Japanese Lexical Complexity for Non-Native Readers: A New Dataset

Yusuke Ide[1] **Masato Mita**[2] **Adam Nohejl**[1] **Hiroki Ouchi**[1,3] **Taro Watanabe**[1]
[1]Nara Institute of Science and Technology [2]CyberAgent Inc. [3]RIKEN
{ide.yusuke.ja6, nohejl.adam.mt3, hiroki.ouchi, taro}@is.naist.jp,
mita_masato@cyberagent.co.jp

Abstract

Lexical complexity prediction (LCP) is the task of predicting the complexity of words in a text on a continuous scale. It plays a vital role in simplifying or annotating complex words to assist readers. To study lexical complexity in Japanese, we construct the first Japanese LCP dataset. Our dataset provides separate complexity scores for Chinese/Korean annotators and others to address the readers' L1-specific needs. In the baseline experiment, we demonstrate the effectiveness of a BERT-based system for Japanese LCP.

1 Introduction

Reading comprehension requires a certain level of vocabulary knowledge. The results reported by Hu and Nation (2000) suggest that most English learners need to understand 98% of tokens in a text to comprehend it. A follow-up study by Komori et al. (2004) estimates the percentage to be 96% for Japanese learners to comprehend text. Acquiring vocabulary to reach such levels, in turn, is a lengthy and challenging task for learners. This opens up opportunities for assistive applications, such as simplification or annotation of complex words. The first step necessary for such applications is to predict the complexity of the words. The task of **lexical complexity prediction (LCP)** is defined as predicting how difficult to comprehend words or phrases in a text are on a continuous scale (Shardlow et al., 2020). This differentiates LCP from complex word identification (CWI), i.e., binary classification of complex words (Yimam et al., 2018). As complexity is naturally perceived as continuous, a continuous scale used in LCP allows to represent it without loss of information.

The LCP research so far has been limited to English, for which two LCP datasets have been constructed (Shardlow et al., 2020, 2022), and no such dataset has been created for Japanese. Meanwhile, there are a number of features specific to the Japanese language that could affect lexical complexity, and their effects have yet to be studied. For example, the Chinese characters, which are used extensively in Japanese, lower text readability (Tateisi et al., 1988).

Previous studies on Japanese lexical complexity used pedagogical word lists to estimate complexity level. Nishihara and Kajiwara (2020) modeled lexical complexity of words based on the Japanese Educational Vocabulary List (Sunakawa et al., 2012). The word list assigns a degree of difficulty to each item, based on the subjective judgment of Japanese language teachers, not learners themselves, and does not consider the learners' L1 background.

In light of this, we present JaLeCoN[1], Dataset of **Ja**panese **Le**xical **Co**mplexity for **N**on-Native Readers. Our dataset has the following key features:

(1) Complexity scores for single words as well as multi-word expressions (MWEs);

(2) Separate complexity scores from Chinese/Korean annotators and others, addressing the considerable advantage of the former in Japanese reading comprehension.

Our analysis reveals that the non-Chinese/Korean annotators perceive words of Chinese origin or containing Chinese characters as especially complex. In the baseline experiment, we investigate the effectiveness of a BERT-based system in the Japanese LCP task, and how it varies according to the word complexity and L1 background.

2 Task Setting

Since Japanese has no explicit word boundaries, word segmentation is the first prerequisite for LCP. We use short unit words (**SUWs**) as the basic word unit, combining them into longer word units in the case of multi-word expressions (**MWEs**):

[1]JaLeCoN is available at https://github.com/naist-nlp/jalecon.

Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 477–487

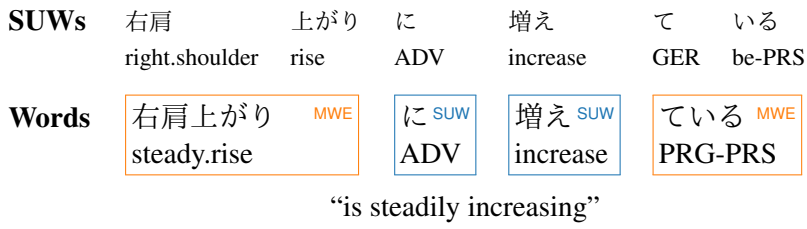

Figure 1: Example of text segmented as SUWs and as words (either SUW or MWE). Semantically opaque sequences are chunked into MWEs. Abbreviations in glosses: ADVerbializer, GERund, PReSent, PRoGressive.

