103–118.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Tomoya Mizumoto, Mamoru Komachi, Masaaki Nagata, and Yuji Matsumoto. 2011. Mining revision log of language learning SNS for automated Japanese error correction of second language learners. In *Proceedings of 5th International Joint Conference on Natural Language Processing*, pages 147–155, Chiang Mai, Thailand. Asian Federation of Natural Language Processing.

NINJAL (National Institute for Japanese Language and Linguistics [Kokuritsu Kokugo Kenkyūjo]). 2016. Construction of the Corpus of Spontaneous Japanese [Nihongo hanashikotoba kōpasu no kōchikuhō] (in Japanese).

Adam Nohejl, Frederikus Hudi, Eunike Andriani Kardinata, Shintaro Ozaki, Maria Angelica Riera Machin, Hongyu Sun, Justin Vasselli, and Taro Watanabe. 2024. Beyond Film Subtitles: Is YouTube the Best Approximation of Spoken Vocabulary? *ArXiv preprint*, arXiv:2410.03240v1 [cs].

JASSO (The Japan Student Services Organization). 2024. Universities (Undergraduate) and Junior Colleges ｜ Study in Japan Official Website.

Masashi Oshika, Makoto Morishita, Tsutomu Hirao, Ryohei Sasano, and Koichi Takeda. 2024. Simplifying translations for children: Iterative simplification considering age of acquisition with LLMs. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8567–8577, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Gustavo Paetzold and Lucia Specia. 2016. 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*, pages 717–727, Osaka, Japan. The COLING 2016 Organizing Committee.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis Salazar, Marcos Zampieri, and Horacio Saggion. 2024a. An extensible massively multilingual lexical simplification pipeline dataset using the MultiLS framework. In *Proceedings of the 3rd Workshop on Tools and Resources for People with REAding DIfficulties (READI) @ LREC-COLING 2024*, pages 38–46, Torino, Italia. ELRA and ICCL.

Matthew Shardlow, Fernando Alva-Manchego, Riza Batista-Navarro, Stefan Bott, Saul Calderon Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, Anna Hülsing, Yusuke Ide, Joseph Marvin Imperial, Adam Nohejl, Kai North, Laura Occhipinti, Nelson Peréz Rojas, Nishat Raihan, Tharindu Ranasinghe, Martin Solis

Salazar, Sanja Štajner, Marcos Zampieri, and Horacio Saggion. 2024b. The BEA 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589, Mexico City, Mexico. Association for Computational Linguistics.

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

Sophia University, Center for Language Education and Research. 2024. Levels of Japanese Language Courses.

James H. Steiger. 1980. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251.

Yuriko Sunakawa, Jae-ho Lee, and Mari Takahara. 2012. The Construction of a Database to Support the Compilation of Japanese Learners' Dictionaries. *Acta Linguistica Asiatica*, 2(2).

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

A Comparison of Word Origins and Parts of Speech in the Test and Trial Sets

		Test	Trial
Word Origin	Chinese	55.4%	50.0%
	English	5.3%	10.0%
Part of Speech	Noun	45.6%	36.7%
	Verb	27.5%	36.7%
	Adjectival Noun	7.4%	3.3%
	MWE	7.0%	6.7%
	Adverb	6.0%	10.0%
	Adjective	2.1%	3.3%
	Particle	1.8%	—
	Pronoun	0.9%	—
	Conjunction	0.7%	3.3%
	Suffix	0.5%	—
	Auxiliary	0.4%	—
	Prefix	0.2%	—

Table 8: Comparison of the test set and trial set in terms of proportions of words containing tokens of Chinese or English origin and parts of speech. The remaining target words are purely of indigenous Japanese origin. We distinguish between adjectives (形容詞, so-called *i*-adjectives) and adjectival nouns (形容動詞 or 形状詞, *na*-adjectives and *to*/*taru*-adjectives). The Particle category excludes conjunctive particles (接続助詞), which we categorize as Auxiliaries together with auxiliary verbs (助動詞). MWE are multi-word expressions, typically noun-verb phrases.

