Translation as Augmentation: Effect of Translated Data on Assessment of Difficulty

Yiheng Wu, Jue Hou, Roman Yangarber
University of Helsinki, Finland
first.last@helsinki.fi

Abstract

Reliable Text Difficulty Assessment is a prerequisite for valid text simplification workflows and personalized learning applications. However, the development of robust assessment models is severely hindered by a critical bottleneck: the scarcity of expert-annotated corpora containing fine-grained difficulty levels (e.g., CEFR), particularly for lower-resource languages. This paper addresses this data scarcity problem in the context of a low-resource European language. We propose a cross-lingual data augmentation strategy that leverages machine translation to transfer labeled resources from high-resource languages to the target low-resource language. We train BERT-based regression models to predict difficulty scores and investigate whether synthetic, translated data can effectively supplement native training sets. Our experiments demonstrate that augmenting scarce native data with machine-translated corpora significantly improves the accuracy of difficulty estimation, offering a viable solution for languages lacking extensive expert annotations.

1. Introduction

Assessment of text complexity—Difficulty Assessment—has become increasingly important for accessible communication practice, driven by policy mandates in the United States[1] and the European Union[2]. It also plays an essential role in personalized second-language (L2) instruction (Nahatame and Yamaguchi, 2026). Compliance with these mandates and the personalization of L2 instruction both hinge on this shared technical bottleneck: the reliable estimation of text difficulty. We argue that difficulty assessment is a prerequisite for the task of automatic text simplification: a simplification system cannot be meaningfully guided or evaluated without reliable metrics to determine whether its output meets the target difficulty specification.

Consequently, this paper focuses on the assessment problem. We model this problem as a regression task, where a *difficulty model* predicts a continuous score, which can then be mapped to the CEFR scale.[3] Such models are critical not only for assessing learner texts but also for acting as critics in Large Language Model (LLM)-based simplification pipelines (Hurst et al., 2024). A central obstacle to building robust difficulty models is data scarcity. High-quality assessment requires sizable corpora annotated by domain experts, a resource that is lacking, e.g., in many lower-resource European languages, such as Finnish.

To address this bottleneck, we propose a cross-lingual approach: leveraging existing expert-annotated corpora from higher-resource languages and applying machine translation to generate synthetic labeled training data in the target language. We address two **Research Questions**:

1. Can training data augmented with machine-translated data from a higher-resource language improve the quality of difficulty assessment in a lower-resource language?
2. To what extent does training on translated data improve cross-lingual generalization between the source and target languages?

We train a BERT-based *regression* model to predict difficulty scores on a continuous scale. Our results indicate that using machine-translated data can substantially improve the accuracy and robustness of difficulty assessment in the target language. The paper is organized as follows. Section 2 presents an overview of related work. Section 3 describes the datasets used for training. Section 4 details the experimental setup and results. Section 5 provides conclusions and future directions.

2. Related Work

Estimating the difficulty of written text—variously referred to as readability, proficiency, or grade-level assessment[4]—has a long history in both educational research and NLP. Early formula-based approaches such as Flesch-Kincaid and the Lexile framework provide simple numeric difficulty scores (Kincaid et al., 1975; Stenner, 1996), but rely on surface-level features and fail to capture deeper lexical or syntactic complexity. Subsequent supervised systems incorporated richer handcrafted linguistic features, including parse depth,

[1] Plain Writing Act of 2010, Pub. L. 111-274

[2] European Accessibility Act

[3] Behindertengleichstellungsgesetz (BGG), Federal Republic of Germany, 2002 (amended 2016)

[4] Throughout this paper, we use these terms interchangeably.

Proceedings of the 2nd Workshop on Evaluating Text Difficulty in a Multilingual Context (DeTermIt! 2026), pages 42–50
11 May 2026.

grammatical constructions, and word-frequency lists (Collins-Thompson and Callan, 2004; Vajjala and Meurers, 2012; Laposhina et al., 2018). More recent neural approaches—from hierarchical attention networks (Azpiazu and Pera, 2019) to fine-tuned BERT models (Martinc et al., 2021)—have substantially outperformed feature-based baselines, and Transformer-based models have been compared against feature-engineered systems for both English and Russian (Sharoff, 2022). Large language models have also been evaluated on readability tasks with competitive results (Imperial and Tayyar Madabushi, 2024).

Difficulty assessment plays an equally important role within text simplification pipelines, both as a signal for identifying complex content (Gasperin et al., 2009; Aluísio et al., 2010) and as an optimization objective in rule-based systems (Woodsend and Lapata, 2011). Recent work has moved toward feedback-driven generation, using readability classifiers with reinforcement learning (Alkaldi and Inkpen, 2023) or controllable generation conditioned on target reading level (Agrawal and Carpuat, 2023)—a paradigm directly relevant to our use of a difficulty model as a critic in an LLM-based simplification loop (Hurst et al., 2024).

