Estimation of Text Difficulty in the Context of Language Learning

Anisia Katinskaia,[†‡] **Anh-Duc Vu,**[†‡] **Jue Hou,**[†‡] **Ulla Vanhatalo,**[◇]
Yiheng Wu,[‡] **Roman Yangarber**[‡]
[†]Department of Computer Science, [‡]Department of Digital Humanities
[◇]Department of Finnish, Finno-Ugrian and Scandinavian studies
University of Helsinki, Finland
first.last@lhelsinki.fi

Abstract

Easy language and text simplification are currently topical research questions, with important applications in many contexts, and with various approaches under active investigation, including prompt-based methods. The estimation of the level of difficulty of a text becomes crucial when the estimator is employed inside a simplification workflow as a quality-control mechanism. It can act as a *critic* in frameworks where it can guide other models, which are responsible for generating text at a specified level of difficulty, as determined by the user's needs. We present our work in the context of simplified Finnish. We discuss problems in collecting corpora for training models for estimation of text difficulty, and our experiments with estimation models. The results of the experiments are promising: the models appear usable both for assessment and for deployment as a component in a larger simplification framework.

1 Introduction

In the US[1] and in the European Union,[2] legal pressures are emerging with laws that require government-affiliated agencies, as well as private-sector organizations in certain situations, to use clear communication that members of the public can understand. Workflows that involve easy language are already in official use at various levels of functioning in the public and private sectors in 20 countries in the EU. Easy language also plays a key role in second-language (L2) education, in particular—simplification of text to a level appropriate for a given learner is a key component of *personalization* in teaching. Simplification itself is a widely researched area in NLP.

In this paper, we take the position that methods for evaluating and *assessing* the difficulty level of a piece of text are *prerequisite* to methods for simplification—since in the absence of effective evaluation, simplification methods cannot be effectively validated or falsified.

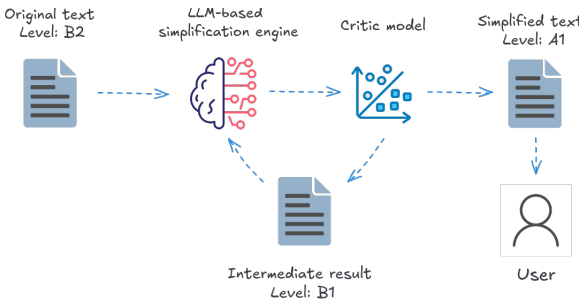

Figure 1: Text simplification using GPT-4o guided by level-aware feedback from a difficulty classifier as critic.

The task of assessing the level of difficulty of the text can be framed as classification or (more appropriately) as regression—labeling a piece of text with a difficulty level, such as, e.g., a CEFR level.[3] We will refer to models performing this task as *difficulty models*. These models can serve various purposes in language learning, such as estimating the difficulty level of texts that learners encounter. In this work, we use difficulty models to guide and evaluate text simplification pipelines performed by a large language model (LLM), specifically GPT-4o from OpenAI (Hurst et al., 2024).

Our simplification pipeline (Figure 1) employs a difficulty model that serves as a *critic*: it evaluates the difficulty level of the LLM output. If the resulting text exceeds the target level, feedback will be sent to the LLM to try again. The feedback includes the resulting text and its estimated level. The pipeline runs several iterations; if the resulting text remains harder than the target level after N iterations, the process terminates, and an error message is returned to the user.

We train two BERT-based difficulty models:

[1] PlanLanguage.gov
[2] European Accessibility Act
[3] CEFR: Common European Framework of Reference for Languages.

Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications, pages 594–611
July 31 - August 1, 2025

one *regression* model, which predicts continuous scores that are later mapped to CEFR levels, and one *ordinal classification* model, which directly predicts the CEFR level of the input text. Our results show that both models improve the performance of the simplification pipeline over a baseline that runs without any critical guidance. The ordinal classification model proves to be a more effective critic for the LLM. Our hypothesis is that it aligns better with the difficulty assessment task because of its ordinal (ranking) nature.

The paper is organized as follows: Section 2 presents an overview of related work; Section 3 discusses the data we use to train the assessment models; Section 4 presents the experimental setup for the difficulty assessment; Section 5 presents the experiments with controlling the behavior of the LLM via a critic that assesses difficulty; Section 6 discusses the results and concludes the paper.

2 Related Work

Assessment of text difficulty, often referred to as readability assessment, has a long history in both education and in NLP. Traditional readability formulas, such as the Flesch-Kincaid Grade Level and Flesch Reading Ease, and the Lexile framework, based on item response theory (IRT), provide simple numeric scores for text difficulty (Kincaid et al., 1975; Stenner, 1996). These methods are easy to apply, but they rely on surface-level features and do not directly account for deeper lexical or syntactic complexity.

Early NLP readability systems used supervised models with hand-crafted linguistic features, including frequency word lists, depth of parse trees, grammatical constructions, and discourse structures. Collins-Thompson and Callan (2004) introduced a language modeling approach to predict reading difficulty for a tutoring system. Vajjala and Meurers (2012) incorporated features from Second Language Acquisition research to better serve language learners. For Russian, Laposhina et al. (2018) introduced a feature-based readability tool available online and widely used by L2 teachers.

Azpiazu and Pera (2019) present a multilingual readability model using a hierarchical attention network that learns to attend to difficult parts of a text and can implicitly learn factors like semantic difficulty or subtle syntactic cues. These models can be trained on proficiency-labeled data (e.g., with CEFR levels) to detect nuances of text difficulty specific to L2 readers (e.g.,, idiomatic language). Recent work has shown that a fine-tuned BERT can outperform strong feature-based baselines by a significant margin in classifying texts by grade level or proficiency level (Martinc et al., 2021). Sharoff (2022) investigated compared the performance of Transformer-based models for predicting text difficulty vs. assessment using linguistic features, such as frequency of conjunctions, discourse particles, etc., for English and Russian.

Early pipeline approaches used readability classifiers to decide when to simplify: for example, Gasperin et al. (2009) trained a model to identify sentences that need simplification based on linguistic complexity features. Aluísio et al. (2010) developed readability assessment tools to support simplifying texts for low-literacy readers. Readability metrics have also served as simplification objectives in rule-based systems— Woodsend and Lapata (2011) incorporate a Flesch-Kincaid grade formula into an optimization-based simplifier.

Readability predictors have been used as feedback in generation loops—Alkaldi and Inkpen (2023) use a readability classifier in a reinforcement learning framework to iteratively simplify a text until it reaches the desired difficulty. More recently, large-scale neural systems have combined reading level prediction with controllable generation techniques (Agrawal and Carpuat, 2023).

3 Data

First, we describe the data used for training and evaluating the difficulty models and for the simplification pipeline. A major challenge is the scarcity of annotated data in Finnish for text simplification and difficulty prediction. To address this, we use a combination of Finnish texts annotated with difficulty levels ("native" data), and Russian texts annotated with difficulty levels and then translated into Finnish using machine-translation models.

3.1 Native Data

We use two collections of native Finnish data. The first consists of 1113 documents manually annotated by teachers of Finnish as a second language (L2), see "Manual" in Table 1. These are primarily informative and literary texts: the former covering topics such as human rights, social benefits, etc.; the latter feature classic Finnish literature and fragments of the Bible. The "Score" column in Table 1 shows the numerical values we assign to CEFR

Source	Level	Score	# Docs	# Words	# Sent.
SM	easy	1.5	153	294	9.3
YLE-selko	medium	3.5	766	249	8.7
HS	hard	5.5	715	598	13.7
YLE	hard	5.5	703	480	14.5
Manual	A2	2.0	363	237	10.9
	B1	3.0	229	204	11.0
	B2	4.0	154	221	11.8
	C1	5.0	192	272	17.5
	C2	6.0	175	189	19.9

Table 1: Native Finnish data.

levels, which are later used in regression models.