B Handling of Words Missing in Data Sources

Data Source	Values	Handling of Missing Values	Formula for One Token or Character x	Sequence of Tokens or Characters $\mathbf{s}$
All Corpora	Log-Frequency	Laplace smoothing	$f(x) = \log\left(\frac{\text{count}(x)+1}{\#\text{tokens} + \#\text{types}}\right)$	$f(\mathbf{s}) = \min_{x \in \mathbf{s}} f(x)$
JEV	Levels 1–6	Dummy values	$f(x) = \begin{cases} \text{level}(x) & \text{if } x \in \text{JEV} \\ 7 & \text{otherwise} \end{cases}$	$f(\mathbf{s}) = \max_{x \in \mathbf{s}} f(x)$
WLSP-Familiarity	$F \subset \mathbb{R}$	Dummy values	$f(x) = \begin{cases} \text{familiarity}(x) & \text{if } x \in \text{WLSP} \\ \min(F) & \text{otherwise} \end{cases}$	$f(\mathbf{s}) = \min_{x \in \mathbf{s}} f(x)$

Table 9: Handling of words (or characters) missing in data sources used for PCC computation in Table 3. For all corpora, we use Laplace smoothing recommended by Brysbaert and Diependaele (2013) to provide log-frequency values even for words missing in the corpora. To words missing in JEV, we assign the value corresponding to a level beyond those present in the data. To words missing in WLSP-Familiarity, we assign the minimum familiarity level present in the data. To sequences consisting of multiple tokens or characters, we assign the minimum or maximum value assigned to the individual items as appropriate.

C Difference of Complexity Perception by Annotators' L1 and Word Origin

C.1 Original Annotation and Chinese L1 Reannotation

Target Word	Word Origin	TUBELEX $\log_{10}$ Frequency	Complexity Original	Complexity Chinese L1	Complexity Difference ↓
掲載した	Chinese	−4.744	0.400	0.100	−0.300
恩を売り	Ch. + Ja.	−5.166	0.700	0.450	−0.250
標題	Chinese	−7.173	0.375	0.125	−0.250
考慮した	Chinese	−4.815	0.400	0.175	−0.225
強盗被害	Chinese	−5.554	0.400	0.225	−0.175
各種の	Chinese	−4.978	0.200	0.025	−0.175
気にかけない	Ch. + Ja.	−3.588	0.475	0.300	−0.175
書き添えられて	Japanese	−6.817	0.600	0.450	−0.150
長大な	Chinese	−6.317	0.475	0.325	−0.150
随所	Chinese	−5.857	0.725	0.600	−0.125
応用した	Chinese	−4.935	0.225	0.125	−0.100
旧	Chinese	−4.613	0.150	0.075	−0.075
市電	Chinese	−6.232	0.475	0.400	−0.075
募集し	Chinese	−4.664	0.100	0.050	−0.050
諌める	Japanese	−7.068	0.775	0.775	0.000
変更されて	Chinese	−4.105	0.100	0.100	0.000
または	Japanese	−2.939	0.075	0.075	0.000
戦闘曲	Chinese	−4.224	0.425	0.425	0.000
ロック	English	−4.245	0.025	0.050	+0.025
はじめ	Japanese	−4.239	0.025	0.075	+0.050
繰り返し	Japanese	−4.232	0.200	0.275	+0.075
小物	Japanese	−5.112	0.225	0.325	+0.100
馴染み深かった	Japanese	−7.913	0.500	0.600	+0.100
再び	Japanese	−4.462	0.075	0.175	+0.100
連れ戻す	Japanese	−6.602	0.300	0.400	+0.100
ピックアップして	English	−4.977	0.050	0.175	+0.125
直ちに	Japanese	−5.383	0.275	0.400	+0.125
なおかつ	Japanese	−5.103	0.500	0.700	+0.200
キレさせる	Japanese	−5.205	0.375	0.625	+0.250
コーナー	English	−4.325	0.100	0.400	+0.300

Table 10: Target words in the trial set of MultiLS-Japanese; their word origin; log-frequency; mean complexity annotated by the original annotators, whose L1 was neither Chinese or Korean, and the Chinese L1 annotators; difference between the former and the latter. The table is sorted by the complexity difference to highlight the overlap between words of Chinese origin and words perceived as less complex by the Chinese L1 annotators compared to the original annotators. "Ch. + Ja." denotes expressions mixing content words of Chinese and Japanese origin. We ignore the origin of common functional words such as particles and light verbs.

