Testing relevant linguistic features in automatic CEFR skill level classification for Icelandic

Isidora Glišić
University of Iceland
Reykjavík, Iceland
isidora@hi.is

Caitlin Laura Richter
Reykjavík University
Reykjavík, Iceland
caitlinr@ru.is

Anton Karl Ingason
University of Iceland
Reykjavík, Iceland
antoni@hi.is

Abstract

This paper explores the use of various linguistic features to develop models for automatic classification of language proficiency on the CEFR scale for Icelandic, a low-resourced and morphologically complex language. We train two classifiers to assess skill level of learner texts. One is used as a baseline and takes in the original unaltered text written by a learner and uses predominantly surface features to assess the level. The other uses both surface and other morphological and lexical features, as well as context vectors from transformer (IceBERT). It takes in both the original and corrected versions of the text and takes into account errors/deviation of the original texts compared to the corrected versions. Both classifiers show promising results, with baseline models achieving between 62.2-67.1% accuracy and dual-version between 75-80.3%.

1 Introduction

Language skill level assessment is a critical component in language education and testing, and accurate and scalable methods for assessing skill levels can facilitate personalized learning, enhance testing systems, and contribute to linguistic research. However, automating this process presents significant challenges, especially for low-resourced languages such as Icelandic. In this paper, we present findings from an ongoing study focused on using linguistic features to train models for automatic skill level assessment in Icelandic as a second language (L2) texts on the CEFR scale, a widely adopted framework in language education. By focusing on Icelandic we aim to contribute to the growing body of work on underrepresented languages in natural language processing (NLP) and highlight the importance of broadening research efforts beyond high-resourced languages.

The Icelandic L2 Error Corpus (IceL2EC), published in 2022 (Ingason et al., 2022), has served as a foundational dataset for analyzing features associated with the CEFR skill level. In particular, manually corrected text versions and error annotation have shown a high value in predicting proficiency levels through machine learning approaches (Glišić, 2023). To explore automatic assessment further, this study builds on IceL2EC and includes additional unpublished texts sourced from the University of Iceland. Using this combined dataset, several models were trained to test the efficacy of various features for the assessment of the CEFR skill level. We present the results of baseline models using K-nearest neighbors (KNN) algorithm which uses the learners' original texts and basic linguistic features, and "dual-version" models (logistic regression - LR) which integrate the corrected versions of the data, and more complex features.

Key research questions addressed in this study include: (1) How accurately can linguistic features predict writing skill levels in Icelandic L2 texts, and (2) To what extent do corrected texts and advanced models like IceBERT (Snæbjarnarson et al., 2022) contribute to improved classification performance? We incorporate surface features, morphological and lexical elements, and IceBERT-derived context vectors to provide a comprehensive approach to automatic skill level assessment.

The paper is organized as follows: Section 2 provides a background on L2 Icelandic, the CEFR scale, and automatic skill level detection. Sections 3 and 4 detail our models and evaluation metrics, while Section 5 presents the experimental results.

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 217–222
March 3-4, 2025 ©2025 University of Tartu Library

2 Background

Icelandic stands out among lesser spoken languages for its relatively robust digital resources (Nikulásdóttir et al., 2020; Nikulásdóttir et al., 2022). However, the resources available for Icelandic as a learner language (L2) are sparse mainly due to L2 Icelandic being a relatively recent phenomenon and collecting written data for L2 Icelandic is challenging. However, in recent years the number of foreign nationals in Iceland has surged. In the mid-1990s, only 2% of Iceland's population were first-generation immigrants; by early 2023, this number reached approximately 17.3%. (Hagstofa Íslands, 2023). This demographic shift has heightened the importance of developing resources for L2 Icelandic. An essential aspect of teaching and assessing a second language is measuring learner skill level. The CEFR standardizes skill level assessment with a six-level scale (A1 to C2), focusing on communicative competencies rather than specific linguistic structures (Council of Europe, 2018). IceL2EC, developed under a government-sponsored language technology initiative (see Nikulásdóttir et al., 2020), is the primary resource available for investigating CEFR-labeled learner errors and interlanguage features; data for a new learner corpus is currently being collected to build on these foundations.