The second collection contains 2337 texts from *Suomen Mestari* (SM), a Finnish textbook, and *YLE selkosuomeksi* news,[4] as well as news articles from the major newspapers *YLE* and *Helsingin Sanomat* (HS). These texts were not manually annotated. Instead, we make a coarse assumption based on the source: all texts from SM are labeled as easy, texts from YLE-selko as medium, and texts from YLE and HS as hard. We then suppose these difficulty levels roughly correspond to CEFR levels A1-A2, B1-B2, and C1-C2, respectively. Although this source-based annotation is a simplification—individual texts may vary in difficulty—it provides a practical heuristic in the context of limited human resources for annotating data.

3.2 Translated Data

Having some amount of Russian data annotated for difficulty, we translate it into Finnish to extend the size of the training set.

We use two sources of annotated Russian texts: 1. the *RuFoLa* corpus (Laposhina, 2020), which contains texts from coursebooks designed for learners of Russian as a foreign language; 2. the *RuAdapt* corpus (Dmitrieva and Tiedemann, 2021), a *parallel* Russian–Simple Russian dataset of texts adapted for learners of Russian as a foreign language. For our study, we use only the literary (*Zlatoust*) and encyclopedic sub-corpora, see Table 2. The "Score" column again shows the mapping between CEFR levels and numeric labels used later for a BERT-based regression model.

We filter out texts shorter than 10 words, as such a short context can negatively affect translation quality. We translated the Russian texts into Finnish using a model from OpusMT.[5] We should note that machine translation does not guarantee that a text in Russian will remain at the same difficulty level after translation into Finnish. This problem merits a dedicated research experiment. The entire dataset was split into 3 sets: training, validation, and test, see Table 3. The test set—860 texts—contains only native documents, and most documents are manually annotated.

[4]News in Simple Finnish: yle.fi/selkouutiset

[5]The Tatoeba model for Slavic-Finnish.

Source	Level	Score	# Docs	# Words	# 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	3300	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.

Split	# Documents	Source
Training	6248	MT
	2221	Native
Validation	1222	MT
	364	Native
Test	865	Native

Table 3: Data splits.

4 Experiments

To establish an interpretable baseline for document-level difficulty prediction, we first train a feature-based regression model. This allows us to evaluate how well linguistic features alone can capture text difficulty, and later compare its performance to that of less interpretable deep-learning approaches.

4.1 Feature-based Regression

In this experiment, we use only native Finnish texts to train a Ridge regression model that predicts the difficulty level of a document. The target labels are mapped to the following numeric values: 0.0 (easy), 1.0 (medium), and 2.0 (hard). Manually annotated documents are mapped to the same numeric values: A1-A2 to 0.0, B1-B2 to 1.0, and C1-C2 to 2.0. We use these numeric values instead of the scores presented in Table 1 for simplicity.

We use 179 features to capture linguistic characteristics of the texts. These include normalized averages of count of POS tags, depth of parse

tree, sentence length, word distribution across ten frequency bins, and the proportion of out-of-vocabulary (OOV) words.[6] The features also include the counts of over 160 linguistic constructs, covering grammatical features—e.g., tense, case, number, etc.—and syntactic patterns—e.g., necessity constructions, government structures, etc. The extraction of constructs from text is performed using the text processing pipeline in the Revita language learning system (Katinskaia et al., 2018, 2017; Hou et al., 2019); see examples of linguistic constructs and how they are extracted from text in (Katinskaia et al., 2023). Details of the features appear in Appendix D.

We evaluate three variants of the baseline model:

(A) using all 179 features,

(B) using a bootstrap selection of 104 features,

(C) performing feature selection by training a Lasso regression model.

More details on the models are presented in Appendix A. As all models exhibited comparable performance, we adopt model (B) as the baseline in subsequent analyses due to its smallest feature set.

4.1.1 Results

Evaluation was performed using only native Finnish texts. The baseline model (B) achieved a mean absolute error (MAE) of 0.27 and a root mean squared error (RMSE) of 0.35. Figure 2 shows the distribution of the predicted scores in the three coarse levels of difficulty. The plot shows that easy texts tend to get scores higher than 0.0. This could be explained by the fact that we have much fewer easy texts in the native corpus, as well as by the assumption that all SM texts should be labeled easy, while in fact some of these texts are of intermediate difficulty. Nevertheless, the results provide a strong baseline for comparison with more complex models used in subsequent experiments, which offer less interpretability.

4.2 BERT-based Regression

We extend the BERT model for regression-based difficulty prediction, integrating custom loss weighting to handle class imbalances in the training data. The model is based on BERT, whose output layer is replaced with: (a) a pre-classification layer that projects BERT's pooled output into a lower-dimensional space, has ReLU activation and Dropout, and (b) a final feedforward regression head that predicts a continuous difficulty score.

[6]Based on a large Finnish corpus, we build a list of words sorted by frequency and grouped into frequency bins.

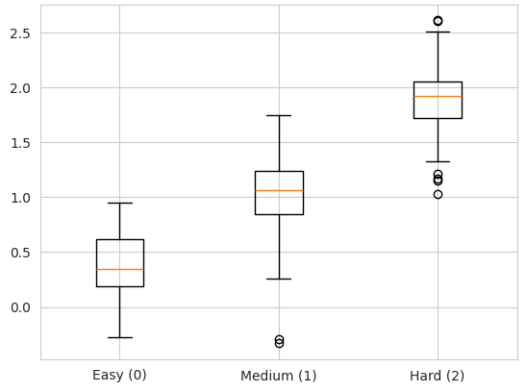

Figure 2: Feature-based regression baseline model (C). Predicted regression scores across difficulty levels: Easy (0), Medium (1), and Hard (2).

The model is trained using weighted mean squared error (MSE) loss. To prevent the model from being biased toward the most frequent difficulty levels in the training data, sample weights are computed inversely proportional to the frequency of each difficulty level. These weights are then normalized to ensure they sum to 1.

The model was trained on all training data presented in Table 3, using the Adam optimizer, separate optimization parameters for the BERT parameters and the linear layers, weight decay = 0.01, cosine scheduler for the learning rate, and early stopping.

4.2.1 Results

The evaluation was again performed on the test set containing native Finnish texts. The BERT-based regression model achieved MAE 0.13 and RMSE 0.29. Figure 3 shows the distribution of the predicted scores at all CEFR levels in the test set. The number of documents per level is shown in the "Support" column of Table 4. As we can see from the plot, for some of the difficulty levels (particularly, for level A1-A2), predicted scores tend to be higher than the true labels, indicating some bias toward overestimation.

To assess the classification performance, we map real-valued predictions to the nearest CEFR level. The resulting confusion matrix is in Figure 4. Class-wise precision, recall, and F1-scores are in Table 4. Overall, the model performs well across most CEFR levels. The lowest F1-score is observed for the A1-A2 level, which also has the smallest number of examples in the test set.

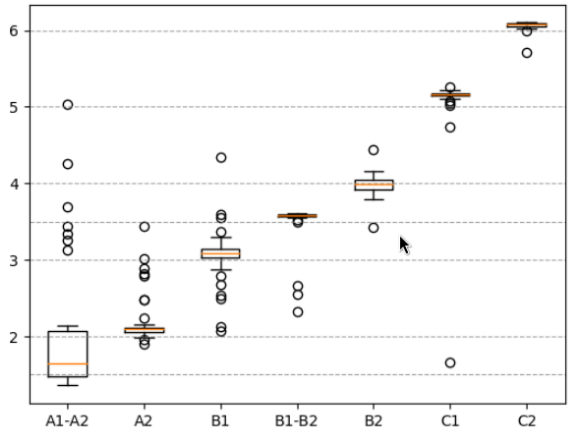

Figure 3: Predicted regression scores across difficulty levels using BERT-based regression model.