A persistent bottleneck across all these approaches is the scarcity of expert-annotated, difficulty-labeled corpora, which is especially acute for lower-resource languages. Even a recent shared task on English simplification provided no training data (Alva-Manchego et al., 2025). For Finnish, we build on existing annotated resources (Dmitrieva and Konovalova, 2023; Katinskaia et al., 2025) and extend them via machine translation of Russian-language corpora (Dmitrieva, 2025), directly targeting this labeled-data bottleneck.

3. Data

We first describe the Finnish- and Russian-language data used for training and evaluating the difficulty models. A major challenge is the scarcity of annotated data in Finnish for prediction of difficulty. To address this, we augment a small collection of texts in Finnish annotated with difficulty levels—"native" data—with a larger collection of texts in Russian that were annotated with difficulty levels, and then translated into Finnish using machine-translation (MT) models.

3.1. Native Data

We compile a dataset for Finnish by combining various native and machine-translated sources, spanning the range of CEFR readability levels (A1–C2). Table 1 provides an overview of the composition of this dataset. Native data is drawn from five main sources:

- *Easy Language* (EL)—a collection of texts from government and NGO websites, written in "Easy Language" for non-native speakers;
- *TextBook*—a collection of texts from textbooks for L2 learning, at various CEFR levels;
- *Helsingin Sanomat* (HS)—a commercial news site with the widest coverage nationally;
- *YLE*—a government news site;
- *YLE Selkouutiset* (Selko)—YLE's simplified news for non-native speakers and L2 learners.

Each source contributes texts at different levels of linguistic complexity and genre. *Easy Language* and *YLE selkouutiset* provide simplified Finnish materials aimed at beginners and intermediate learners, while *YLE* and *Helsingin Sanomat* offer authentic journalistic texts at advanced CEFR levels (C1–C2). In total, the native corpus contains 4544 texts distributed across the CEFR scale.

We assign each text in this collection to one of 11 classes—these correspond to the 6 "principal" CEFR levels, plus 5 *intermediate* levels, i.e., A1+, A2+, B1+, etc. The rationale for introducing the intermediate levels is as follows. Some sources, such as textbooks, provide fine-grained assignment of the texts to the CEFR levels. However, other sources (e.g., news sites) yield only a coarse-grained *estimate* of difficult. Thus, we assume that newspaper texts are on a hypothetical level "C", approximately between C1 and C2.

It is also important to note that in modeling we make the assumption that the CEFR levels are *evenly* spaced on a linear difficulty scale. This is done because an exact spacing of the levels on the CEFR scale is *latent* and not known explicitly. For modeling, in Section 4.1, we map these levels onto a continuous numerical scale ranging from 1 to 6, preserving their relative order while enabling regression-based prediction of text difficulty.

To address the problem of data scarcity in Finnish—especially at the lower readability levels—we augment the corpus with machine-translated (MT) texts, which have CEFR annotations in the original. This process is described in detail in Section 3.2. The translated subset (labeled "MT ← RU" in Table 1) contains 8321 documents, which mirror the CEFR distribution of the native data to support balanced training. Including translated data allows us to examine whether CEFR-labeled content from a high-resource language can improve performance in a low-resource target language.

Across both native and translated sub-corpora, the dataset covers all CEFR levels (A1–C2), with a total of 12,865 texts. Lower levels (A1–A2+) are primarily sourced from *Easy Language*, *TextBook*, and translated materials, while mid-level

Level	EL	TextBook	HS	YLE	Selko	Native total	MT ← RU	Overall total
A1	0	1	0	0	0	1	294	295
A1+	153	0	0	0	0	153	282	435
A2	0	363	0	0	0	363	465	828
A2+	0	0	0	0	0	0	96	96
B1	0	229	0	0	0	229	3301	3530
B1+	0	0	0	0	766	766	1672	2438
B2	0	163	0	0	0	163	834	997
B2+	0	0	0	0	0	0	0	0
C1	0	192	0	0	0	192	484	676
C1+	0	0	715	703	0	1418	0	1418
C2	0	175	0	0	0	175	29	204
Total	153	1123	715	703	766	3460	7457	10917

Table 1: Number of documents in Finnish *native* and machine-translated (MT) datasets by CEFR Level

Source	CEFR	Level	Total # Docs	Average # Words	Average # Sent.
RuFoLa	A1	1.0	301	136	8.8
Encyclop.	A1-A2	1.5	282	31	12.3
RuFoLa	A2	2.0	466	183	10.5
Zlatoust	A2-B1	2.5	96	50	8.2
RuFoLa	B1	3.0	3306	91	12.2
Zlatoust	B1-B2	3.5	1677	54	15.8
Zlatoust	B2	4.0	834	228	12.8
RuFoLa	C1	5.0	485	363	14.9
RuFoLa	C2	6.0	29	385	16.5

Table 2: Annotated documents in Russian.

texts (B1–B2) are drawn from *YLE selkouutiset* and corresponding MT data. The advanced levels (C1–C2) are mainly represented by authentic Finnish news and literary texts from *HS* and *YLE*. This distribution ensures that the dataset reflects both pedagogically simplified and naturally complex usage of Finnish, enabling robust analysis of cross-lingual transfer and translation-based augmentation across readability levels. We compile a comprehensive Finnish dataset by combining multiple native and machine-translated sources, spanning a wide range of CEFR readability levels (A1–C2).

