EUROCALL 2023: CALL for all Languages
15-18 August 2023, University of Iceland, Reykjavik
DOI: https://doi.org/10.4995/EuroCALL2023.2023.16991

A machine-learning approach to Czech readability

Peter Williams[a] and Robert Reynolds[b]

[a]Department of Linguistics, Brigham Young University, , peterjwms@gmail.com and [b]Office of Digital Humanities, Brigham Young University, , robert_reynolds@byu.edu

How to cite: Williams, P.; Reynolds, R. (2023). A machine-learning approach to Czech readability. In *CALL for all Languages - EUROCALL 2023 Short Papers.* 15-18 August 2023, University of Iceland, Reykjavik. https://doi.org/10.4995/EuroCALL2023.2023.16991

Abstract

We present a new corpus of Czech texts labeled for second-language readability, and show results of experiments to train machine-learning classifiers to automatically label new texts according to reading level. We report results comparing the performance of traditional machine-learning models (including Random Forest, XGBoost, Linear Discriminant Analysis, and XGBoost Random Forest) and a neural network (XLM-RoBERTa). The results of our research can be implemented in tools to support learning Czech, a less commonly taught language. We extract 46 linguistic features in various categories for use with traditional machine-learning algorithms. We train models on these features and evaluate their performance with recursive feature elimination to determine how informative each feature is for each model. We then compare those results to those of a transformer trained for the same task on the same corpus. XGBoost achieves the highest accuracy at 0.81, suggesting that these traditional models can still perform as well as, or better, than newer models on this task. Notably, the transformer has the lowest mean F1 at 0.74.[1]

Keywords: *readability, machine learning, Czech, transformer, corpus.*

1. Introduction

Traditional approaches to readability assessment have used formulae based on average word length and average sentence length. Although these features are useful, they represent an impoverished view of readability. Recent years have seen a marked increase in interest in research surrounding machine-learning approaches to second-language readability, with many studies focused on commonly taught languages, such as English (Vajjala & Meurers, 2012; Xu et al., 2015; Xia et al., 2016; Vajjala & Lučić, 2018), German (Hancke et al., 2012), Russian (Reynolds, 2016), French (Lee & Vajjala, 2022), Italian (Azpiazu & Pera, 2019), and Spanish (Vásquez-Rodríguez et al., 2022). In addition to studying the readability of individual languages, some researchers have worked to identify language-agnostic properties that can be used for multilingual readability classification (Azpiazu & Pera, 2019).

As research in this field continues, it expands both in methods and languages. Some research has been done recently to adapt classic readability metrics to Czech, including the Flesch Reading Ease, Flesch-Kincaid Grade Level, Coleman-Liau Index, and the Automatic Readability Index (Bendová & Cinková, 2021). In addition to traditional machine-learning models, neural networks have also been applied to the readability assessment problem in English and Slovenian, and recently achieved similar levels of success as traditional machine-

[1] Code available at https://github.com/peterjwms/czech-readability.

learning approaches (Deutsch et al., 2020; Martinc et al., 2021; Lee et al., 2021). In order to push the limits of both traditional machine-learning approaches with linguistic features and neural approaches, a combination of linguistic features with a neural network has been used with similar results (Lee et al., 2021).

We contribute to these threads of research in two ways. First, our study is the first to apply machine-learning methods to readability in Czech, a less commonly taught language which is typologically distinct from most languages already studied in the readability literature. Second, a labeled second-language readability corpus in any new language represents a contribution toward efforts to build multilingual readability models.

Based on this corpus, we train a number of machine-learning classifiers to label texts according to readability level. On the one hand, we are interested in maximizing classification accuracy, but we also use the models to help determine which document features are most informative for this task. To this end, we use both neural and traditional machine-learning approaches. For each document, we use natural language processing to extract features in the following categories: Lexical Variability, Lexical Familiarity, Morphology, and Syntax.

First, we use these features to create a text classification model using traditional machine-learning algorithms. Using the coefficients/importances of these models, we evaluate our features, identifying which features are most informative for this task. Second, we compare these results against those of a transformer. As the first study of Czech L2 readability using machine learning, this study lays the groundwork for both theoretical and technical research supporting the learning of a less commonly taught language.