Automatic classification of skill level in written texts remains challenging due to the subjective nature of language proficiency scales like the CEFR. A critical component in skill assessment involves the selection of linguistic features that effectively capture learners' proficiency. Thus, with learner corpora and error tagging, researchers can identify relevant linguistic patterns that correspond to specific CEFR levels. In English, for example, accuracy rates for automatic CEFR classification range from 62.7% to 83.8% (Kerz et al., 2021). Using features derived from lexical, morphological, and syntactic patterns, classifiers like logistic regression and more advanced approaches have achieved promising results in multilingual proficiency assessment tasks, as seen in studies with L2 German, Swedish, and Estonian (Kerz et al., 2021; Vajjala and Lõo, 2014). Importantly, model evaluation metrics must consider the proximity between CEFR levels, recognizing that misclassifications between adjacent levels (e.g., C1 and B2) are less severe than those between distant levels (e.g., C1 and A1). Additionally, language proficiency assessment carries significant implications, as its results can influence the learner's educational and professional opportunities. In this context, predicting a higher level is generally less harmful to the learner than predicting a lower one.

3 Model training

This study establishes preliminary models for automated skill level classification. Baseline models, utilizing only original texts, are compared with dual-version models that use both original and corrected texts. Feature-based approaches yield high prediction accuracy, especially for morphologically rich languages (see Weiss et al., 2021, Reynolds, 2016), and this study combines surface, morphologic, and lexical features, as suggested in recent research (see Pilán and Volodina, 2018, Yekrangi, 2022, Curto et al., 2015), as well as combining context vectors from transformers and perplexity score, typically used to evaluate the performance of language models. For Icelandic proficiency classification, we adapt representative models for these approaches, whose established performance in other languages provides additional context for the results we observe in Icelandic. In this section we introduce the dataset, models and features selected for our task.

3.1 Dataset

Training data consists of IceL2EC, the first published corpus of L2 Icelandic which has 101 student essays categorized by skill level, manually corrected and annotated for errors. Initial CEFR level labels were made based on the students' academic progress and assessment by a human annotator (Glišić and Ingason, 2022). To validate these levels, inter-annotator agreement was reached with five experienced Icelandic L2 instructors, and the final level assignments reflect the averaged ratings from this team. The corpus includes writing assignments of varying lengths, from 150–200 word beginner texts to several thousand words advanced essays, leading to an uneven distribution of data across skill levels. To create a more balanced dataset for model training, 83 additional unpublished texts from the Practical Diploma Program in Icelandic (A1/A2) were added. Additionally, the texts were cleaned by removing all non-Icelandic sentences, and longer texts (in particular full BA and MA theses) were chunked into 40–50 sentences segments to fit

BERT maximum token length, resulting in a total of 276 texts with a more even training support across levels.

Level	Texts	Total Words	Sentences
A1	73	7,820	913
A2	38	10,204	913
B1	37	21,960	1,229
B2	31	22,457	1,052
C	97	84,730	3,873

Table 1: Distribution of data for each level

Given that the ongoing project on Icelandic CEFR alignment currently emphasizes levels A1 through B2, the advanced levels C1 and C2 were merged into a single advanced category, labeled "C." The final dataset thus spans a five-level scale, with A1 and A2 representing beginner, B1 and B2 intermediate, and C advanced levels, as depicted in Table 1.

3.2 Feature selection for baseline

Baseline models used only features that can be computed from shallow analysis of the text. Minimal feature sets were selected from those suggested by (Yekrangi, 2022), inspired by older formulas for assessing text complexity. The total length of the text, along with features like type-token ratio that are considered excessively influenced by it according to consensus in cross linguistic literature (McCarthy and Jarvis, 2010), were excluded from baseline models as confounds.

Baseline-Minimal uses two features: average word length, the number of letters per token; and HD-D, the hypergeometric distribution of lexical (word) diversity, an alternative to type-token ratio (McCarthy and Jarvis, 2010).

Baseline-Lemma requires lemmatisation (stemming) (Ingason et al., 2008) and a frequency list of the language's vocabulary (Arnardóttir and Ingason, 2023), but no further language processing technology or resources. PoS-tagging and lemmatization for all models tested was conducted with ABL Tagger (Steingrímsson et al., 2019) and the Nefnir lemmatizer (Ingólfsdóttir et al., 2019). Some features expect CEFR aligned vocabularies, but lacking one for Icelandic, this implementation assigns the 1000 most frequent words to A1, and so on, following the teaching resource RÚV Orð[1].

The features included are: average word length and HD-D as in Baseline-Minimal; ATTR; CLI; average vocabulary level of tokens' lemmas, advanced vocabulary percentage, the percentage of the text's lemmas not in CEFR A or B inventories; and Dale-Chall readability score (DCRS), a formula combining the proportion of "difficult" tokens (lemmas not in A-B1) with the average number of words per sentence.