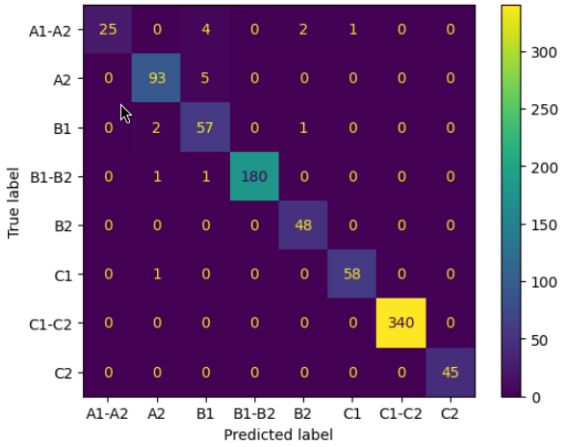

Figure 4: Confusion matrix after mapping difficulty scores to CEFR levels.

Only 1.3% of the test documents were assigned a predicted level that differs from the true level by *more than one* CEFR level. We consider deviations within one level to be acceptable, given the inherent difficulty and subjectivity of the task.

We examine the agreement between the feature-based regression model and the BERT-based regression on the test set, and the agreement of both models with the true labels. The predictions of the two models show a strong correlation: Spearman's rank correlation is 0.76, and Pearson's correlation is 0.83; Quadratic Weighted Kappa (QWK) is 0.84.[7] This suggests a high degree of both rank-order and linear agreement between the models, despite being trained on different datasets and using different features.

The BERT-based model achieves near-perfect

[7]Predictions from the feature-based model were linearly rescaled to match the 1–6 scale of the BERT-based regression.

Level	Precision	Recall	F1-score	Support
A1-A2	1.00	0.78	0.88	32
A2	0.96	0.95	0.95	98
B1	0.85	0.95	0.90	60
B1-B2	1.00	0.98	0.99	183
B2	0.94	1.00	0.97	48
C1	0.98	0.98	0.98	59
C1-C2	1.00	1.00	1.00	340
C2	1.00	1.00	1.00	45

Table 4: Performance on the test set after mapping difficulty scores to CEFR levels.

Model	Pearson	Spearman	QWK
BERT vs. True	0.98	0.95	0.98
Feature vs. True	0.84	0.83	0.81

Table 5: Correlation and agreement of feature-based baseline model and BERT-based regression model with true labels.

agreement with the true difficulty labels (see Table 5), where the gain in QWK suggests that BERT is particularly better at matching difficulty levels. In contrast, the feature-based model demonstrates good but notably lower performance (0.98 vs. 0.81). Both models are available for testing.[8]

4.3 BERT-based Ordinal Classification

This model extends BERT for rank-consistent ordinal regression (Cao et al., 2020), a task in which labels have a meaningful order but unknown interval distances. Unlike standard classification, ordinal regression models the probability of a response exceeding certain thresholds, making it particularly useful for difficulty assessment.

The model predicts $\mathbb{P}(Y > k)$ for each threshold k using a modified BERT architecture, where a linear classifier estimates the probability that the input exceeds a set of ordinal thresholds. In particular, given an input sequence, we pass the pooled output of BERT through a dropout layer and a linear classification head of size $(\text{hidden\_dim} \rightarrow K-1)$, where K is the number of CEFR levels.

For K ordinal labels, the model outputs $K-1$ logits for each threshold. Each logit represents the probability:

$$\mathbb{P}(Y > k \mid X)$$

for each difficulty threshold k, where X represents the BERT-generated input representation.

[8]revita.helsinki.fi/selkomitta

Since ordinal regression differs from standard classification, we use a binary cross-entropy (BCE) loss adapted for ordinal constraints:

- **Ordinal Target Construction:** For a batch of size N and K classes, we construct a binary target matrix $\mathbf{T} \in \{0,1\}^{N\times(K-1)}$, where each element $T_{i,k} = \mathbb{I}[y_i > k]$ indicates whether the true label exceeds threshold k.
- **Weighting Mechanism:** A weight matrix $\mathbf{W} \in \mathbb{R}^{N\times(K-1)}$ assigns higher penalties to more severe misclassifications. This can be scaled by a global hyperparameter α to control the influence of the weighting.

The weighted ordinal loss function is defined as:

$$L = \frac{1}{N}\sum_{i=1}^{N} s_i \sum_{j=1}^{K-1} w_{i,j} \cdot \mathrm{BCE}(\sigma(z_{i,j}), t_{i,j})$$

where:

- N is the batch size.
- $z_{i,j}$ is the model logit for level j.
- $\sigma(z)$ is the sigmoid function.
- $t_{i,j}$ is the binary target: 1 if the true label exceeds threshold j, 0 otherwise.
- $w_{i,j}$ is a weight penalty based on label distance
- s_i is an additional sample-level weight to address class imbalance.

The weights w_{ij} are given by:

$$w_{ij} = 1 + \alpha \cdot |y_i - j|, \quad \alpha > 0$$

where y_i is the true ordinal label. This weighting penalizes predictions that are farther from the correct class more heavily. In our experiments, we set $\alpha = 0.5$.

By modeling thresholds rather than treating classes as independent, the loss preserves ordinal relations. Furthermore, the model learns a probability distribution over ranks, capturing uncertainty rather than committing to hard class decisions.

To obtain the predicted ordinal class, we apply a sigmoid activation to the model's output logits, yielding threshold probabilities $\mathbb{P}(Y > k)$ for each $k = 1, \ldots, K-1$. The predicted class $\hat{y}$ is then calculated by counting how many of these probabilities exceed the threshold of 0.5:

$$\hat{y} = \sum_{k=1}^{K-1} \mathbb{I}\left[\mathbb{P}(Y > k) > 0.5\right]$$

Accuracy	0.76	RMSE	0.57
MAE	0.28	ρ	0.89
QWK	0.87	τ	0.83

Table 6: Results of ordinal classification.

Here, $\mathbb{I}[\cdot]$ denotes the indicator function, which returns 1 if the condition is true and 0 otherwise.

Intuitively, this approach treats the predicted class as the number of ordinal thresholds that the input is likely to exceed with confidence greater than 0.5—higher classes correspond to exceeding more difficulty levels.

During training, we apply different learning rates for BERT layers and for the classifier head. Optimization is performed using AdamW. The learning rate is scheduled using a cosine annealing strategy with a linear warm-up over the first 10% of the training steps. The model is trained using the same data as for BERT-based regression. Since our data is not balanced over many classes for classification, we map the labels to 6 classes only: A1, A2, B1, B2, C1, and C2.

4.3.1 Results

The ordinal critic performs worse in terms of standard classification metrics on the same test set of 865 documents, see Table 10 and Figure 7 in Appendix B. The model achieves an accuracy of 0.76, see Table 6. However, metrics such as accuracy do not fully capture ordering information.

To better account for the severity of misclassifications, we report the Mean Absolute Error (MAE), which measures the average absolute difference between the predicted and the true labels—penalizing larger mistakes more heavily than smaller ones. MAE of 0.28 indicates that, on average, the predicted level deviates from the ground truth by about a quarter of a CEFR level. Analyzing the prediction errors in more detail, we find that 76% of the predictions exactly match the true levels, while 20% of the predictions are within one level of the ground truth. Only 4% of the documents are misclassified by more than one level—a deviation we consider "intolerable" due to the impact on downstream applications. The RMSE of 0.57, which penalizes larger errors more heavily, confirms the relatively low deviation.