2. Method

2.1. Corpus collection

Our corpus comes from two sources. First, we collected graded readers with parallel translations of Czech and other languages labeled with CEFR readability level of either A1-A2 or B1-B2 by the publisher. Because these labels span sublevels 1-2, we simplify the labels to ’A’ and ’B’. Each of these books came from the Czech publisher Edika[2], part of Albatros Media. Each chapter was treated as its own document in order to normalize the lengths of each document and increase the number of documents in the training corpus. Second, we filled the gap of C-level text using Czech news articles from the Czech Text Document Corpus 2.0 (CTDC) (Kral & Lenc, 2017). The CTDC includes newspaper articles from the Czech News Agency labeled topically at the document level. We assume that all documents from CTDC are at the C reading level. We randomly selected a subsection of the CTDC to have a mostly balanced number of texts for each level, with a variety of topics, themes, and styles present in each level. Our corpus has 229 A-level documents, 209 B-level documents, and 230 C-level documents, with a total of 426,834 tokens. The number and level of classes were determined by their availability in our dataset.

2.2. Feature extraction

In total, we extracted 46 features split into several categories: lexical variability, lexical familiarity, morphology, and syntax. Lexical variability includes features like type-token ratio. Lexical familiarity includes word and lemma frequencies and ranks, obtained from the corpus SYN2015 from the Czech National Corpus[3] (Český národní korpus, 2016). Morphology includes the average number of certain parts of speech per sentence, and part-of-speech to token ratios. Syntax includes mean and max dependency lengths for various parts of speech. These features were extracted for each chapter primarily using Stanza (Qi et al., 2020).

Additionally, we used several classic measures of readability, with formulas adapted to Czech by Bendová and Cinková (2021). These formulas include adjustments to the coefficients of the Flesch Reading Ease, Flesch-

[2] https://www.edika.cz/

[3] https://wiki.korpus.cz/doku.php/seznamy:srovnavaci_seznamy

Kincaid Grade Level, Automatic Readability Index, and the Coleman-Liau Index. The function to count Czech syllables for the purpose of these readability metrics was adapted from Bendová[4] (2021).

2.3. Choice of models

For the machine-learning models, we used Linear Discriminant Analysis (LDA) and Random Forest (RF) models from scikit-learn (Pedregosa et al., 2011), and XGBoost (XGB) and XGBoost Random Forest (XGBRF) models from XGBoost (Chen & Guestrin, 2016). For the transformer, we used XLM-RoBERTa (base size) from HuggingFace as the pre-trained model on which we fine-tuned (Conneau et al., 2019).

2.4. Folds

The chapters in each book are likely to share similar vocabulary, grammar, and style, so we structured our cross-validation folds to avoid having chapters from the same book in both the training and validation. We grouped the texts in our corpus according to the book that they originally belonged to, such that all the texts belonging to the B-level 'Jana Eyrova' are part of group 2, and a random selection of articles from the CTDC are group 33, and so on. Using these groups, we manually created folds for cross-validation that combined one group each of A, B, and C texts. One fold was kept out as a test set. The rest of the data was used for cross-validation. In this way, we were able to control for specific variations in style and content that would affect reading level. This method of cross-validation was applied to both the traditional machine-learning models as well as the transformer.

We also performed recursive feature elimination on each model, and then trained the models again by adding the next most informative feature for that model each loop. Each model was trained with the same folds and cross-validation method as before.

3. Results

Table 1. Performance metrics on cross-validation

Model	Accuracy	Mean precision	Mean recall	Mean F1
LDA	0.77	0.91	0.77	0.80
RF	0.80	0.90	0.80	0.83
XGB	**0.83**	0.91	0.83	**0.85**
XGBRF	0.78	0.89	0.78	0.80
XLM-RoBERTa	0.75	0.78	0.74	0.74

The results from training each model are shown in Table 1. F1 is the harmonic mean of precision and recall. These standard metrics are calculated as a weighted average of the performance of each fold on the validation set, and then a macro average was taken of those to find an overall score for each model. Across all metrics, XGB achieves the best performance, with a mean accuracy of 0.83 and mean F1 of 0.85. XLM-RoBERTa performs slightly worse on all metrics than all of the traditional machine-learning models. Notably, each traditional model has a much higher precision than recall, which means that the models are returning very few false positives, but are all missing some of the true positives.