3.2. Translated Data

We use text resources in Russian that have been manually annotated for difficulty, and translate them into Finnish to augment the training dataset. We use a collection of Russian simple-language corpora, introduced in (Dmitrieva, 2025). Two corpora of annotated Russian texts, shown in Table 2:

- the *RuFoLa* corpus (Laposhina, 2020), which contains texts from coursebooks designed for learners of Russian as a foreign language;

Split	FI	RU (MT)	Both
Train	2,332	5,961	8,293
Dev	263	748	1,001
Test	865	748	1,613
Total	3,460	7,457	10,917

Table 3: Data splits by source language

- the *RuAdapt* corpus (Dmitrieva and Tiedemann, 2021), a *parallel* corpus of Russian–Simple Russian, with authentic texts adapted for learners of Russian as a foreign language. In this paper, we use only the literary (*Zlatoust*) and encyclopedic sub-corpora (*Encyclop.*).

We translate the Russian texts into Finnish using models from OpusMT.[5] It is critical to note that machine translation does not *guarantee* that a text in Russian will remain at the same difficulty level after translation into Finnish. This question—under what conditions and to what extent do MT models preserve the difficulty level of the original text—deserves detailed investigation on its own; we would expect that this would depend heavily on how the MT model is trained. However, these particular MT models with which we experiment do seem to exhibit a strong ability to preserve the difficulty level across the translation, as confirmed by manual inspection of a sample by native language experts.

Examining the instances in the Russian corpus, we find that some instances that are too short or too long. Therefore, we removed some outlier instances, and hence the number of *MT ← RU* documents in Table 1 is somewhat lower than the original Russian texts in Table 2. All documents—native and translated—were split into 3 sets: train-

[5] Tatoeba MT model for Slavic–Finnish.

Sample	TTR	Lexical Density	Mean Word Length	Mean Sent. Length	Mean Clause Length	POS Diversity
FI	0.6993	0.6930	7.40	8.65	19.67	1.90
RU	0.8262	0.5934	5.30	12.71	10.76	2.04
FI+MT	0.8598	0.6338	6.61	9.31	10.68	1.85

Table 4: Linguistic features for FI, RU and FI+MT samples.

Exp	Lang	Train & Dev Set	Test Set	MSE (%)	RMSE (%)	MAE (%)	R^2 (%)
1	FI	Native	Native	5.12	22.63	12.17	**97.30**
2	FI	Native	MT	146.84	121.18	102.59	-99.71
3	FI	MT	Native	155.00	124.50	109.25	18.24
4	FI	MT	MT	24.54	49.54	26.80	66.93
5	FI	Native + MT	Native	7.57	27.51	9.43	96.01
6	FI	Native + MT	MT	**4.22**	**20.55**	**8.24**	94.26
7	RU	Native	Native	19.19	43.81	26.04	73.90
8	RU	Native + MT	Native	**16.72**	**40.89**	**19.43**	**77.26**

Table 5: Model performance on difficulty prediction for Finnish and Russian text. MT denotes data autmentation via machine translation (Russian to Finnish). Experiments 1–8 test various training vs. test combinations of native vs. translated datasets.

ing, validation, and test, as shown in Table 3. For the experiments with Finnish, we translated the Russian training and validation sets, and added them to the corresponding Finnish sets, to augment the limited size of the Finnish sets. In contrast, for the experiments with Russian, we translated only the Finnish training set and incorporated it into the Russian training set, as the Russian development set was already sufficiently large.

Table 4 reports six metrics of linguistic complexity, which are commonly used to characterize text difficulty:

- Type-Token Ratio (TTR)—is a measure of vocabulary richness,
- Lexical Density—measures the proportion of content words,
- Mean Word Length—measures morphological variety,
- Mean Sentence Length—reflects syntactic variety,
- Mean Clause Length—reflects the structural complexity of sentences, capturing both the elaboration of clause-internal constituents and the depth of syntactic embedding,
- POS Diversity—indicates part-of-speech variety within the text.

These features roughly reflect the lexical and structural properties of the texts in the dataset. They are not used in modeling (at present), and are presented to give the reader an intuitive description of the data.

The main point in this Section is that the dataset augmented with translated data is 3 times larger than the original "native" dataset—which we hope will help train a more accurate regresion model.