In addition to accuracy, we report three metrics that better reflect the ordinal nature of CEFR levels; they include absolute- and rank-based measures, as well as agreement-based metrics.

Setup	# Documents	# Simplifications	Accuracy (%)
Baseline (no critic)	209	627	41.18
Regression Critic	212	634	50.00
Ordinal Classifier Critic	196	588	**71.12**

Table 7: Accuracy of simplification across different critic strategies. Each document is simplified to 3 target levels: A1, A2, and B1. A simplification is considered correct if the critic assesses it to match the target level.

- **Spearman's rank correlation coefficient** ($\rho = 0.89$), which suggests a strong monotonic relationship between the predicted and true rankings. A higher ρ value indicates better ordinal agreement.
- **Kendall's Tau** ($\tau = 0.83$), which confirms high ordinal agreement and is especially robust for small test sets.
- **QWK** of 0.87, which reflects substantial agreement between the predicted and true labels, while penalizing larger errors more heavily than smaller ones.

Taken together, these results indicate that the model not only achieves a high proportion of exact matches, but also preserves the ordinal structure of the CEFR scale with strong rank correlation and consistent agreement.

5 LLM-based Text Simplification

In this section, we describe how we use BERT-based difficulty models to assist LLM-based text simplification. These models act as critics to guide the simplification pipeline (see Figure 1):

- The original level of the input text is either assessed by the critic or manually labeled.
- The LLM receives the input text, the target level, and a prompt describing the target level.
- The LLM attempts to generate a simplified version of the text.
- The critic assesses the difficulty level of the output.
- If the target level is reached, the process is terminated.
- Otherwise, the LLM receives its previous output, the achieved level, the target level, and an updated prompt.
- The process is repeated for a maximum of 5 iterations.

When using the BERT-based regression model as a critic, its continuous difficulty scores are mapped to discrete CEFR levels for compatibility with the feedback loop. When using the ordinal classification model, predictions can be used directly without mapping.

If the output is still above the target adjective after 5 iterations, the process stops. At each step, the LLM gets the feedback: `This is your previous attempt to simplify the text to level X. The critic says your simplification is Y. Try harder to reach X.`

5.1 Evaluation with and without Critic

We evaluate three variants of our guided text simplification pipeline: (1) **Baseline**, where the model performs one-shot simplification without critic feedback; (2) **Regression-based (REG) Critic**, where the critic is a BERT-based regression model; and (3) **Ordinal Classification (ORD) Critic**, where the critic is an ordinal classification model.

The evaluation was conducted on 220 manually annotated documents from the test set, whose original levels are above B2. Simplifications were generated to 3 target CEFR levels: A1, A2, and B1. The results are summarized in Table 7.[9]

The baseline system frequently produced simplifications that were off by one CEFR level, with common confusions such as A1 vs. A2 or A2 vs. B1. Adding the REG critic led to a moderate improvement in accuracy (+9%), suggesting that iterative refinement is beneficial. However, the most substantial improvement came from the ORD critic, which achieved 71.12% accuracy—nearly 30 percentage points higher than the baseline.

These results indicate that feedback from the ordinal critic aligns more effectively with the CEFR framework and better guides the LLM toward the target level. Table 8 shows that the LLM generates more correctly simplified outputs with the ordinal critic than with the regression critic, except for the B1 target level—where it tends to generate more

[9] Several simplification pipelines failed due to random reasons; they were not restarted, hence the number of simplification experiments in Table 7 is different for different critics.

Target	Generated	BL	REG	ORD
A1	A2	18.9	**7.9**	8.0
A1	A2-B1	—	8.7	—
A1	B1	5.6	7.9	**0.0**
A1	B1-B2	—	1.6	—
A2	A1	4.6	**0.0**	3.4
A2	B1	9.9	9.8	**1.1**
A2	B1-B2	—	5.5	—
A2	B2	0.0	1.3	**0.0**
B1	A1	1.9	**0.0**	0.9
B1	A2	13.6	**1.4**	15.1
B1	B2	2.9	3.9	**0.0**
B1	B2-C1	—	1.4	—

Table 8: Percentage of generating simplifications at an incorrect level, across three simplification pipelines. Each row indicates "incorrect" simplifications, where the generated level does not match the target level.

Target Level	Critic	Average Iterations	Maximum Iterations
A1	Regression	2.95	5
	Ordinal	**2.71**	5
A2	Regression	2.86	5
	Ordinal	**2.21**	5
B1	Regression	2.74	5
	Ordinal	**1.87**	**4**

Table 9: Average number of simplification iterations per target CEFR level using regression vs. ordinal critic.

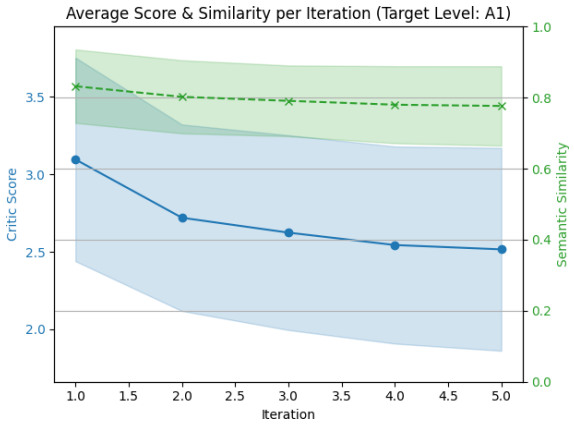

Figure 5: Regression critic with target level A1: average score and cosine similarity per simplification iteration

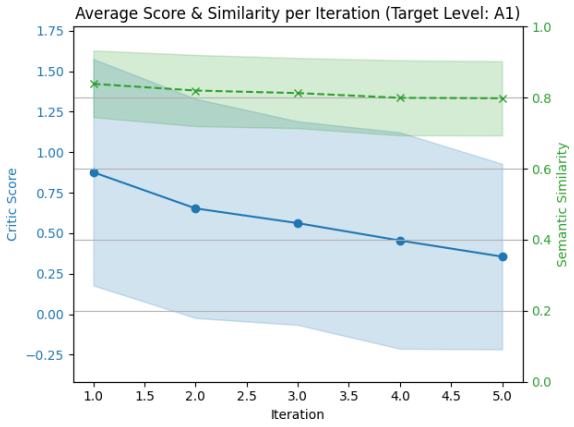

Figure 6: Ordinal critic with target level A1: average score and cosine similarity per simplification iteration

A2-level outputs when guided by the ordinal critic. We also report the average number of iterations required to reach the target level in Table 9: using the ORD critic requires fewer iterations on average, especially for the B1 target.

For all test documents, we tracked the simplification process performed by the LLM by measuring the *intermediate* CEFR levels at each iteration, and the cosine similarity between the intermediate simplification and the original input.[10] Figures 5 and 6 present the mean and standard deviation of difficulty and similarity scores across all documents, with the X-axis representing the iteration number, the left Y-axis showing difficulty scores, and the right Y-axis showing similarity scores.

Note that unlike in Tables 1 and 2, the scores produced by the ORD critic range from 0 to 5. Target level A1 in Figure 5 should be around 1 and in Figure 6—around 0. The plots show that, with the ORD critic, the difficulty of the first intermediate output is already below A2 (i.e., below 1.0), while for the REG critic it remains around B1 (around 3.0). We see a similar gain in performance of ORD over REG critic when the target level is A2 (Figures 8, 9) and B1 (Figures 10, 11) in the Appendix. Cosine similarity stays consistently at or above 0.8 in both pipelines, with slightly higher values when using the ORD critic.