SUW: SUWs consist of one or two smallest lexical units (Ogura et al., 2011), and are commonly used for segmentation of Japanese.

MWE: We understand MWEs as multi-*SUW* expressions that are fixed or semantically opaque (see Appendix C) and consequently may have higher complexity than their components. We identify MWEs either using long unit word (LUW)[2] segmentation, or manually (see Section 3).

Consequently, a **word**, can be either an SUW or an MWE (see Figure 1 for examples).

A **complexity score** represents perceived complexity based on the annotators' judgment on a scale from 0 (least complex) to 1 (most complex). We exclude proper nouns from our target because their complexity is influenced by factors unrelated to reading proficiency or vocabulary knowledge.[3]

We annotate the words in an **in-context dense** setting. In-context here means including both intra-sentence and extra-sentence context of each word. Context is important for lexical complexity for two reasons (Gooding and Kochmar, 2019; Shardlow et al., 2021): (1) As polysemous words can have different complexity levels for each sense, context is necessary to differentiate between possible meanings of these words. (2) Presenting a word without context could increase its complexity. In particular, the recognition of abstract words relies on context (Schwanenflugel et al., 1988). **Dense** means annotating each word of the text with a complexity label, instead of annotating one specific word in each sentence (Shardlow et al., 2022). We adopt the dense setting to avoid any bias that could arise from targeting specific words.

3 Construction of JaLeCoN

In order to include both written and spoken language and a variety of vocabulary, we sourced texts from two different genres:

News comes from the Japanese-English data of the WMT22 General Machine Translation Task (Kocmi et al., 2022). It contains a variety of news texts written for the general Japanese reader.

Government is composed of press conference transcripts from Japanese ministries or agencies.[4]

The whole dataset is composed of sequences of sentences constituting either the beginning of an article (News) or a question-answer pair (Government). We restricted the length of the sequences to at least 6 and at most 11 sentences to obtain similar amounts of text, and presented each sequence as a whole for annotation.

3.1 Word Segmentation

We used Comainu 0.80[5] (Kozawa et al., 2014) to perform two-level segmentation. The low-level SUW segmentation was done using MeCab (Kudo et al., 2004), a Japanese morphological analyzer, and the UniDic 2.3.0 (Den et al., 2007) dictionary. At the second level, Comainu chunked the SUWs into LUWs. Based on the two segmentations, we segmented the text into words as follows:

(1) If an LUW is a noun, we use the constituting SUWs as words. Transparent noun compounds are ubiquitous in Japanese (e.g., 次期 | 気象 | 衛星[6] “next meteorological satellite”), and we do not consider them MWEs.

(2) If an LUW is not a noun, we use the LUW as a word. Such an LUW may be a single SUW, or a sequence of SUWs, which we consider an MWE. Such MWEs most importantly include functional words, such as compound particles (e.g., に | 比べ | て “compared to”) and auxiliary verbs (e.g., なけれ | ば | なら | ない “have to”).

We also identified other MWEs manually, as explained in Section 3.3.

[2]The LUW is defined as a syntactic word by Omura et al. (2021).

[3]Sequences containing segmentation errors are also excluded (see Appendix D).

[4]The transcripts were retrieved from the websites of five organizations: JMA, JTA, MOJ, MOFA, and MLHW.

[5]https://github.com/skozawa/Comainu

[6]The vertical bars denote boundaries between SUWs.

Genre	Sentences	Words	MWE Ratio	CK		Non-CK	
				All Words	MWEs	All Words	MWEs
News	400	10,256	7.9%	.009	.020	.024	.072
Government	200	7,964	14.4%	.005	.009	.028	.047

Table 1: Statistics of JaLeCoN. The CK and Non-CK columns show the mean complexity scores by L1 group.