4. Experiments

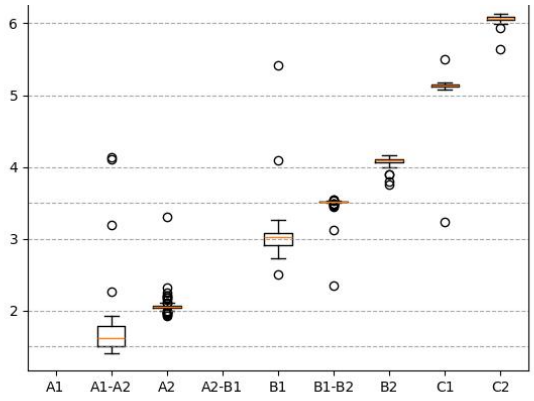
To examine whether machine-translated data can improve text difficulty prediction, we train regression models on native Finnish texts and translated Russian texts. This allows us to explore the research questions: assess how MT-based data augmentation influences model performance and cross-lingual generalization in predicting document-level difficulty.

4.1. Model settings

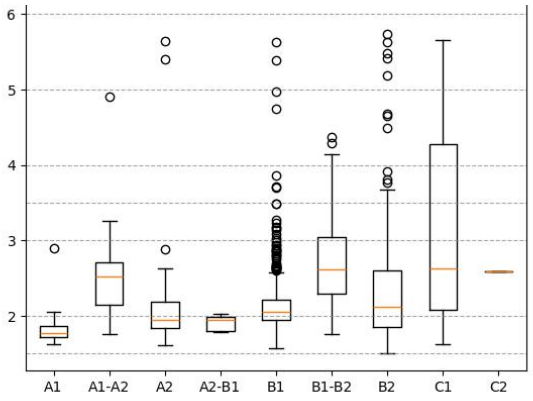
We build a BERT-based regression model to predict text difficulty. For Finnish, we use the TurkuNLP/bert-base-finnish-cased-v1 model; for Russian, we use ai-forever/ruBert-large. Both models are fine-tuned with a regression objective.

4.2. Experiments on RQ1

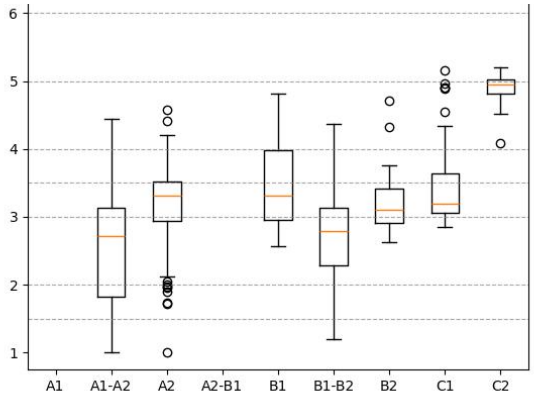
Table 5 presents the results in terms of key performance measures—MSE (mean quared error), RMSE (root mean square error), MAE (mean absolute error) and R^2 (coefficient of determination)—across different training and testing dataset configurations. Overall, the results show that incorporating machine-translated data from Russian improves text difficulty prediction for Finnish, com-