5.2 Evaluation on Parallel Data

We further evaluate our approach using the Parallel Corpus of Standard Finnish–Easy Finnish (Dmitrieva and Konovalova, 2023).[11] The Easy Finnish dataset includes news articles from the Yle archive, and consists of 1,919 manually verified pairs, each comprising an article in Easy Finnish and its corresponding article in Standard Finnish (the *source* article). We extracted 300 document pairs that are longer than 10 words and have Lev-

[10] huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[11] clarino.uib.no/comedi/editor/lb-2022111625

enshtein distance greater than 10, in order to focus on longer contexts that can be meaningfully simplified. As the dataset does not include difficulty annotations, we estimate the difficulty levels of the selected documents using the REG model and the ORD model. Both models indicate that in approximately 65% of the selected pairs, the source document is indeed more difficult than its simplified version.

We processed all source documents through the simplification pipeline once with REG and once with the ORD critic. The outputs generated by the LLM were then compared to the Easy Finnish articles using the SARI metric (Xu et al., 2016), which has been shown to correlate with human judgment of simplicity. The SARI metric is 40.6 for simplification with the BERT-based regression, and 43.1—with the ordinal model.

5.3 Manual Evaluation

An expert in teaching Finnish performed a preliminary manual analysis of the simplification results described above. We randomly selected 24 pairs of source texts and their simplified versions, generated by the pipeline with the REG and ORD models as critics. The annotator's task was to assess whether the simplified text was indeed simpler than the source in terms of lexicon, grammar, sentence structure, and content. Although a more systematic analysis would require a larger sample and deeper investigation, several qualitative patterns emerged. Table 13 and 14 present manually analyzed pairs generated by these two simplification pipelines.

Both pipelines generally demonstrate a strong ability to simplify text: in all 24 cases, at least some parts of each sentence were successfully simplified, and in many cases, the entire sentence was made simpler (e.g., see Example 3 in Table 13). Lexical simplifications include, for instance, "tiivistää vientiponnisteluja" (*intensify export efforts*) → "lisätä vientiä" (*increase exports*), "kehittyvät taloudet" (*developing economies*) → "kehitysmaat" (*developing countries*).

The REG pipeline frequently adds explanatory or contextual information, e.g., by fronting reporting clauses or expounding on the original content (see Examples 8 and 10 in Table 14). While longer texts are not necessarily more complex, such additions may increase the risks of hallucinations. In contrast, the ORD pipeline is often more effective at removing redundant information, resulting in more concise sentences. In some instances, however, the simplifications were simply paraphrases that did not reduce the overall difficulty. Whether a change constitutes a genuine simplification often depends on the reader and may require closer inspection. Both pipelines also occasionally miss clear opportunities for simplification.

In several cases, both models produced "simplified" sentences that were arguably more complex than the original; such cases are highlighted in red in the tables. For example, the verb "tuplaantua" (*to double*) may be easier for L2 learners than the synonym "kaksinkertaistua," even though both are correct. Also, a few minor grammar problems are seen in the outputs, such as incorrect case usage in Finnish noun phrases. In other cases, the simplified sentence introduced factual ambiguities or errors, due to the model's misunderstanding of the context or reference. More details on the results are in Appendix E.

6 Discussion and Conclusion

Our experiments with difficulty models demonstrate that small models can effectively guide text simplification performed by a large language model. Although both BERT-based difficulty models were trained on a *mix* of native and translated data, they significantly improve over the zero-shot baseline.

While the ordinal classifier performs worse on standard classification metrics, it proves more effective as critice in the simplification pipeline. We hypothesize several reasons for this. First, the regression model requires mapping floating-point difficulty scores to discrete CEFR levels, which may lose meaningful distinctions—especially during iterative simplification, where small improvements may be obscured by rounding. Second, regression assumes linear distances between levels, e.g., that the distance between A1 and A2 is equal to the distance between C1 and C2. This assumption is not required by ordinal classification.

An additional benefit of the ORD critic, currently unused, is its ability to estimate *probabilities* for CEFR thresholds—which could be interpreted as a confidence of a text being A1, A2, etc., and enable more fine-grained feedback for the LLM.

In future work, we plan to integrate feature-based and Transformer-based models, enabling the LLM to receive targeted feedback about which linguistic features in the intermediate texts do not match the desired difficulty level.

7 Acknowledgements

This work was supported in part by Business-Finland: Agency for Technology and Innovation, Project "*Easy Language for accessible workplace communication*" (Grant 4173/31/2024). We are grateful to Tiina Onikki-Rantajääskö for her insightful feedback.

8 Limitations and Ethical Considerations

While our results show that difficulty models can effectively guide LLM-based text simplification, several limitations remain. First, the models are trained and evaluated on a small dataset. Working only with Finnish may limit generalizability to other languages or domains. Second, the mapping from regression scores to CEFR levels introduces discretization errors that may obscure nuanced improvements. Third, the simplification pipeline is constrained to five iterations, which may be insufficient for particularly complex texts, and more iterations are expensive to run. Finally, we use a fixed prompt template for LLM interactions; future work could explore adaptive or dynamically generated prompts.

This work focuses on improving language accessibility, particularly for second-language (L2) learners, and aims 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, over-reliance on automated systems may inadvertently reduce the role of human educators in assessing learner needs. Lastly, misuse of simplification systems—e.g., to manipulate or oversimplify critical content—could have adverse effects. We emphasize that these systems should be used as assistive tools, not 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 NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 1–9, Los Angeles, California.

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.

Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka. 2020. Rank consistent ordinal regression for neural networks with application to age estimation. *Pattern Recognition Letters*, 140:325–331.

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 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.

Jue Hou, Maximilian W Koppatz, José Maria Hoya Quecedo, Nataliya Stoyanova, Mikhail Kopotev, and Roman Yangarber. 2019. Modeling language learning using specialized Elo ratings. In *BEA: 14^{th} Workshop on Innovative Use of NLP for Building Educational Applications, ACL: 56th annual meeting of Association for Computational Linguistics*.

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*.

Anisia Katinskaia, Jue Hou, Anh-duc Vu, and Roman Yangarber. 2023. Linguistic constructs represent the domain model in intelligent language tutoring. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 136–144, Dubrovnik, Croatia.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2017. Revita: a system for language learning and supporting endangered languages. In *6th Workshop on NLP for CALL and 2nd Workshop on NLP for Research on Language Acquisition, at NoDaLiDa*, Gothenburg, Sweden.

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2018. Revita: a language-learning platform at the intersection of ITS and CALL. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan.

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.

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 the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 409–420, Edinburgh, Scotland, UK.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Level	Precision	Recall	F1-score	Support
A1	0.70	0.22	0.33	32
A2	0.65	0.63	0.64	98
B1	0.72	0.77	0.75	243
B2	0.14	0.19	0.16	48
C1	0.89	0.93	0.91	399
C2	0.95	0.47	0.63	45

Table 10: Performance of ORD classifier on test set

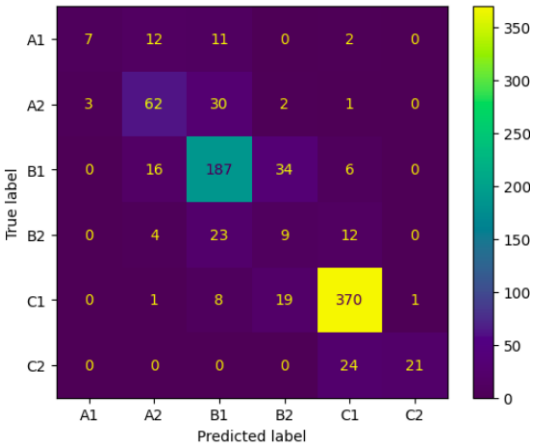

Figure 7: Confusion matrix for ORD classifier

A Baseline classification: feature-based regression

Models (A) and (B) were trained with a regularization strength α = 1.0. For (B), we fit a Ridge regression model to $N_{\text{boot}} = 1000$ bootstrap samples of the training set, each time recording the feature coefficients. For each feature, we calculate the mean and standard deviation of its coefficient across bootstraps. The signal-to-noise ratio is defined as the absolute mean divided by the standard deviation. Features with a signal-to-noise ratio above a threshold (e.g., ≥ 1) are selected, ensuring selection of features with stable and consistently strong effects across resampled datasets.