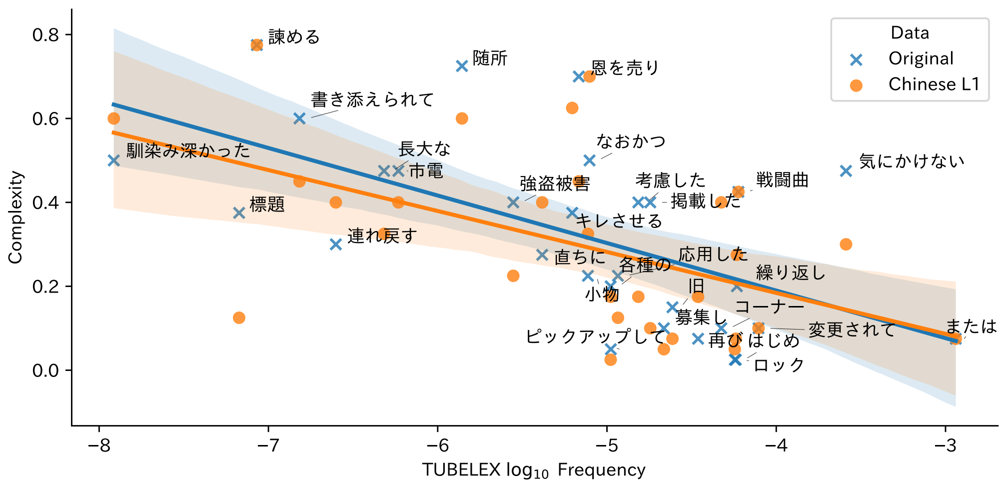


Figure 3: Mean complexity of target words in the the trial set of MultiLS-Japanese and in the Chinese L1 reannotation, plotted against log-frequency. Lines show linear fit with 95% confidence interval as a shaded area.

C.2 Original Annotation and Replication

Target Word	Word Origin	TUBELEX $\log_{10}$ Frequency	Complexity Original	Complexity Replication	Complexity Difference ↓∓
長大な	Chinese	−6.317	0.475	0.300	−0.175
繰り返し	Japanese	−4.232	0.200	0.050	−0.150
気にかけない	Ch. + Ja.	−3.588	0.475	0.375	−0.100
考慮した	Chinese	−4.815	0.400	0.325	−0.075
市電	Chinese	−6.232	0.475	0.400	−0.075
変更されて	Chinese	−4.105	0.100	0.050	−0.050
各種の	Chinese	−4.978	0.200	0.150	−0.050
再び	Japanese	−4.462	0.075	0.025	−0.050
書き添えられて	Japanese	−6.817	0.600	0.550	−0.050
随所	Chinese	−5.857	0.725	0.700	−0.025
または	Japanese	−2.939	0.075	0.050	−0.025
はじめ	Japanese	−4.239	0.025	0.000	−0.025
恩を売り	Ch. + Ja.	−5.166	0.700	0.700	0.000
連れ戻す	Japanese	−6.602	0.300	0.325	+0.025
ピックアップして	English	−4.977	0.050	0.075	+0.025
強盗被害	Chinese	−5.554	0.400	0.425	+0.025
直ちに	Japanese	−5.383	0.275	0.300	+0.025
小物	Japanese	−5.112	0.225	0.275	+0.050
旧	Chinese	−4.613	0.150	0.200	+0.050
馴染み深かった	Japanese	−7.913	0.500	0.550	+0.050
掲載した	Chinese	−4.744	0.400	0.475	+0.075
諌める	Japanese	−7.068	0.775	0.850	+0.075
応用した	Chinese	−4.935	0.225	0.325	+0.100
コーナー	English	−4.325	0.100	0.225	+0.125
標題	Chinese	−7.173	0.375	0.525	+0.150
なおかつ	Japanese	−5.103	0.500	0.700	+0.200
戦闘曲	Chinese	−4.224	0.425	0.625	+0.200
募集し	Chinese	−4.664	0.100	0.325	+0.225
ロック	English	−4.245	0.025	0.275	+0.250
キレさせる	Japanese	−5.205	0.375	0.725	+0.350

Table 11: Target words in the trial set of MultiLS-Japanese; their word origin; log-frequency; mean complexity annotated by the original annotators, and the replication annotators; difference between the mean complexities perceived by the two groups. Neither annotator group contained native Chinese or Korean speakers, hence compared to Table 10, there is not any clear tendency for words of Chinese origin.