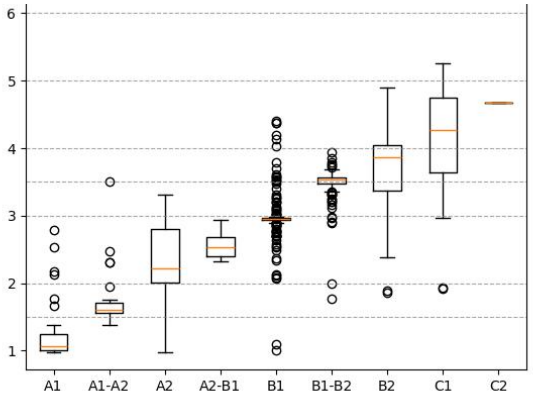
(1) FI train and test on native data



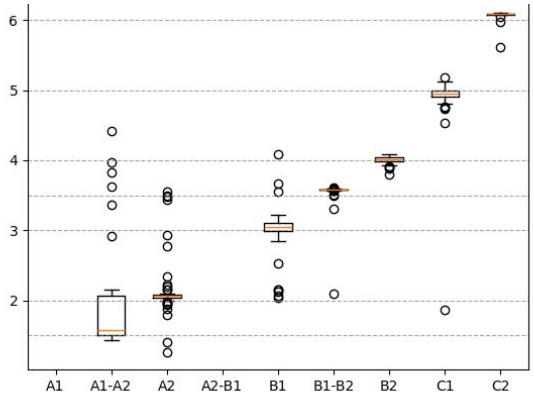
(2) FI train on native, test on MT



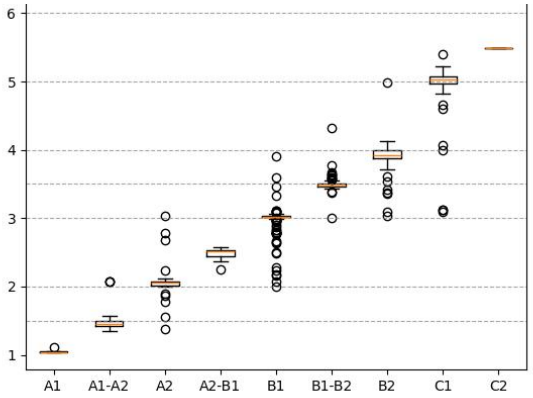
(3) FI train on MT, test on native data



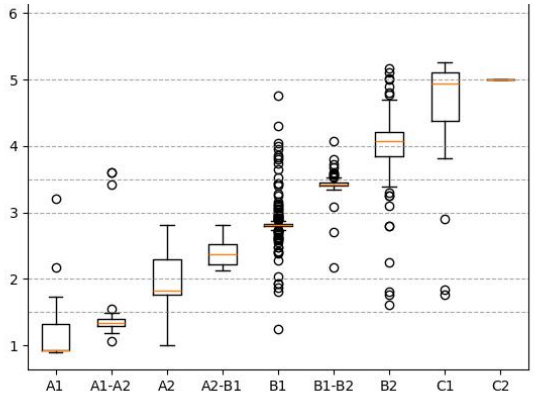
(4) FI train and test on MT



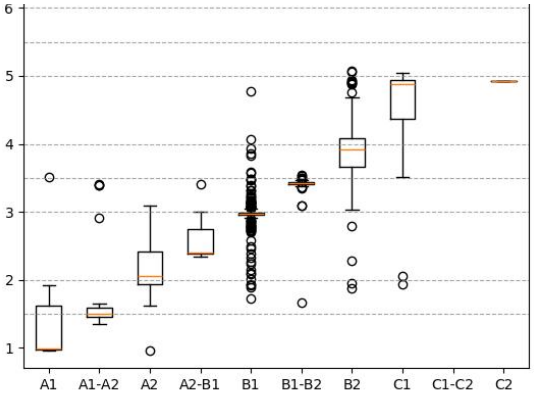
(5) FI train on all data, test on native,



(6) FI train on all data, test on MT,



(7) RU train and test on native data



(8) RU train on all data and test on native data

Figure 1: Prediction of difficulty, corresponding to experiments in Table 5.

Exp.	Lang.	Train Set	Dev Set	Test Set	MSE (%)	RMSE (%)	MAE (%)	R^2 (%)
Label Injection (whole-label inclusion)								
9	FI	Native + MT (A1–A2)	Native + MT	Native	29.29	54.12	41.45	84.55
10	FI	Native + MT (B1–B2)	Native + MT	Native	6.29	25.08	11.40	96.68
11	FI	Native + MT (C1–C2)	Native + MT	Native	16.96	41.18	25.62	91.06
Label-wise Proportional Sampling								
12	FI	Native + 20% MT	Native + MT	Native	7.83	27.97	16.81	95.87
13	FI	Native + 40% MT	Native + MT	Native	7.20	26.84	10.66	96.20
14	FI	Native + 60% MT	Native + MT	Native	8.22	28.68	14.72	95.66
15	FI	Native + 80% MT	Native + MT	Native	8.71	29.51	12.74	95.41
16	FI	Native + 100% MT	Native + MT	Native	7.57	27.51	9.43	96.01

Table 6: Model performance on Finnish text difficulty prediction using machine-translated (MT) Russian data. **Label Injection:** Entire CEFR-level subsets (A1–A2, B1–B2, C1–C2) of MT data are injected into the training set. **Label-wise Proportional Sampling:** Russian MT data are sampled proportionally within each CEFR label (20%–100%) and added to Finnish native data. Note: experiment 16 is exactly the same as line 5 in Table 5 (repeated for clarity).

pared to training on native data alone, or on translated data alone.

When using only translated Russian data (MT RU→FI), model performance remains modest, with MSE of 24.54% and R^2 of 66.93% (Figure 1.4). However, combining translated and native Finnish data—while applying dataset balancing—yields a dramatic performance gain , reaching MSE of 4.22%, MAE of 8.24%, and R^2 of 94.26%. The box plots in Figure 1.6 show notably tighter error distributions in this combined setting, indicating that the model may benefit from both the larger volume and the additional diversity provided by the translated corpus.

The best overall performance is achieved when the model is trained on the native + MT and evaluated on native Finnish data, reaching an R^2 of 96.01% in Figure 1.5. This demonstrates that machine-translated data not only contributes to improved prediction accuracy in a low-resource setting, but also supports generalization to unseen native Finnish texts. By contrast, models trained only on translated data perform poorly when tested on native Finnish data—MSE 155.00%, R^2 18.24%—which suggests that direct transfer without native exposure is insufficient (Figure 1.3).

Training exclusively on native Finnish data yields strong in-domain performance (R^2 97.30% in Figure 1.1), confirming the quality of native annotations. However, the combined model's comparable accuracy, despite including automatically translated material, shows that translation-based augmentation is an effective strategy for low-resource difficulty prediction.