We fit a Lasso regression model (C), which was employed for feature selection due to its ability to perform both regularization and automatic variable selection. Features with nonzero coefficients are selected, while those with coefficients shrunk to zero are excluded. The regularization parameter α for the Lasso model was selected via cross-validation using the `LassoCV` procedure, optimizing for mean squared error on held-out validation folds.

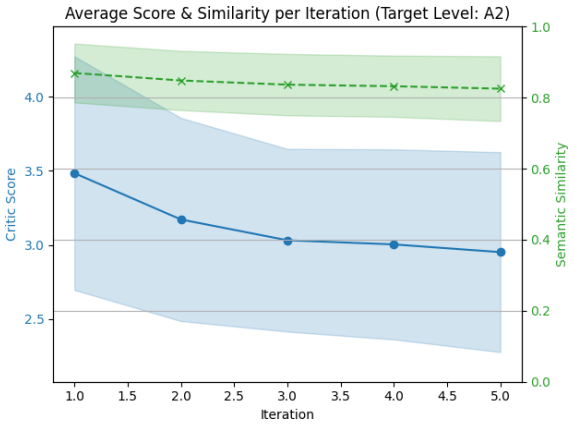

Figure 8: Regression critic with target level A2: average score and cosine similarity per simplification iteration.

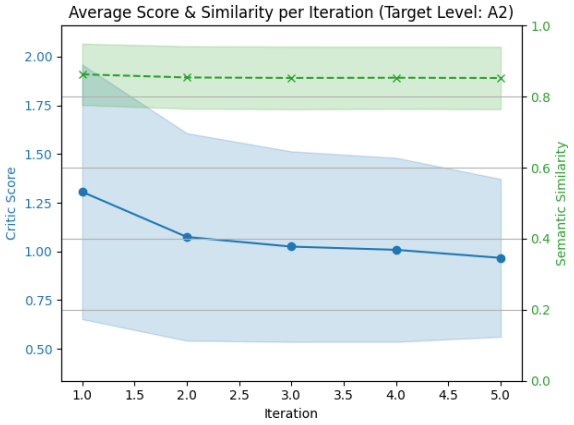

Figure 9: Ordinal critic with target level A2: average score and cosine similarity per simplification iteration.

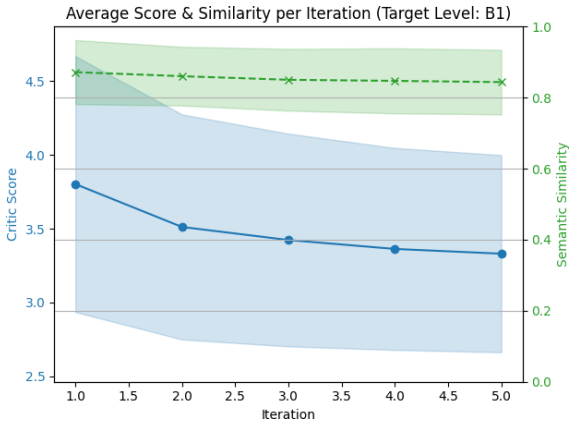

Figure 10: Regression critic, target level B1: average score and cosine similarity per simplification iteration.

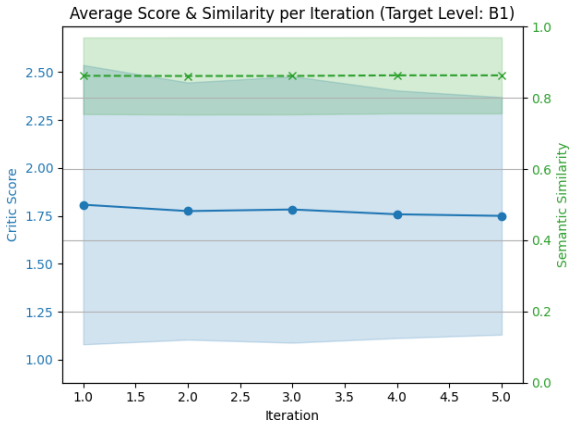

Figure 11: Ordinal critic with target level B1: average score and cosine similarity per simplification iteration.

B Ordinal classification performance

Table 10 and Figure 7 show classification metrics for the BERT-based ordinal classification difficulty model.

C LLM Prompt Templates

Below we list the CEFR-level-specific prompts used to guide GPT-4o in the simplification task. The prompts were formulated based on the definitions of CEFR levels.[12] Each prompt instructs the model to return a JSON object containing a single key `"SIMPLIFICATION"`, with text adapted to the specified proficiency level.

Common Prompt Structure:

> You must always output a JSON object with a `"SIMPLIFICATION"` key. You are an expert in Finnish language and language teaching. You will be given a text in Finnish. Your task is to read it first and then to provide an adaptation into CEFR level `X`. Please do not significantly change the meaning of the input text. [*Level-specific instructions*] This is the text to simplify:
> `{text}`
>
> Imagine that you are teaching a `X` learner, your adaptation should fit their proficiency level.

Level-specific Instructions:

A1 Prompt

> A1 is the simplest level for beginners. The texts in A1 should be simple, with short sentences and easy grammar. The definition of a learner with A1 level is: "Can understand and use familiar every-

[12] www.coe.int/en/web/common-european-framework-reference-languages/table-1-cefr-3.3-common-reference-levels-global-scale

day expressions and very basic phrases aimed at the satisfaction of needs of a concrete type. Can introduce themselves and others and can ask and answer questions about personal details such as where someone lives, people they know and things they have. Can interact in a simple way provided the other person talks slowly and clearly and is prepared to help."

A2 Prompt

A2 is just above the beginner level. The text in A2 should be simple and have relatively easy grammar. The definition of a learner with A1 level is: "Can understand sentences and frequently used expressions related to areas of most immediate relevance (e.g. very basic personal and family information, shopping, local geography, employment). Can communicate in simple and routine tasks requiring a simple and direct exchange of information on familiar and routine matters. Can describe in simple terms aspects of his/her background, immediate environment and matters in areas of immediate need."

B1 Prompt

B1 is an intermediate level. The definition of a learner with B1 level is: "Can understand the main points of clear standard input on familiar matters regularly encountered in work, school, leisure, etc. Can produce simple connected text on topics which are familiar, or of personal interest. Can describe experiences and events, dreams, hopes & ambitions and briefly give reasons and explanations for opinions and plans."

D Feature List

Tables 11 and 12 present the features used to train the feature-based models. Features shown in bold in both tables were selected via bootstrap feature selection. The features include morphophonemic, grammatical, lexical, and syntactic features. Details regarding how these features (or constructs) are detected in text can be found in (Katinskaia et al., 2023).

Consonant gradations features are identified using rule-based methods. The label "*Inactive*" indicates that gradation *does not* occur in the given form, e.g.: compare "nu**kk**ua" (infinitive *to sleep*) and "nu**kk**uu" (3rd person singular *she/he sleeps*)—no gradation. The label "*Active*" indicates that the gradation is present, e.g.: compare "nu**kk**ua" (*to sleep*) and "nu**k**un" (1st person singular *I sleep*)—gradation **kk** → **k**.