3.2 Complexity Annotation

To capture the lexical complexity for a non-native Japanese reader with intermediate or advanced reading ability, we recruited 15 annotators per sentence with Japanese reading proficiency ranging from CEFR (Common European Framework of Reference for Languages) level B1 to C2. We required at least intermediate proficiency, as it has been shown that complexity judgments made by intermediate or advanced learners can be used to adequately predict the needs of beginners but not vice versa (Gooding et al., 2021). The proficiency levels were self-reported (see Appendix A for details). We used the annotations made by 14 of them, after removing one outlier, whose annotations had over 70% higher mean than those of any other annotator, clearly not corresponding to the reported reading proficiency.

Approximately half of the annotators we recruited have a Chinese/Korean L1 background (CK).[7] CK learners have a considerable advantage in comprehension of words of Chinese origin, which also form a large part of Chinese and Korean vocabulary (Koda, 1989).

The annotators were asked to assign one of the following labels to each span if they find it complex: 3 (Very Difficult), 2 (Difficult), or 1 (Not Easy); otherwise the annotators were to leave the span unlabeled and we interpreted it as 0 (Easy).[8] Annotators could label a span of any length if it was complex as a whole, but were asked to create as short a span as possible. To calculate the average, the labels were converted to numerical values as follows: $3 \rightarrow 1$, $2 \rightarrow 0.67$, $1 \rightarrow 0.33$, $0 \rightarrow 0$. The averaging hinges on the assumption that the labels have an equal distance between them. We always presented the labels together with the values 0 to 3 to reinforce the perception of equal distance.

[7] On average, the CK annotators reported higher Japanese reading proficiency than the non-CK (see Appendix A).

[8] See Appendix B for detailed definitions of each label.

3.3 MWE Annotation

In parallel with the complexity annotation, we annotated MWEs not identified by LUW segmentation (see Section 3.1). Given the absence of an MWE detector for Japanese of sufficient quality, the annotation was performed manually by a native Japanese speaker and a non-native speaker with a degree in the Japanese language. The expression categories we consider MWEs are described in Appendix C.

3.4 Complexity Scoring

Using annotations from the previous steps, we assigned complexity scores to words according to the following rules:

(1) If a span contains one or more words, each word receives the complexity value of the span.

(2) If an MWE (manually annotated according to Section 3.3) overlaps with or contains multiple spans, the MWE receives the maximum of the complexity values of the spans.

Finally, for each word, we calculated the complexity score for each L1 group as the average of the individual values from the annotators in that group.

4 Statistics and Analysis

Overall statistics for both genres and L1 groups are shown in Table 1.[9] MWEs have higher mean complexity than single words for both L1 groups and are more frequent in the Government genre. There is a tendency towards perceiving higher complexity in the non-CK group, which corresponds to slightly lower average Japanese proficiency of the non-CK annotators (see Appendix A).

We measured inter-annotator agreement (IAA) using Krippendorf's α for interval values (Krippendorff, 1970). The IAA is 0.32 in the CK group, and 0.31 in the non-CK group, while it would be 0.19 if we merged the groups. As lexical complexity is

[9] See Appendix E for the complexity scores and annotation distributions of several words in the non-CK group.

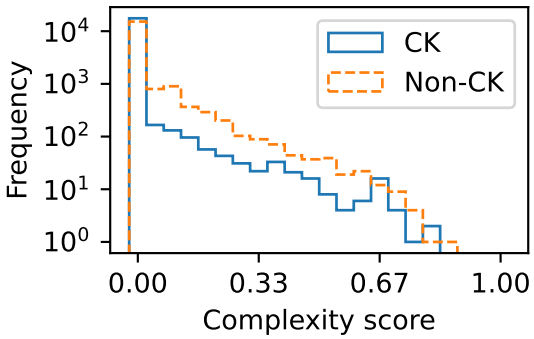

Figure 2: Histogram of complexity scores by L1 group.