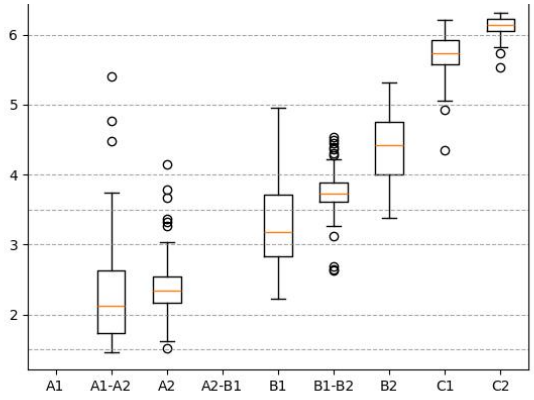
In sum, the box plots and quantitative results jointly confirm that (1) machine-translated data can substantially enhance low-resource Finnish performance.

4.3. Experiments on RQ2

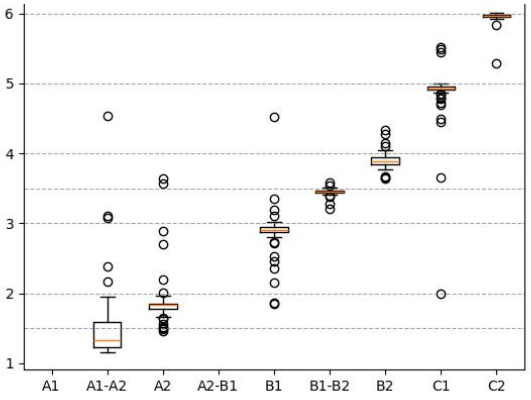
Table 6 presents the results of ablation studies, combining the native Finnish data with varying amounts of machine-translated (MT) data from the Russian corpus. We explored two strategies: label injection, where entire CEFR-level subsets of translated data (A1–A2, B1–B2, C1–C2) were added to the training set, and label-wise proportional sampling, where translated data were added in increasing proportions (20–100%) across all levels.

Under the label injection setting, we observe that including B1–B2-level translated texts yields the strongest improvement, achieving R^2 of 96.68% and the lowest overall error rates (Figure 2.2). This suggests that mid-level translated data contribute most effectively to modeling Finnish text difficulty, possibly because they provide balanced lexical and syntactic diversity without overwhelming the model with extreme examples from beginner or advanced levels. In contrast, adding low-level (A1–A2) (Figure 2.1) or high-level (C1–C2) (Figure 2.3 translated data results in noticeably higher error, indicating limited transferability at the edges of the CEFR scale.

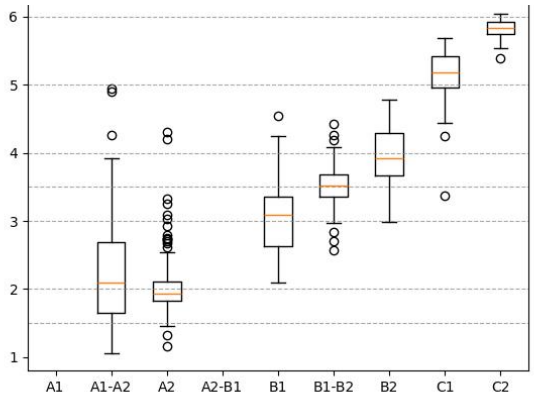
In the experiments with proportional sampling, performance is consistently high across all proportions, with minor fluctuations. The best result (Figure 2.5) is obtained at 40% translated data, reaching R^2 of 96.20%, slightly outperforming full (100%) augmentation ((Figure 2.8). This pattern implies that moderate infusion of translated data effectively regularizes the model, enhancing generalization without introducing domain noise from excessive MT input.