Lexical features include groups denoting temporal concepts, e.g.: time of the day in allative case—"aamu**lla**" (*in the morning*), "yö**llä**" (*at night*); but months in inessive case: "elokuu**ssa**" (*in August*), "kesäkuu**ssa**" (*in June*), etc.

Vocabulary bags—from 1 to 10—represent frequency bins constructed from a list of over 20,000 lemmas, sorted by their frequency. The feature `OOV coverage` measures the proportion of words in the text whose lemmas are not found in any frequency bins, averaged over the text length.

E Simplification Results

Table 13 shows results of simplification with LLM-based pipeline guided by the ordinal classification model.

E.1 Simplification Pipeline guided by Ordinal Classification

In Example 1, the simplification was achieved by splitting the original sentence into two, grammar was simplified by replacing conditional mood with indicative: "maksaisi" (*would cost*) → "maksaa" (*costs*); "pienenisivät" (*would decrease*) → "pienenevät" (*decrease*).

Examples 2 and 3 demonstrate the removal of unnecessary information. The simplification in Example 4 resulted in a simpler information structure, as "Brittania" was moved to the beginning of the sentence. However, two sentences were combined into one, which made the overall structure more complex. The lexicon was also simplified; see the blue highlights in the simplified sentence.

Example 5 illustrates a case where the "simplified" version was actually more complex in some respects: "aiotaan nostaa" (*is going to be increased*) → "suunnitellaan korotettavaksi" (*is planned to be raised*); "prosenttia" (*percent*) → "prosentilla" (*by percent*).

In Example 7, the grammar was improved: "katsoo päätöksessään" (*considers in its decision*) → "päätti" (*decided*); "ettei" (*that not—contracted*)

→ “että ei” (*that not—expanded*); “ei ollut syytä epäillä” (*had no reason to suspect*) → “ei voinut epäillä” (*could not suspect*). However, some parts were made more difficult: “rekrytointia hoitanut mies” (*the man who handled the recruitment*) → “rekrytoinnista vastannut mies” (*the man responsible for recruitment*).

E.2 Simplification Pipeline guided by Regression

Although the simplification in Example 8 improved contextual information—by adding “puolue” (*party*) and “lakialoite” (*legislative initiative*)—it also contains an error in orthography (a missing hyphen between “Perussuomalaiset” and “puolue”). Additionally, it introduces unnecessary and grammatically complex information, such as “lain voimaantulon jälkeen sen aiheuttamat [kustannukset]” (*the [costs] caused by it after the law comes into force*).

Example 9 demonstrates changes that made the lexicon more difficult: “yli” (*over*) → “ylittäen” (*exceeding*); “on kasvanut paljon” (*has grown a lot*) → “lisääntynyt huomattavasti” (*increased significantly*).

The simplified text in Example 10 illustrates the removal of the unnecessary word “käytännössä” and the simplification of some grammatical forms: “voisivat” (*could*) → “voivat” (*can*); “tiivistää vientiponnisteluja” (*intensify export efforts*) → “parantaa yhteistyötä viennissä” (*improve cooperation in exports*). However, it also introduces new information not present in the source (see red highlight).

The red highlights in Examples 11–13 indicate cases where the forms were made lexically, grammatically, or syntactically more complex than in the source texts.

Feature Set 1	Feature Set 2
Comparative adjective form	**Consonant gradation (A type, nouns, active)**
Positive adjective form	**Consonant gradation (A type, nouns, inactive)**
Superlative adjective form	**Consonant gradation (A type, verbs, active)**
Abessive case	Consonant gradation (A type, verbs, inactive)
Ablative case	Consonant gradation (“lki” A type, active)
Accusative case	**Consonant gradation (“lki” A type, inactive)**
Adessive case	Consonant gradation (A type ending with “uku”, active)
Allative case	Consonant gradation (A type ending with “uku”, inactive)
Comitative case	Consonant gradation (B type, nouns, active)
Elative case	**Consonant gradation (B type, nouns, inactive)**
Essive case	**Consonant gradation (B type, verbs, active)**
Genitive case	**Consonant gradation (B type, verbs, inactive)**
Illative case	Compound noun inflection
Inessive case	**Lists of confusable nouns**
Instructive case	**Noun paradigm “aihe”**
Nominative case	Noun paradigm “bussi”
Partitive case	Noun paradigm “kala”
Translative case	Noun paradigm “kannel”
Clitics of emphasis	**Noun paradigm “koditon”**
Clitics of negation	**Noun paradigm “koira”**
Clitics of question	**Noun paradigm “kysymys”**
Clitics han	**Noun paradigm “maa”**
Clitics pa	**Noun paradigm “manner”**
Construction with differen tactors	**Noun paradigm “mansikka”**
Construction of type “ESSA” (Temporaalirakenne)	**Noun paradigm “nainen”**
Construction with “Että”, perfect	**Noun paradigm “olut”**
Construction with “Että”, present	Noun paradigm “ovi”
Construction with “Että”, different actors	**Noun paradigm “puhelin”**
Construction with “Että”, same actors	**Noun paradigm “talo”**
Existential construction	**Noun paradigm “uusi”**
Existential construction, negative	Noun paradigm “uutuus”
Existential construction, positive	**Noun paradigm “valas”**
Negative construction	**Noun possessive suffixes**
Necessity Construction	Noun of time
Permission Construction	**Noun of time (day, essive)**
Construction of possession	Nouns of time (hour, adessive)
Construction of possession, negative	**Noun of time (month, inessive)**
Construction of possession, positive	**Noun of time (season, adessive, essive)**
Construction with same actors	**Noun of time (time of the day, adessive, essive)**
Construction with “TUA” (Temporaalirakenne)	Noun of time (week, adessive)
Government by adjective	Noun of time (year, essive)
Government by noun	**Plural number**
Government by verb	**Singular number**
Government by adposition	Cardinal numeral
Infinitive 1	**Cardinal numeral, long**
Infinitive 2	**Cardinal numeral, short**
Infinitive 3	Ordinal numeral
Infinitive 4	Ordinal numeral, long
Infinitive 5	Ordinal numeral, short
Infinitive TUA	**Agentive participle**
Conditional mood	**Perfect active participle**
Conditional passive mood	**Perfect passive participle**
Imperative mood	**Participle with possessive suffixes**
Indicative mood	Present active participle
Potential mood	**Present passive participle**
Potential passive mood	**Person 1**
Possessiveness	**Person 2**
Negative polarity	Person 3
Average dependency tree depth	**OOV coverage**

Table 11: Combined linguistic feature sets for the feature-based regression model.

Feature Set 3
Demonstrative pronoun
Indefinite pronoun
Indefinite pronoun “joku”
Indefinite pronoun “kukaan”
Interrogative pronoun
Interrogative pronoun “kumpi”
Personal pronoun
Reflexive pronoun
Relative pronoun
Active object
Object of infinitive
Genitive modifier
Object of imperative
Object of passive
Object in ablative of “sense” verbs
Object in ablative of “source” verbs
Object in adessive “instrument” verbs
Object in allative of “sense” verbs
Object in allative of “communication” verbs
Object in allative of “possession” verbs
Complement construction
Object of negative
Object of verb of “feeling” (Tunnekausatiivi)
Imperfect tense
Imperfect indicative tense
Perfect tense
Perfect Indicative tense
Pluperfect tense
Pluperfect indicative tense
Present tense
List of confusable verbs
Infinitive
Infinitive bases
Verb paradigm “haluta”
Verb paradigm “juosta”
Verb paradigm “nousta”
Verb paradigm “saada”
Verb paradigm “sanoa”
Verb paradigm “tarvita”
Verb paradigm “tehdä”
Verb paradigm “tulla”
Verb paradigm “vanheta”
Participle
Active voice
Passive voice
Passive imperfect voice
Passive perfect voice
Passive pluperfect voice
Passive present voice
Average number of POS
Average sentence length
Vocabulary bag 1
Vocabulary bag 2
Vocabulary bag 3
Vocabulary bag 4
Vocabulary bag 5
Vocabulary bag 6
Vocabulary bag 7
Vocabulary bag 8
Vocabulary bag 9
Vocabulary bag 10

Table 12: Linguistic features used for the feature-based regression model.

