



## Japanese readability assessment using machine learning

Tyler Ivie<sup>a</sup> and Robert Reynolds<sup>b</sup>

<sup>a</sup>Department of Linguistics, Brigham Young University, , [tyler.j.ivie@gmail.com](mailto:tyler.j.ivie@gmail.com) and <sup>b</sup>Office of Digital Humanities, Brigham Young University, , [robert\\_reynolds@byu.edu](mailto:robert_reynolds@byu.edu)

How to cite: Ivie, T.; Reynolds, R. (2023). Japanese readability assessment using machine learning. In *CALL for all Languages - EUROCALL 2023 Short Papers*. 15-18 August 2023, University of Iceland, Reykjavik. <https://doi.org/10.4995/EuroCALL2023.2023.16989>

---

### Abstract

*We present a new corpus of Japanese texts, labeled according to six second-language readability levels. We also show the results of experiments training machine-learning classifiers to automatically label new texts according to reading level. The resulting models can be used in language-learning websites and applications to enhance Japanese language learning. The best-performing model, Random Forest, achieved an F1 score of 0.86, with an adjacent accuracy of 0.97. Of the 114 features used, we identify a small subset of five features that are sufficient to achieve an F1 score of 0.74. The corpus, code, and resulting models are free and open-source.<sup>1</sup>*

**Keywords:** readability, machine learning, Japanese.

---

## 1. Introduction

The goal of assigning reading levels to texts has been approached through a variety of methods. Readability metrics and subsequent classifications have been produced using simple formulas, such as syllable to word ratio as seen with the Flesch Reading Ease scale (Flesch, 1948), statistical analysis (e.g. Lee & Hasebe, 2016), and machine-learning (e.g. Hancke et al., 2012).

All methods have their unique merit, and the results of each are distinct. Formulaic approaches are simple enough to be computed by hand, providing a helpful insight and often being more language-agnostic (Bendová & Cinková, 2021). The Flesch scale, for example, was used for predicting the readability of Czech. Statistical approaches allow far more factors to be considered without being too cumbersome and can produce very useful results, albeit statistical models require more expertise to execute than a formula, but they can be automated and even made available online (Hasebe & Lee, 2015).

Machine-learning approaches are much like statistical models but can be fed additional features to consider, giving them the potential to be even more versatile, whereas formulaic approaches like the Flesch scale, considering the least amount of features, are most easily tricked. Many formulas have been created for the English readability assessment over the years to overcome genre-specific and demographic-specific classification challenges (Klare, 1974), where machine-learning models, considering more features, can perform well in more varied contexts. Traditional machine learning models cannot easily be adapted between languages, however, and are time-consuming to create.

---

<sup>1</sup>[https://github.com/reynoldsnlp/japanese\\_readability\\_corpus](https://github.com/reynoldsnlp/japanese_readability_corpus)

In our search we were not able to find examples where machine-learning was implemented to classify the readability of Japanese documents for L2 readers. Most current literature regarding the automated reading level assessment of Japanese leverages statistical models involving features carefully selected from large corpora (Lee & Hasebe, 2016), often using features like the proportion of different Japanese scripts and length of sentences. There is also record of non-automated, social approaches to document classification, where different versions of the same document are given to test-groups of native speakers to collect human-reported difficulty metrics (Sakai, 2011).

It is our aim to help expand resources available to Japanese language learners and professionals with machine-learning readability classification using open-source software and freely available corpora. The code used to obtain and analyze our corpus will be included in links to our Github repository along with a simple script to predict the grade level of texts from outside of the training dataset. Most of the corpus used in this paper is in the creative commons and is provided in our codebase, but for one to replicate the results of the study, part of the corpus will need to be downloaded from the publisher.

## **2. Method**

The labeled corpus was created by collecting all of the freely available graded readers from two websites: tadoku.org and jgrpg-sakura.com. Each document was given a level 0-6, 0 being the simplest and 6 being the most difficult. Levels 1-5 correspond to JLPT levels 5-1, whereas level 0 in our corpus represents a below JLPT N5 level reader. A sampling of academic articles and news stories were also added manually. Image count was also manually collected but was not considered in the final machine-learning model.

Document text from tadoku.org was stripped from graded-reader PDFs. Furigana, pronunciation guide characters, were removed systematically by comparing font size ratios. Text was manually cleaned and audited to ensure at least 98% accuracy. Image-only PDFs were digitized using Tesseract optical character recognition. In a few cases purely manual work was necessary because of poor implementation of vertical text in some file formats.

Documents from jgrpg-sakura.com were downloaded in .html format and parsed, removing `<ruby></ruby>` tags to ensure no furigana were swaying the data. These texts were also manually reviewed but required less rigorous auditing due to the less extensive and non-manual edits made to the documents. Because jgrpg-sakura.com uses a 9-level system as opposed to the 6-level system used by jreadability.net and tadoku.org, individual documents from jgrpg-sakura.com were assigned a 6-level equivalent beforehand, following a brief manual review.