	Japanese		Chinese		Other	
	All	CC	All	CC	All	CC
CK	.003	.009	.004	.004	.071	.000
Non-CK	.010	.032	.062	.072	.007	.143
Frequency	52%	10%	26%	22%	4%	0%

Table 2: Mean complexity (by L1 group) and frequency, according to (1) word origin: Japanese (*wago*), Chinese/Sino-Japanese (*kango*), and Other (*gairaigo*, borrowings from languages other than Chinese), and (2) whether the words contain Chinese characters (denoted by CC). The origin was classified using MeCab and Comainu (see Section 3.1), excluding words of mixed or unknown origin.

highly subjective (Gooding et al., 2021), the low agreement does not imply low reliability, but it indicates that perception of complexity is more alike within the L1 groups than across all annotators.

The complexity score distribution in each L1 group is shown in Figure 2. No words achieved a score greater than 0.81 and 0.86 in the CK and non-CK groups, respectively, which reflects that words are rarely labeled as Difficult or Very Difficult by all annotators in a group.

In addition to the aforementioned difference in proficiency, there is also a clear difference in how the two L1 groups perceive complexity of words based on their origin and whether they contain Chinese characters[10], as analyzed in Table 2. For the CK group, the mean complexity of words of Japanese and Chinese origin was similar. For the non-CK group, however, words of Chinese origin were markedly more complex (0.062) than words of Japanese origin (0.010), and both categories of words were more complex when they contained Chinese characters.[11]

5 Experiments

The newly created dataset can be used to evaluate performance of LCP for non-native Japanese readers of different L1 backgrounds (CK and non-CK). We developed a baseline system based on a fine-tuned BERT (Devlin et al., 2019) model, and evaluated it using cross-validation. We fine-tuned a Japanese pre-trained BERT model released by Tohoku University, namely the base model for UniDic Lite segmentation[12].

For each word w in our dataset and the sentence s that contains it at token indices i to $j-1$, we construct an input sequence (`[CLS]`, s_0^{i-1}, `<Unused1>`, w, `<Unused2>`, $s_j^{|s|-1}$, `[SEP]`, w, `[SEP]`). The target word occurs first delimited by unused tokens (`<Unused`n`>`) in the sentence context, and then on its own following the first `[SEP]` token.[13] To predict the complexity score, we feed the final hidden representation of the `[CLS]` token into a linear layer with a single output. A similar fine-tuning approach, but without the special tokens, was used for English LCP by Taya et al. (2021), achieving one of the highest R^2 values in the single-word subtask of SemEval-2021 Task 1 (Shardlow et al., 2021).

We fine-tune and evaluate models for CK and non-CK complexity separately. See Appendix G for the hyperparameters and cross-validation scheme.

The results are reported in Table 3. In addition to R^2 (coefficient of determination)[14], we report mean average error (MAE) by complexity score tiers to draw the full picture of the models' performance at different complexity levels. The score ranges of the

[10]Japanese vocabulary consists of words of Japanese origin, Chinese (Sino-Japanese) origin, and foreign words from other languages (*gairaigo*). The first two categories can be written using Chinese characters (*kanji*), Japanese syllabary (*kana*), or a combination thereof, while other foreign words are usually written in syllabary only. (See Appendix F for examples.)

[11]The opposite tendency for *gairaigo* (foreign words mostly from English) to be perceived as more complex in the CK group coincides with lower English proficiency among annotators in this group (see Appendix A), and therefore should not be explained by their L1 background.

[12]Available from `https://huggingface.co/cl-tohoku/bert-base-japanese-v2`.

[13]Due to a different segmentation (version of UniDic) used by Tohoku BERT and our dataset, we have to enforce segmentation at the word's boundaries using spaces.

[14]Compared to correlation coefficients, R^2 is more appropriate for LCP, since it also captures deviations in mean and variance. Compared to MAE or MSE, it is easier to interpret, as $R^2 = 0$ corresponds to the mean regressor, while $R^2 = 1$ corresponds to a perfect model.