Ordinal classifier as critic:
ID: 1 **Source:** Perussuomalaiset arvioivat, että “Lex Lindström” maksaisi noin 20-30 miljoonaa euroa, ja menot pienenisivät vuosittain. **Level:** C1 **Simplified:** Perussuomalaiset-puolue arvioi, että “Lex Lindström” maksaa noin 20-30 miljoonaa euroa. He arvioivat, että kulut pienenevät vuosittain. **Level:** B2
ID: 2 **Source:** Jo täällä asuvilla maahanmuuttajilla on Nafisa Yeasminin mukaan tärkeää rooli uusien tulijoiden sopeutumisessa Suomeen. — Me tunnemme jo suomalaisen kulttuurin ja kielen. **Level:** B2 **Simplified:** Suomessa asuvat maahanmuuttajat auttavat uusia ihmisiä. — Me tunnemme Suomen kulttuurin ja kielen. **Level:** A2
ID: 3 **Source:** Käytännössä Pohjoismaat voisivat esimerkiksi tiivistää puhtaan teknologian vientiponnisteluja kehittyvissä talouksissa. **Level:** B2 **Simplified:** Pohjoismaat voisivat esimerkiksi lisätä puhtaan teknologian vientiä kehitysmaihin. **Level:** A2
ID: 4 **Source:** Siirtymäkauden ajan Britannia säilyy EU:n sisämarkkinoilla, maksaa maksuja ja soveltaa EU-lainsäädäntöä. Se ei kuitenkaan osallistu päätöksentekoon. **Level:** B2 **Simplified:** Britannia on vielä vähän aikaa EU:ssa mukana, noudattaa EU:n sääntöjä ja maksaa EU:lle, mutta ei voi osallistua päätöksiin. **Level:** A2
ID: 5 **Source:** Korkeinta kokopäivähoidon maksua aiotaan nostaa peräti 22 prosenttia. Tällä hetkellä ylin maksu on 283 euroa kuukaudessa ensimmäisestä lapsesta. **Level:** C1 **Simplified:** Korkeinta kokopäivähoidon maksua suunnitellaan korotettavaksi jopa 22 prosentilla. Tällä hetkellä maksetaan enintään 283 euroa kuukaudessa ensimmäisestä lapsesta. **Level:** B2
ID: 6 **Source:** Oikeus katsoo päätöksessään, ettei rekrytointia hoitaneella miehellä ollut syytä epäillä saamaansa ohjetta. — Olemme tyytyväisiä. **Level:** C1 **Simplified:** Oikeus päätti, että rekrytoinnista vastannut mies ei voinut epäillä saamaansa ohjetta. — Olemme tyytyväisiä. **Level:** B1
ID: 7 **Source:** Suomalaisten testamenttilahjoittajien sydäntä lähellä on erityisesti lasten ja naisten hyvinvointi ja koulutus niin Suomessa kuin kehitysmaissa. — Lahjoittajat sanovat, että kun satsaa tyttöihin ja naisiin, niin yhteiskuntakin muuttuu vahvasti. **Level:** B2 **Simplified:** Suomalaiset lahjoittajat haluavat auttaa lapsia ja naisia Suomessa ja muissa maissa. He sanovat, että kun auttaa tyttöjä ja naisia, myös yhteiskunta paranee. **Level:** A2

Table 13: Pairs of source texts and their simplified versions. Simplifications were guided by the ordinal classifier. Underlined text in the source indicates which parts were changed; blue highlights mark successful simplification. Red highlights parts which are erroneous or harder than the source.

Regression model as critic:
ID: 8 **Source:** Perussuomalaiset arvioivat, että "Lex Lindström" maksaisi noin 20-30 miljoonaa euroa, ja menot pienenisivät vuosittain. **Level:** B2-C1 **Simplified:** Perussuomalaiset puolue on arvioinut, että 'Lex Lindström' — lakialoitteen taloudellinen vaikutus olisi noin 20-30 miljoonaa euroa. Lain voimaantulon jälkeen sen aiheuttamat kustannukset pienentyisivät vuosittain. **Level:** B2-C1
ID: 9 **Source:** Kaiken kaikkiaan lemmikkien lääkintä — ja hoitomenot ovat tuplaantuneet kymmenessä vuodessa yli 200 miljoonaan euroon. Lemmikkien määrä on kasvanut samassa ajassa paljon hitaammin. **Level:** B2-C1 **Simplified:** Yhteensä lemmikkien lääkintä — ja hoitokulut ovat kasvaneet kaksinkertaisiksi kymmenessä vuodessa ylittäen 200 miljoonaa euroa. Samaan aikaan lemmikkien määrä on lisääntynyt huomattavasti hitaammin. **Level:** B2
ID: 10 **Source:** Käytännössä Pohjoismaat voisivat esimerkiksi tiivistää puhtaan teknologian vientiponnisteluja kehittyvissä talouksissa. **Level:** B2-C1 **Simplified:** Pohjoismaat voivat parantaa yhteistyötä puhtaan teknologian viennissä kehittyviin talouksiin esimerkiksi jakamalla innovaatioita ja edistämällä yhteisiä hankkeita. **Level:** B2-C1
ID: 11 **Source:** Korkeinta kokopäivähoidon maksua aiotaan nostaa peräti 22 prosenttia. Tällä hetkellä ylin maksu on 283 euroa kuukaudessa ensimmäisestä lapsesta. **Level:** B2-C1 **Simplified:** Kokopäivähoidon korkeinta maksua suunnitellaan nostettavaksi jopa 22 prosenttia. Nyt korkein maksu on 283 euroa kuukaudessa ensimmäisestä lapsesta. **Level:** B2
ID: 12 **Source:** Jämsä ja Liukkokin ovat muuttaneet kulutustottumuksiaan. — Kaupassa voin jättää muovipussin ostamatta ja ruokaillessa syön mieluummin kasviksia tai kanaa naudanlihan sijaan, Jämsä kertoo. **Level:** B1-B2 **Simplified:** Jämsä ja Liukko ovat muuttaneet tapojaan kuluttaa. Jämsä sanoo, että kaupassa hän voi päättää olla ostamatta muovipussia ja ruokailutilanteissa hän valitsee mieluummin kasvikset tai kanan naudanlihan sijasta. **Level:** B1-B2
ID: 13 **Source:** Suomalaisten testamenttilahjoittajien sydäntä lähellä on erityisesti lasten ja naisten hyvinvointi ja koulutus niin Suomessa kuin kehitysmaissa. — Lahjoittajat sanovat, että kun satsaa tyttöihin ja naisiin, niin yhteiskuntakin muuttuu vahvasti. **Level:** B2 **Simplified:** Suomalaiset testamenttilahjoittajat haluavat parantaa lasten ja naisten elämää ja koulutusta Suomessa ja kehitysmaissa. He uskovat, että kun tukee tyttöjä ja naisia, koko yhteiskunta voi kehittyä vahvemmaksi. **Level:** B1

Table 14: Pairs of source texts and their simplified versions. Simplifications were guided by the regression model. Underlined text in the source indicates which parts were changed; blue highlights mark successful simplification. Red highlights parts which are erroneous or harder than the source.