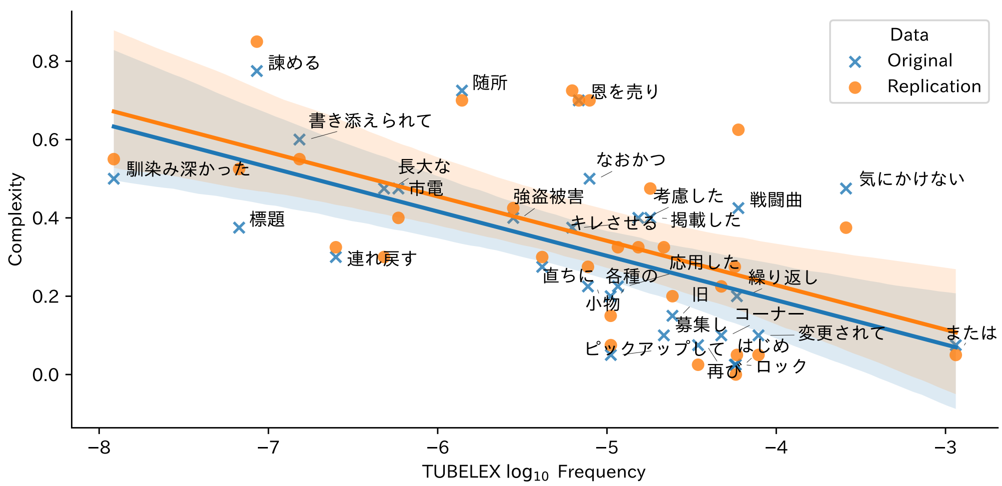


Figure 4: Mean complexity of target words in the the trial set of MultiLS-Japanese and in the replication. Lines show linear fit with 95% confidence interval as a shaded area.

D Model Details

Task	Model Description	Implementation	Postprocessing
LCP	Linear regression with L2 regularization ($\alpha = 1$)	`Ridge()`	Clip values to the valid range [0, 1].
CWI	Logistic regression with balanced class weights	`LogisticRegression( class_weight='balanced')`	—

Table 12: Details of the models used for experiments in Section 3, implemented using the `scikit-learn` Python package, namely classes from `sklearn.linear_model`. For baselines, the only feature is log-frequency in TUBELEX (see Appendix B). For the BERT-based models, the features are (1) output of the finetuned BERT model and optionally (2) log-frequency in TUBELEX. The BERT models are exactly as described by Ide et al. (2023), except that we finetuned them for CK and non-CK complexity of the whole JaLeCoN dataset (not using any train-test split of the data).

E Results of the BERT-based Model Variants

Test Train	Group	Individual
Group	**0.14**	−0.24 (0.40)
Individual	−0.30 (0.62)	**−0.01 (0.13)**

Table 13: LCP results (R^2) using output of the BERT model trained on JaLeCoN-non-CK.

Model	Test Train	Group	Individual
LCP-CWI	Group	0.47	0.47 (0.09)
	Individual	0.46 (0.17)	0.47 (0.12)
CWI	Group	0.73	0.63 (0.06)
	Individual	**0.73 (0.01)**	**0.64 (0.06)**

Table 14: CWI results (F1) using output of the BERT model trained on JaLeCoN-non-CK.

Test / Train	Group	Individual
Group	**0.41**	−0.10 (0.41)
Individual	−0.06 (0.64)	**0.13 (0.15)**

Table 15: LCP results (R^2) using TUBELEX log-frequency and output of the BERT model trained on JaLeCoN-CK.

Model	Test / Train	Group	Individual
LCP-CWI	Group	0.71	0.65 (0.05)
	Individual	0.56 (0.18)	0.56 (0.11)
CWI	Group	**0.78**	0.67 (0.06)
	Individual	0.77 (0.01)	**0.67 (0.04)**

Table 16: CWI results (F1) using TUBELEX log-frequency and output of the BERT model trained on JaLeCoN-CK.

Test / Train	Group	Individual
Group	**0.00**	−0.32 (0.39)
Individual	−0.43 (0.58)	**−0.09 (0.13)**

Table 17: LCP results (R^2) using output of the BERT model trained on JaLeCoN-CK.

Model	Test / Train	Group	Individual
LCP-CWI	Group	0.34	0.35 (0.08)
	Individual	0.34 (0.01)	0.37 (0.07)
CWI	Group	**0.62**	0.59 (0.06)
	Individual	0.62 (0.01)	**0.59 (0.07)**

Table 18: CWI results (F1) using output of the BERT model trained on JaLeCoN-CK.