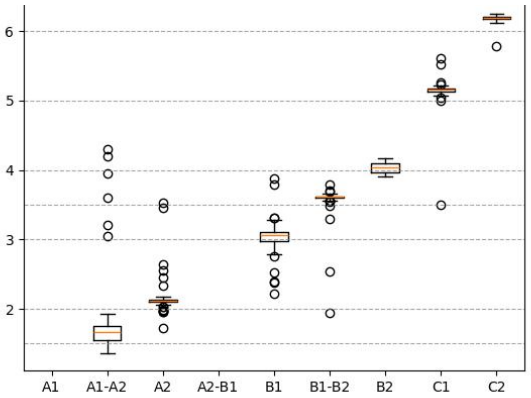
(1) Exp. 9



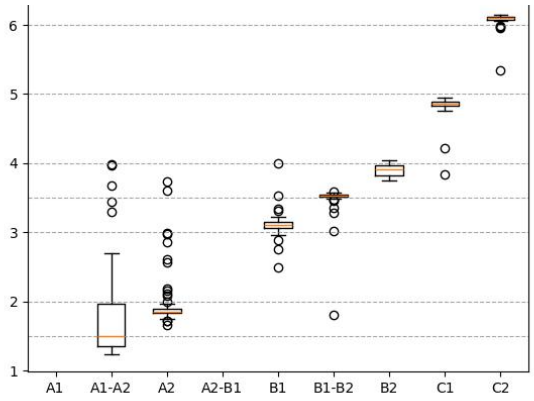
(2) Exp. 10



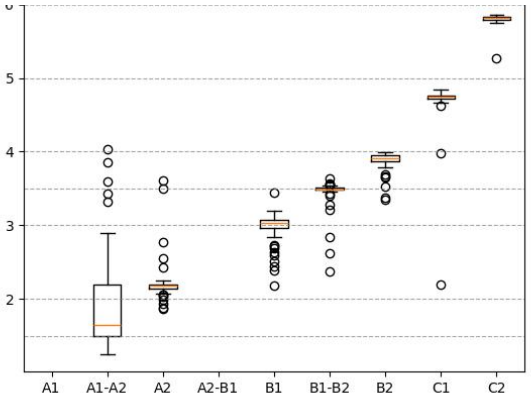
(3) Exp. 11



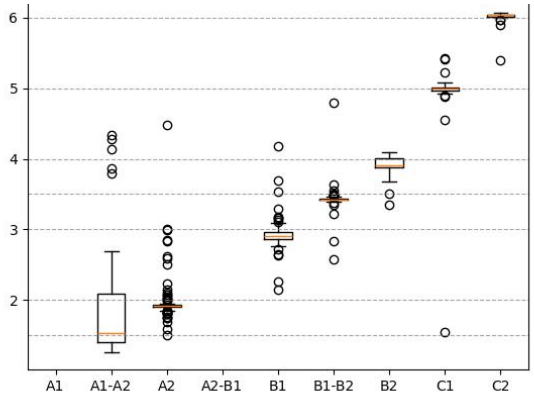
(4) Exp. 12



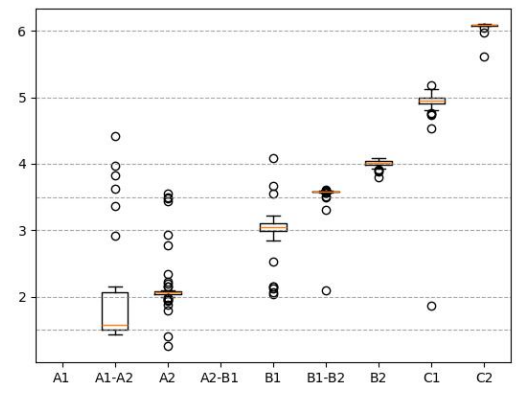
(5) Exp. 13



(6) Exp. 14



(7) Exp. 15



(8) Exp. 16

Figure 2: Box plots for experiments (Exp. 9–16) showing error distributions across models trained with different proportions and CEFR-level subsets of machine-translated data.

5. Conclusion

We set out to explore two research questions: (1) machine-translated data from a high-resource language can substantially improve prediction of text difficulty in a low-resource setting, and (2) the improvement generalizes across languages, especially when the quantity of translated data are well-balanced across linguistic levels.

The results confirm that carefully selected MT data can act as a strong proxy for native material in low-resource language modeling. The experimental findings support both research questions. Data augmentation via translating text annotated with difficulty from another language can indeed improve the performance of the difficulty model. A crucial caveat is that we must assure that the particular MT model we use for translation is able to preserve the CEFR level of its input reasonably well. This is far from a foregone conclusion, since many modern LLMs are trained to "improve" on the text while translating, including simplifying the text or making it more "standard." Care must be taken that the MT model preserves the level reasonably accurately, and additional techniques need to be explored to ensure this in a systematic way.

In future work, we plan to pursue several directions. One plan is to integrate feature-based and Transformer-based models, which could also help with the interpretability of the resulting models in terms of easily understood features. We plan to explore whether can provide reasonable guarantees that MT preserves the levels. This would highlight the importance of our results, since difficulty- and CEFR-annotated data of high quality are very difficult to find, and creating such data is highly resource-intensive. Data augmentation via MT would allow us to grow our training (and test) datasets considerably, to yield substantial improvements in performance on this complex task. At the same time will explore multilingual models of difficulty, to study to what extent the transformer can identify language-independent features that impact on text difficulty.

6. Limitations and Ethical Considerations

While our results show that difficulty models trained on labeled data from multiple languages can be effective, several limitations remain. First, the models are trained and evaluated on small datasets. Working only with Finnish may limit generalizability to other languages or domains, and additional languages should be explored. Second, the mappings that we apply—from continuous regression scores to CEFR levels—introduce discretization errors that may obscure improvements on a more nuanced level.