	MAE by Gold Complexity Score Tier				
	Zero	Easy > 0	Not Easy	(Very) Difficult	R^2
CK	0.0034	0.0676	0.1913	0.2954	0.4351
Non-CK	0.0066	0.0510	0.1169	0.2932	0.6142

Table 3: Results of the fine-tuned BERT model by L1 group (means over 5 cross-validation folds).

	Zero	Easy > 0	Not Easy	(Very) Difficult
CK	17,563	393	223	41
Non-CK	15,209	2,067	837	107

Table 4: Word counts in the whole dataset by L1 group and MAE tier.

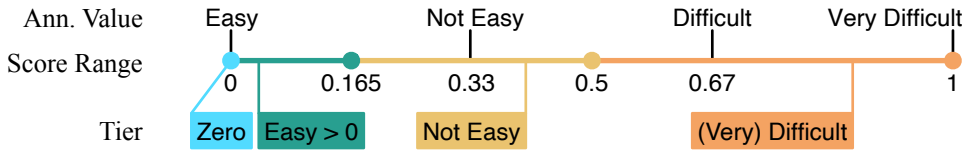

Figure 3: Illustrated score ranges of the MAE tiers: $\{0\}$ for Zero, $(0, 0.165]$ for Easy > 0, $(0.165, 0.5]$ for Not Easy, and $(0.5, 1]$ for (Very) Difficult.

tiers are centered at annotation values as illustrated in Figure 3. We handle zero as a special tier, and merge Very Difficult with Difficult due to a low number of words.

The fine-tuned BERT model for CK and Non-CK achieves R^2 of 0.4351 and 0.6142, respectively. For both L1 groups, the MAE value increases markedly in each successive complexity tier, as the number of training examples (shown in Table 4) diminishes. Similarly, the CK model achieves lower error than non-CK only in tier Zero, where it has more examples available than the non-CK model. This suggests that the scarcity of words with complexity above zero is a factor contributing to worse performance on CK data, as measured by R^2.

6 Conclusion

In this paper, we presented the first dataset for Japanese LCP. It provides separate complexity scores based on the CK/non-CK distinction of annotators' L1 background. Our analysis corroborates our conjecture that special consideration of L1 background is useful for the Japanese LCP task in particular. We believe it could benefit LCP in other languages as well.

In the baseline experiment, we demonstrated the efficacy of our BERT-based system for both CK and non-CK readers. Even after separating CK and non-CK annotators, however, notable inter-annotator disagreement remains within these groups. Therefore personalized systems analogous to Gooding and Tragut (2022) could improve on our system. Future research should study this possibility, analyzing both its costs and benefits.

Models trained on JaLeCoN can be used as part of a lexical simplification pipeline for Japanese, both to identify complex words and to rank candidate simplifications. JaLeCoN itself can be further used as a basis for a lexical simplification dataset targeting words actually perceived as complex, similar to TSAR-ST datasets for English and Spanish (Štajner et al., 2022).

Limitations

Our task setting and baseline system requires that the input is already segmented into words including MWEs. The MWE identification step in the construction process of our dataset involved time-consuming manual annotation. Building a high-quality system that fully automates the process is an issue for future work. Our dataset can be used to evaluate such a Japanese MWE identification system.

Additionally, as shown in Section 5, our baseline model performed relatively poorly in the higher complexity tiers. This is an effect of the dense annotation setting; it results in uneven distributions of

complexity as shown in Figure 2, where easy words greatly outnumber difficult words. One possible solution would be creating another LCP dataset using sparse annotation, where target words are selected using frequency bands so that the words are distributed across a wide range of frequency (Shardlow et al., 2022). Our data could provide insights as to what kind of words should be targeted by sparse annotation for such a dataset.

Acknowledgments

We would like to express our gratitude to Justin Vasselli and the anonymous reviewers for their insightful feedback. This work was supported by JSPS KAKENHI grant number JP19K20351 and NAIST Foundation.