Academic articles and news were assumed to be native-level and were assigned level 5. They received the least amount of processing, containing no furigana. Characters that triggered unresolved escaped in, however, namely backslashes and ascii control characters were removed. The final document count in this small labeled corpus is 167.

For machine learning, we extracted 114 linguistic features from each document. Stanza (Qi et al., 2020), a python library for natural-language-processing was used for part-of-speech and basic dependency tagging. The features extracted consist of basic orthographic features (e.g. ratio of hiragana, katakana, and kanji): vocabulary and frequency features (e.g. proportion of words at a certain level on the Japanese Language Proficiency Test (JLPT)); and grammatical features (e.g. particle count, certain conjugations, and a sampling of specific morpho-syntactic patterns).

Token frequency lists were taken from <http://corpus.leeds.ac.uk/frqc/internet-jp.num> and kanji frequency lists were sourced from <https://scriptin.github.io/kanji-frequency/wikipedia/>. Since official JLPT vocabulary lists

have not been posted since the Japanese government added a fifth level, JLPT N3 in 2010,<sup>2</sup> vocabulary level lists were sourced from <https://tangorin.com/vocabulary/>, one of many informal but researched resources for assigning JLPT levels to Japanese vocabulary.

After features were extracted, we trained models using 5-fold cross-validation, including the following algorithms from Scikit Learn (Pedregosa et al, 2011) and xgboost (Chen & Guestrin, 2016): XGBoost (XGB), XGBoost Random Forest (XGBRF), Logistic Regression (LR), Linear Discriminant Analysis (LDA), K-Nearest Neighbor (KNN), Classification And Regression Trees (CART), Naïve Bayes (NB), and Support Vector Machine (SVM).

Later, importances were generated and used in tandem with recursive feature elimination in another Random Forest model to track the performance of the corpus RF model with and without different features in order to determine the features that were most important.

### 3. Results

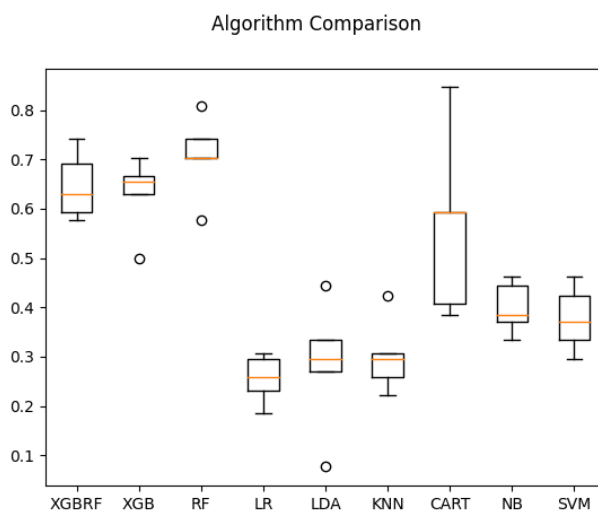


Figure 1. F1 scores on 5-fold cross-validation.

As seen in Figure 1, the Random Forests classifier was found to consistently outperform other models, so further evaluation focuses on this model. A full classification report for the Random Forest model is given in Table 1. Precision measures the accuracy of positive predictions, while recall measures the completeness of positive predictions. The F1-score is the harmonic mean of precision and recall. Table 1 shows that the average F1 score for all levels is 0.86. The lowest performance was for levels 1 and 2.

<sup>2</sup> <https://www.jlpt.jp/e/topics/list2010.html>

	Precision	Recall	F1-score	Support
Level 0	0.88	0.88	0.88	8
Level 1	0.57	0.80	0.67	5
Level 2	1.00	0.60	0.75	5
Level 3	0.86	0.86	0.86	7
Level 4	1.00	1.00	1.00	4
Level 5	1.00	1.00	1.00	5
Accuracy	---	---	0.85	34
Macro Avg	0.88	0.86	0.86	34
Weighted Avg	0.88	0.85	0.86	34

**Table 1.** Classification report for Random Forest model on 5-fold cross-validation

The confusion matrix in Table 2 shows where individual predictions of the model relate to their actual readability levels. Gray cells along the diagonal are correct predictions. This shows that almost all of the Random Forest model’s mistakes are only off by one level, yielding an adjacent accuracy score of 0.97. This indicates that the model is not only making good predictions in general, but that even its mistakes are reasonably close to the actual readability level.