[4] https://github.com/vanickovak/ReadabilityFormula/tree/main

When performing recursive feature elimination on the results of each model, we see that each model reaches similar performance, but using different combinations and numbers of features to reach peak performance, as is visible in Figure 1. RF reaches peak performance with only 12 features with peak mean F1 score of 0.84. XGBRF is close behind with a peak mean F1 of 0.82 at 20 features. With similarly few features, LDA and XGB have F1 scores within 0.05, and then both peak much later with 31 and 46 features at 0.82 and 0.84.

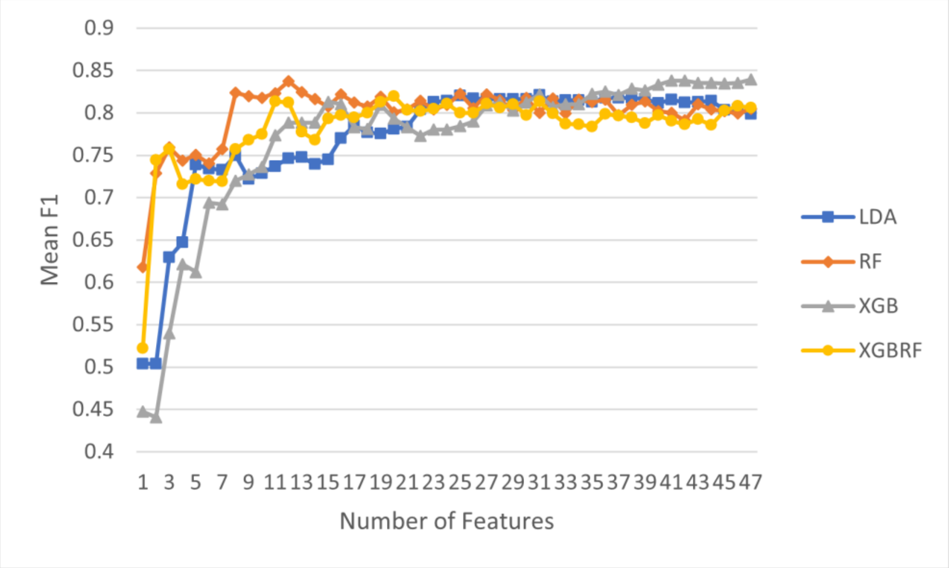

Figure 1. Model performance by number of features

We see that of the ten most informative features for XGB, RF, and XGBRF, many are the same for all three, including: mean lemma frequency, pronoun-token ratio, mean number of nouns per sentence, mean number of indicative verbs per sentence, mean number of adpositions per sentence, mean dependency length, Automatic Readability Index, and Coleman-Liau Index.

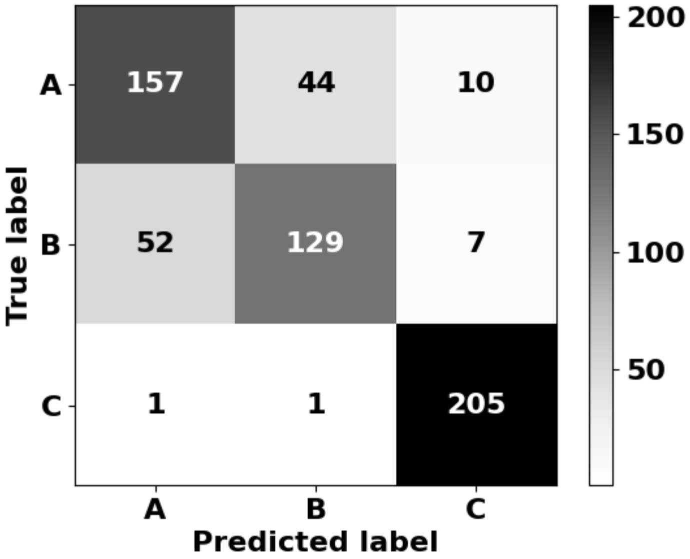

Figure 2. XGBoost cross-validation confusion matrix

4. Discussion

Our results suggest that for this problem, XGB is the best-performing model, especially considering the higher mean F1 score and consistently high performance in standard training, and when considering feature importances. RF performs comparably when considering feature importances and training the model to use as few resources as possible. The two random forest architectures perform better on low resources than our other models, as expected, although XGB outperforms them both with more data. The consistently lower results from XLM-RoBERTa suggest that the traditional models with handcrafted features are currently better-suited for this task in Czech. We could potentially improve performance by fine-tuning a different multilingual or Czech-specific model, or by augmenting the transformer architecture with linguistic features.