References

Yasuharu Den, Toshinobu Ogiso, Hideki Ogura, Atsushi Yamada, Nobuaki Menematsu, Kiyotaka Uchimoto, and Hanae Koiso. 2007. The development of an electronic dictionary for morphological analysis and its application to japanese corpus linguistics. *Japanese Linguistics*, 22(5):101–123.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sian Gooding and Ekaterina Kochmar. 2019. Complex word identification as a sequence labelling task. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1148–1153, Florence, Italy. Association for Computational Linguistics.

Sian Gooding, Ekaterina Kochmar, Seid Muhie Yimam, and Chris Biemann. 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*, pages 4439–4449, Online. Association for Computational Linguistics.

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

Marcella Hu and I.S.P Nation. 2000. Unknown Vocabulary Density and Reading Comprehension. *Reading in a Foreign Language*, 13(1):403–30.

Ekaterina Kochmar, Sian Gooding, and Matthew Shardlow. 2020. Detecting multiword expression type helps lexical complexity assessment. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4426–4435, Marseille, France. European Language Resources Association.

Tom Kocmi, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Thamme Gowda, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Rebecca Knowles, Philipp Koehn, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Michal Novák, Martin Popel, and Maja Popović. 2022. Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 1–45, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

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

Kazuko Komori, Junko Mikuni, and Kondoh Atsuko. 2004. Bunshō rikai o sokushin suru goi chishiki no ryōteki sokumen : Kichigo ritsu no ikichi tansaku no kokoromi [What percentage of known words in a text facilities reading comprehension? : A Case Study for Exploration of the Threshold of Known Words] (in Japanese). *Nihongo Kyōiku [Journal of Japanese Language Teaching]*, 120:83–92.

Shunsuke Kozawa, Uchimoto Kiyotaka, and Yasuharu Den. 2014. Adaptation of long-unit-word analysis system to different part-of-speech tagset. *Journal of Natural Language Processing*, 21(2):379–401.

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

Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. 2004. Applying conditional random fields to japanese morphological analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 230–237, Barcelona, Spain. Association for Computational Linguistics.

Suguru Matsuyoshi, Satoshi Sato, and Takehito Utsuro. 2007. A dictionary of japanese functional expressions with hierarchical organization. *Journal of Natural Language Processing*, 14(5):123–146.

Daiki Nishihara and Tomoyuki Kajiwara. 2020. Word complexity estimation for Japanese lexical simplification. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3114–3120, Marseille, France. European Language Resources Association.

Hideki Ogura, Hanae Koiso, Yumi Fujiike, Sayaka Miyauchi, Hikari Konishi, and Yutaka Hara. 2011. Gendai nihongo kakikotoba kinkō kōpasu keitairon kiteishū dai 4 ban jō [Regulations of morphological information for balanced corpus of contemporary

written Japanese 4th edition volume 1] (in Japanese). *NINJAL Internal Reports*.

Mai Omura, Aya Wakasa, and Masayuki Asahara. 2021. Word delimitation issues in UD Japanese. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 142–150, Sofia, Bulgaria. Association for Computational Linguistics.

Paula J Schwanenflugel, Katherine Kip Harnishfeger, and Randall W Stowe. 1988. Context availability and lexical decisions for abstract and concrete words. *J. Mem. Lang.*, 27(5):499–520.

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

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

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

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Front Artif Intell*, 5:991242.

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):97.

Yuka Tateisi, Yoshihiko Ono, and Hisao Yamada. 1988. A computer readability formula of japanese texts for machine scoring. In *Coling Budapest 1988 Volume 2: International Conference on Computational Linguistics*.

Yuki Taya, Lis Kanashiro Pereira, Fei Cheng, and Ichiro Kobayashi. 2021. OCHADAI-KYOTO at SemEval-2021 task 1: Enhancing model generalization and robustness for lexical complexity prediction. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 17–23, Online. Association for Computational Linguistics.