3.3 Feature selection for dual version models

The dual-version models incorporate two versions of the data — original and corrected — and include surface features, morphological features derived from PoS-tagging and lemmatization, lexical diversity metrics (word frequencies, tf-idf weighted words), and NLP-based features like contextual embeddings from transformers and text perplexity extracted from originals. Key features that highlight differences between text versions are cosine similarity and average error count per sentence.

Dual-ling uses linguistic features primarily inspired by Pilán and Volodina's feature set (2018), with several adaptations. Key features include average sentence length, percentage of long words (over six characters), average error counts per sentence, and cosine similarity between original and corrected texts; morphological features include proportions of pronouns, past participles, conjunctions, articles, and subjunctive forms; lexical features include average lemma count, average vocabulary level of lemmas, and tf-idf weighted terms for uni-, bi-, and trigrams in both original text and PoS tags.

Dual-expand is supplemented by incorporating IceBERT, an Icelandic language model based on the BERT architecture, which estimates word probability given its context (Snæbjarnarson et al., 2022). The IceBERT-igc feature selection pipeline was applied to derive embeddings and extract relevant features from the dataset, and the model was used to calculate the perplexity of the original texts.

4 Evaluation

Both baseline and dual-version models were tested on several algorithms, including linear regression, SVM, KNN, LR, and MLP. After initial testing, K-nearest neighbors (K=10) was viewed for baseline evaluations, while logistic regression was selected

[1] https://ord.ruv.is/

for the dual-version evaluations. Each model and feature set was assessed using an 80/20 train-test split, with stratified sampling. It was repeated 1000 times with different random splits, and the reported metrics represent averages of these runs. Accuracy, as the percentage of correct classifications, is sensitive to data distribution and may overlook false positives, disproportionately affecting smaller classes. Additionally, accuracy does not account for the "distance of prediction," where predicting C1 instead of B2, for example, is a less severe error than predicting A1 instead of B2. For a more comprehensive evaluation, F1 scores were also calculated to provide a balance between precision and recall. Alongside exact accuracy, we assessed adjacent accuracy, i.e. also viewing predictions within one level above or below the true level (e.g., A2 predicted as either A1, A2, or B1) as correct. This metric reflects the CEFR scale's flexibility and the frequent disagreements between human evaluators.

5 Results

Baseline models showed varied performance, with the highest exact accuracy achieved by the Baseline-Lemma KNN model, which recorded 67.1% exact accuracy and 89.7% with adjacent accuracy. Interestingly, the linear regression model, although performing lower on exact matches at 63.5%, had the highest adjacent accuracy among all models, achieving 97.6%. This suggests that while linear regression may struggle with precise classification, it is particularly effective at capturing a close approximation to the true level. All tested models varied in performance between CEFR levels, with levels A1 and C showing better performance across the board, as seen in Tables 2 and 3.

Class	Precision	Recall	F1	Support
A1	0.79	0.97	0.87	15
A2	0.59	0.36	0.43	8
B1	0.33	0.30	0.30	7
B2	0.42	0.23	0.28	6
C	0.75	0.84	0.79	20

Table 2: Average Performance Statistics for the Baseline-Lemma KNN Model

Table 4 presents a comparative overview of the average accuracy (exact and adjacent) across models. The Dual-Ling model, which combines original and corrected text features without IceBERT embeddings, achieved the highest exact accuracy at 80.3% and 96.4% when including the one-level deviation. In addition, the introduction of lexical features, especially tf-idf weights, notably improved the models' performance, with tf-idf for PoS tags alone contributing an average 4% boost in accuracy.

Class	Precision	Recall	F1	Support
A1	0.89	1.00	0.94	16
A2	0.75	0.50	0.60	6
B1	0.78	0.70	0.74	10
B2	1.00	0.11	0.20	9
C	0.62	1.00	0.77	15

Table 3: Average Performance Statistics for Dual-expand LR Model

Model	Exact(%)	Adjacent(%)
Baseline-Minimal	62.2	85.9
Baseline-Lemma	67.1	89.7
Dual-Ling	80.3	96.4
Dual-Expand	75.0	94.6

Table 4: Comparative accuracy of KNN and LR models, exact and adjacent

6 Discussion

Findings from this study show that baseline models can achieve moderate classification accuracy, with the Baseline-Lemma KNN model reaching the highest baseline performance (67.1% exact, 89.7% adjacent). Models performed best at A1 and C levels, likely due to both their highest data support as well as distinctiveness, while B1 and B2 had lower F1 scores, reflecting their similarity to adjacent levels and greater classification difficulty. This level of performance aligns with accuracy rates reported for other languages, suggesting that even simple feature sets can yield effective results for Icelandic, despite the limited resources available for L2. We also note that these results are robust to variation in the feature set, as e.g. other two-feature models with different lexical diversity measures in place of HD-D perform about as well as Baseline-Minimal, without clear preference among a few reasonable alternatives at least within present statistical power.