The confusion matrix in Figure 2 exhibits a pattern seen in all of our models: the majority of the errors confuse A and B texts. To ensure that these errors were not due to a bad assumption that all of the news documents are level C, we trained binary classifiers with only documents at levels A and B. Results were similar, which suggests that either some of the A- and B-level texts are mislabeled, or that we have failed to identify features that distinguish between these levels. Additionally, when certain books are used as the validation set, all of the models perform much worse, suggesting that these books in particular might be poorly labeled, or otherwise unique enough in their features that our models were not able to account for them. These errors might have been mitigated by collection of a larger or more varied corpus, or by confirming that each text was labeled accurately.

From our recursive feature elimination, we notice that the most informative features for XGB, XGBRF, and RF span all of our categories of features, including readability metrics, suggesting that none of the categories is considerably less informative. LDA, on the other hand, overlaps with the other three only in its use of pronoun-token ratio, and otherwise heavily relies on part-of-speech to token ratios and word length metrics.

5. Conclusions

We trained traditional and neural machine-learning models for labeling Czech documents according to three readability levels. XGBoost achieved an accuracy of 0.83. Further work can focus on increasing the size and accuracy of the labeled corpus, and testing other models, including a hybrid architecture that combines linguistic features with transformers, as demonstrated on English by Lee et al. (2021).

Acknowledgements

We would like to thank the BYU College of Humanities for their assistance in funding the corpus collection and travel to the conference.

References

Azpiazu, I. M., & Pera, M. S. (2019). Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics, 7*, 421-436.

Bendová, K. (2021). Using a parallel corpus to adapt the Flesch Reading Ease formula to Czech. *Jazykovedný časopis, 72*(2), 477-487.

Bendová, K., & Cinková, S. (2021, August). Adaptation of classic readability metrics to Czech. In *International Conference on Text, Speech, and Dialogue* (pp. 159-171). Cham: Springer International Publishing.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... & Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Český národní korpus: *Srovnávací frekvenční seznamy*. Ústav Českého národního korpusu FF UK, Praha 2016. Dostupné z WWW: http://www.korpus.cz

Deutsch, T., Jasbi, M., & Shieber, S. (2020). Linguistic features for readability assessment. *arXiv preprint arXiv:2006.00377*.

Hancke, J., Vajjala, S., & Meurers, D. (2012, December). Readability classification for German using lexical, syntactic, and morphological features. In *Proceedings of COLING 2012* (pp. 1063-1080).

Lee, B. W., Jang, Y. S., & Lee, J. H. J. (2021). Pushing on text readability assessment: A transformer meets handcrafted linguistic features. *arXiv preprint arXiv:2109.12258*.

Lee, J., & Vajjala, S. (2022). A neural pairwise ranking model for readability assessment. *arXiv preprint arXiv:2203.07450*.

Král, P., & Lenc, L. (2017). Czech text document corpus v 2.0. *arXiv preprint arXiv:1710.02365*.

Martinc, M., Pollak, S., & Robnik-Šikonja, M. (2021). Supervised and unsupervised neural approaches to text readability. *Computational Linguistics, 47*(1), 141-179.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (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). Stanza: A Python natural language processing toolkit for many human languages. *arXiv preprint arXiv:2003.07082*.

Reynolds, R. (2016, June). Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 289-300).

Vajjala, S., & Lučić, I. (2018, June). OneStopEnglish corpus: A new corpus for automatic readability assessment and text simplification. In *Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications* (pp. 297-304).

Vajjala, S., & Meurers, D. (2012, June). On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the seventh workshop on building educational applications using NLP* (pp. 163-173).

Vásquez-Rodríguez, L., Cuenca-Jiménez, P. M., Morales-Esquivel, S., & Alva-Manchego, F. (2022, December). A Benchmark for Neural Readability Assessment of Texts in Spanish. In *Proceedings of the Workshop on Text Simplification, Accessibility, and Readability (TSAR-2022)* (pp. 188-198).

Xia, M., Kochmar, E., & Briscoe, T. (2019). Text readability assessment for second language learners. *arXiv preprint arXiv:1906.07580*.

Xu, W., Callison-Burch, C., & Napoles, C. (2015). Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics, 3*, 283-297.