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

A Annotators

	Japanese		English	
	B1/B2	C1/C2	B1/B2	C1/C2
CK	4	3	7	0
Non-CK	6	1	2–3	4–5

Table 5: Annotator counts per sentence in each L1 group, by Japanese and English reading proficiency category. The proficiency levels were determined by self-reports with reference to an assessment grid either in Japanese[15] or in English.[16] Overall, our CK annotators are better at Japanese reading and poorer at English reading than the non-CK.

CK	Chinese: 6,	Korean: 1		
Non-CK	English: 2–3,	Thai: 2–3,	Indonesian: 1,	Lao: 1

Table 6: Annotator counts per sentence of each L1 group.

B Complexity labels

3 (Very Difficult):	You hardly understand its meaning in the context.
2 (Difficult):	You can infer its meaning, but you are not confident.
1 (Not Easy):	You understand its meaning with confidence, but it is quite difficult among the expressions you can understand.
0 (Easy):	None of the above.

Table 7: Complexity labels. An annotator can label spans with complexity 3, 2, 1, or 0.

[15] https://jfstandard.jp/pdf/self_assessment_jp.pdf
[16] https://rm.coe.int/CoERMPublicCommonSearchServices/DisplayDCTMContent?documentId=090000168045bb52

C MWE Categories

Category	Description	Example
Lexicalized expressions	Non-compositional expressions whose meaning as a whole cannot be completely inferred by the meaning of their components.	使い | 勝手 (ease of use)
Institutionalized expressions	Compositional expressions whose components cannot be replaced without distorting the meaning of the whole expression or violating the language conventions.	感染 | 症 (infectious disease)
Functional expressions	Expressions that behave like single function words.	に | つき | まし | て (as for)

Table 8: Categories we regard as MWEs. See Kochmar et al. (2020) for lexicalized and institutionalized expressions, and Matsuyoshi et al. (2007) for functional expressions. The vertical bars in the examples denote boundaries between SUWs.

D Excluded categories

Category	Identification Approach	Example
Proper nouns	Proper nouns are first identified by MeCab. We also manually annotate proper noun phrases.	関東 | 大 | 震災 (The Great Kantō Earthquake)
Segmentation errors	We manually annotate sequences with segmentation errors.	も | や (mist)

Table 9: Categories of words or spans we exclude from our target. The vertical bars in the examples denote boundaries between SUWs. The correct segmentation for も \| や is もや.