This work aims at improving language accessibility, particularly for second-language (L2) learners, and seeks to reduce linguistic barriers in education and communication. However, several ethical considerations must be acknowledged. First, automated simplification tools may reinforce biases present in the training data, especially if texts from specific groups or dialects are underrepresented. Second, in general, over-reliance on automated systems may reduce the role of human educators in assessing learner needs, which is not the intent, and is not productive. Lastly, indiscriminate use or misuse of simplification systems—e.g., to manipulate or oversimplify critical content—can have adverse effects. We emphasize that these systems should be used as *assistive* tools, rather than as replacements for human judgment in the context of education or public communication.

References

Sweta Agrawal and Marine Carpuat. 2023. Controlling pre-trained language models for grade-specific text simplification. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12807–12819, Singapore.

Wejdan Alkaldi and Diana Inkpen. 2023. Text simplification to specific readability levels. *Mathematics*, 11(9):2063.

Sandra Aluísio, Lucia Specia, Caroline Gasperin, and Carolina Scarton. 2010. Readability assessment for text simplification. In *Proceedings of the 5th Workshop on Innovative Use of NLP for Building Educational Applications*, Los Angeles, California.

Fernando Alva-Manchego, Regina Stodden, Joseph Marvin Imperial, Abdullah Barayan, Kai North, and Harish Tayyar Madabushi. 2025. Findings of the TSAR 2025 shared task on readability-controlled text simplification. In *Proceedings of the 4th Workshop on Text Simplification, Accessibility, and Readability*, Suzhou, China.

Ion Madrazo Azpiazu and Maria Soledad Pera. 2019. Multiattentive recurrent neural network architecture for multilingual readability assessment. *Transactions of the Association for Computational Linguistics*, 7:421–436.

Kevyn Collins-Thompson and James P. Callan. 2004. A language modeling approach to predicting reading difficulty. In *HLT-NAACL 2004:*

Proceedings of the Human Language Technology Conference of the NAACL, pages 193–200.

Anna Dmitrieva. 2025. *Resources and Tools for Automatic Text Simplification: Cases of Russian and Finnish*. Ph.D. thesis.

Anna Dmitrieva and Aleksandra Konovalova. 2023. Creating a parallel Finnish-Easy Finnish dataset from news articles. In *Proceedings of the 1st Workshop on Open Community-Driven Machine Translation*, pages 21–26, Tampere, Finland.

Anna Dmitrieva and Jörg Tiedemann. 2021. Creating an aligned Russian text simplification dataset from language learner data. In *Proceedings of the 8th Workshop on Balto-Slavic Natural Language Processing*, pages 73–79, Kiyv, Ukraine. Association for Computational Linguistics.

Caroline Gasperin, Lucia Specia, Tiago F. Pereira, and Sandra M. Aluísio. 2009. Learning when to simplify sentences for natural text simplification. In *Proceedings of the Encontro Nacional de Inteligência Artificial (ENIA)*, Bento Gonçalves, Brazil.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

Joseph Marvin Imperial and Harish Tayyar Madabushi. 2024. Part of the problem: Toward a finer-grained analysis of LLMs for readability assessment. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications*.

Anisia Katinskaia, Anh-Duc Vu, Jue Hou, Ulla Vanhatalo, Yiheng Wu, and Roman Yangarber. 2025. Estimation of text difficulty in the context of language learning. In *BEA: 20th Workshop on Innovative Use of NLP for Building Educational Applications*, Vienna, Austria.

J. Peter Kincaid, Robert P. Fishburne, Richard L. Rogers, and Benjamin S. Chissom. 1975. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Technical report, Naval Air Station Memphis (Research Branch Report 8-75).

Antonina Laposhina. 2020. A corpus of Russian textbook materials for foreign students as an instrument of an educational content analysis. *Russian Language Abroad*, 6(283):22–28.

Antonina Laposhina, Tatiana Veselovskaya, Maria Lebedeva, and Olga Kupreshchenko. 2018. Automated text readability assessment for Russian second language learners. In *Computational Linguistics and Intellectual Technologies*, pages 403–413.

Matej Martinc, Senja Pollak, and Marko Robnik-Šikonja. 2021. Supervised and unsupervised neural approaches to text readability. *Computational Linguistics*, 47(1):141–179.

Shingo Nahatame and Katsuyoshi Yamaguchi. 2026. Revisiting text readability and processing effort in second language reading: Bayesian analysis of eye-tracking data. *Language Learning*.

Serge Sharoff. 2022. What neural networks know about linguistic complexity. *Russian Journal of Linguistics*, 26(2):371–390.

A. Jackson Stenner. 1996. Measuring reading comprehension with the Lexile framework. Technical report, MetaMetrics Inc., Durham, NC.

Sowmya Vajjala and Detmar Meurers. 2012. On improving the accuracy of readability classification using insights from second language acquisition. In *Proceedings of the 7th Workshop on Building Educational Applications Using NLP (BEA)*.

Kristian Woodsend and Mirella Lapata. 2011. Learning to simplify sentences with quasi-synchronous grammar and integer programming. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, Edinburgh, Scotland.