		Predicted Level					
		Lv 0	Lv 1	Lv 2	Lv 3	Lv 4	Lv 5
Actual Level	Lv 0	7	1	-	-	-	-
	Lv 1	1	4	-	-	-	-
	Lv 2	-	1	3	1	-	-
	Lv 3	-	1	-	6	-	-
	Lv 4	-	-	-	-	4	-
	Lv 5	-	-	-	-	-	5

**Table 2.** Confusion matrix for Random Forest model on 5-fold cross-validation

To evaluate which features are most important to the model, we used Recursive Feature Elimination (RFE) to rank features. Then, to determine the minimum number of features needed for a viable model, we iteratively trained models on more and more features, using the order from RFE.

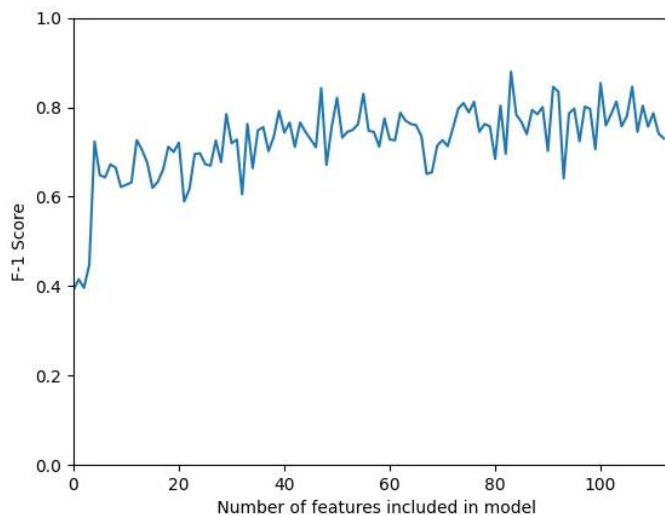


Figure 2. F1 score of models trained with top-N number of features, ranked by importance

The results in Figure 2 show that reasonable performance can be achieved with only five features: type-token ratio, type-lemma ratio, kanji type-token ratio, unique verbs to token ratio, and mean sentence length. These five features represent three different categories of features: lexical variation, orthographic variation, and syntactic complexity.

#### 4. Conclusions

We have collected a small second-language readability corpus of Japanese with 34 documents labeled for six readability levels. We also trained a Random Forest classifier that achieves an average F1 score of 0.86 on cross-validation. This classifier can be implemented in websites and applications to support Japanese language learning.

This study is limited by a small corpus size, but the consistent performance of the Random Forest model, especially its near-perfect adjacent accuracy, suggests that the textual features included in the model are valid identifiers of readability at these six reading levels. In particular, the five features with the highest importance are crucial to identifying the readability of Japanese texts: type-token ratio, type-lemma ratio, kanji type-token ratio, unique verbs to token ratio, and mean sentence length. Especially noteworthy is the importance of orthographic variation, which is almost completely absent from readability research with other languages.

Future work is needed to build a larger corpus to increase confidence in the findings reported here. Although many commercial resources exist, we intend to focus on adding texts with licenses that allow publishing the corpus with an open license.

To our knowledge, this is the first published research using a machine-learning approach to automate Japanese readability classification. Despite limited resources, the results are quite promising, and future work in this domain is likely to see significant gains in automated readability classification of Japanese texts.

#### Acknowledgements

We are grateful to Brigham Young University's College of Humanities for funding this research.

## References

- Bendová, K., & Cinková, S. (2021). Adaptation of classic readability metrics to czech. Paper presented at the *International Conference on Text, Speech, and Dialogue*, 159-171.
- Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. Paper presented at the *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785-794.
- Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology*, 32(3), 221-233. <https://doi.org/10.1037/h0057532>
- Hancke, J., Vajjala, S., & Meurers, D. (2012). Readability classification for german using lexical, syntactic, and morphological features. Paper presented at the *Proceedings of COLING 2012*, 1063-1080.
- Hasebe, Y., & Lee, J. (2015). Introducing a readability evaluation system for japanese language education. Paper presented at the *Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese*, 19-22.
- Klare, G. R. (1974). Assessing readability. *Reading Research Quarterly*, , 62-102.
- Lee, J., & Hasebe, Y. (2016). Readability measurement of japanese texts based on levelled corpora. *The Japanese Language from an empirical perspective*, 143.
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12, 2825-2830.
- Qi, P., Zhang, Y., Zhang, Y., Bolton, J., & Manning, C. D. (2020, July). Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101-108).
- Reynolds, R. (2016). Insights from russian second language readability classification: Complexity-dependent training requirements, and feature evaluation of multiple categories. Paper presented at the *Proceedings of the 11th Workshop on Innovative use of NLP for Building Educational Applications*, 289-300.
- Sakai, Y. (2011). Improvement and evaluation of readability of Japanese health information texts: an experiment on the ease of reading and understanding written texts on disease. *Library and information science*, (65), 1-35.