E Distributions of Annotation

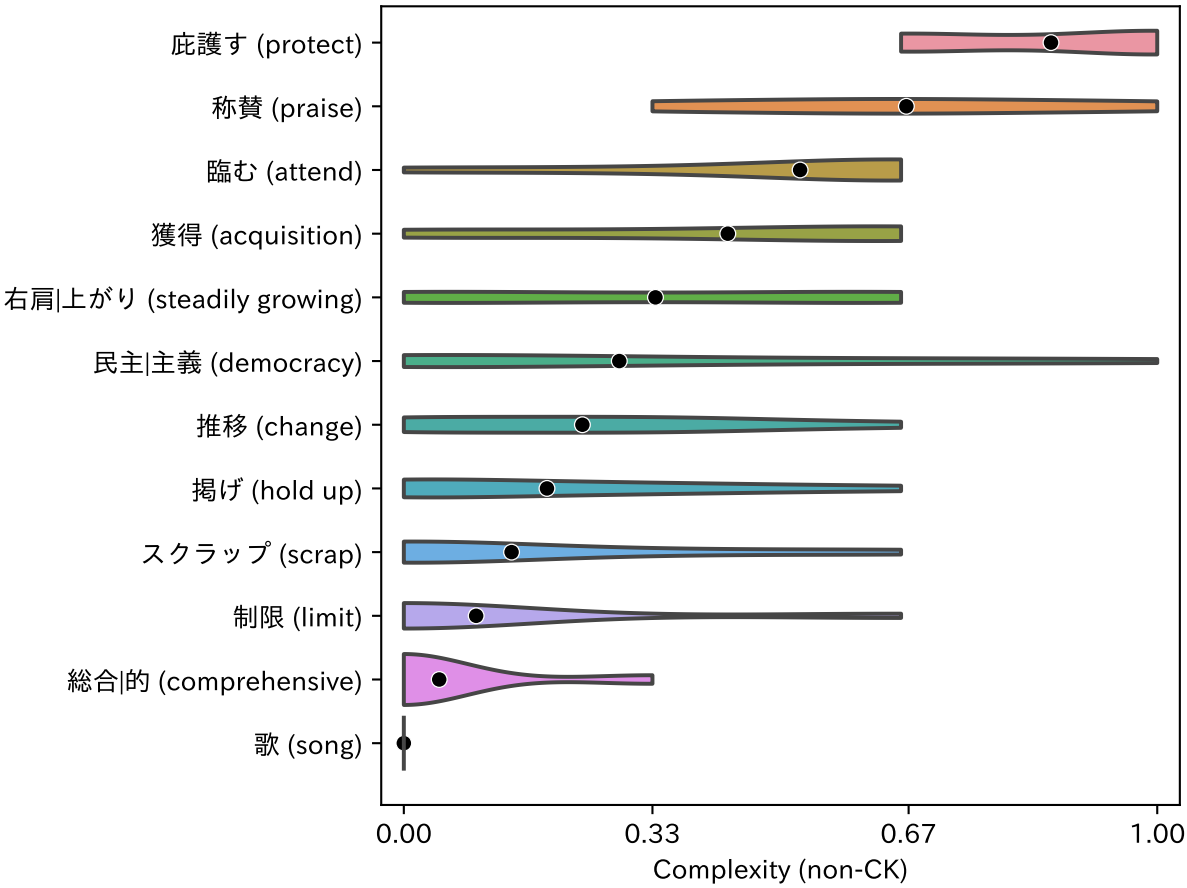

Figure 4: Violin plot showing annotation distributions of several words with dot markers showing the complexity scores, both for non-CK annotators. Words are shown in their surface forms; the vertical bars in them denote boundaries between SUWs.

F Examples of Words by Origin

Origin	Containing Chinese characters (*kanji*)?	
	Yes	No
Japanese (*wago*)	歌 (song) 臨む (attend)	けれど|も (although) ふさわしい (suitable)
Chinese/Sino-Japanese (*kango*)	今回 (this time) 民主|主義 (democracy)	よう (it seems †様) もちろん (of course †勿論)
Other (*gairaigo*)	旦那 (husband <Skt)	スクラップ (scrap <Eng) ホーム|ページ (home page <Eng)

Table 10: Examples of words in JaLeCoN categorized by word origin and whether they contain Chinese characters. † marks a variant of the word written using Chinese characters documenting the Sino-Japanese origin; < marks the word's origin (Sanskrit or English). The vertical bars in the examples denote boundaries between SUWs. All categories except Other (*gairaigo*) written using Chinese characters are relatively common.

G Experimental Setting

Optimizer	Adam ($\beta_1 = 0.9$, $\beta_2 = 0.999$)
– learning rate	5e-5
– schedule	no warm-up, linear decay
– L2 weight decay	0.01
Epochs	5
Loss function	Mean squared error
Dropout	0.1
Batch size	16
Weight initialization	$\mathcal{N}(\mu = 0, \sigma = 0.02)$ truncated to $\pm 2\sigma$
Bias initialization	0
Gradient L2 norm clipping	2

Table 11: Hyperparameters used for fine-tuning the BERT model. We have chosen the combination of learning rate (from 8e-6, 5e-5, 3e-5, and 2e-5), warm-up (from no warm-up and 10% steps), and the number of epochs (from 1 to 5) achieving the highest mean R^2 in a nested 4-fold cross-validation on the training data of the first outer cross-validation split. The optimal combination was identical for CK and non-CK complexity.

Folds	5
Stratification	by genre (News and Government)
Grouping	by sequence of sentences (see Section 3)

Table 12: Cross-validation scheme used for fine-tuning and evaluation of the BERT model.