Enhanced models using dual versions and more sophisticated linguistic features outperformed baseline models, with the exception of the linear regression baseline model's strong adjacent accuracy (97.6%). The Dual-ling model demonstrated the highest exact accuracy at 80.3% and 96.4% for adjacent. This supports the hypothesis that including corrected versions can improve classifier accuracy by providing insights into corrective changes that reveal interlanguage patterns. Additionally, incorporating lexical features, specifically tf-idf weights for both lexical terms and PoS tags, proved influential in boosting accuracy, underscoring the importance of lexical diversity and usage patterns in prediction. The absence of IceBERT embeddings in the top-performing Dual-ling model suggests that raw contextual embeddings may not be essential for achieving strong performance in this task. However, it remains a question for future research whether using IceBERT for classification or extracting embeddings comparatively could improve results in a more balanced dataset or with a larger corpus.

7 Conclusion

We have demonstrated that integrating surface and deeper linguistic features is notably effective in skill level classification, showing that a blend of lexical, morphological, and contextual data can meaningfully reflect learner proficiency. We found that baseline models performed moderately well, with the Baseline-Lemma KNN model achieving the highest exact accuracy (67.1%) and 89.7% when adjacent accuracy was considered. The Dual-Ling model, relying on both original and corrected text features, achieved the highest overall performance with 80.3% exact and 96.4% adjacent accuracy. These findings have significant implications for future automated tools assessing Icelandic learners' skill levels. However, a challenge with the CEFR lies in its broad descriptors, which lack specific grammatical and lexical competencies for each level, making it difficult to map concrete linguistic features directly to skill levels. The forthcoming Icelandic learner corpus, specifically designed for skill level analysis with balanced data, marks an important step forward. It promises to provide an empirically grounded dataset for further development of automated tools, enabling more accurate skill level assessments.

References

Þórunn Arnardóttir and Anton Karl Ingason. 2023. Frequency lists for icelandic 23.06. CLARIN-IS.

Council of Europe. 2018. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors*. Council of Europe Publishing, Strasbourg.

Pedro Curto, Nuno Mamede, and Jorge Baptista. 2015. Automatic text difficulty classifier - assisting the selection of adequate reading materials for european portuguese teaching. pages 36–44.

Isidora Glišić. 2023. Towards automated icelandic skill level evaluation: A deep dive into l2 error corpus patterns and classification. Master's thesis, University of Iceland, September. Master's Thesis.

Isidora Glišić and Anton Karl Ingason. 2022. The nature of icelandic as a second language: An insight from the learner error corpus for icelandic. pages 23–33.

Hagstofa Íslands. 2023. Yfirlit mannfjölda. Accessed: 2023-06-04.

Anton Karl Ingason, Sigrún Helgadóttir, Hrafn Loftsson, and Eiríkur Rögnvaldsson. 2008. A mixed method lemmatization algorithm using a hierarchy of linguistic identities (holi). In *Advances in natural language processing: 6th international conference, GoTAL 2008 Gothenburg, Sweden, August 25-27, 2008 Proceedings*, pages 205–216. Springer.

Anton Karl Ingason, Lilja Björk Stefánsdóttir, Þórunn Arnardóttir, Xindan Xu, Isidora Glišić, and Dagbjört Guðmundsdóttir. 2022. The icelandic l2 error corpus (IceL2EC) 1.3 (22.10). CLARIN-IS.

Svanhvít Lilja Ingólfsdóttir, Hrafn Loftsson, Jón Friðrik Daðason, and Kristín Bjarnadóttir. 2019. Nefnir: A high accuracy lemmatizer for Icelandic. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 310–315, Turku, Finland. Linköping University Electronic Press.

Elma Kerz, Daniel Wiechmann, Yu Qiao, Emma Tseng, and Marcus Ströbel. 2021. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. In *Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 199–209, Online. Association for Computational Linguistics.

Philip M McCarthy and Scott Jarvis. 2010. "mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment". *Behavior research methods*, 42(2):381–392.

Anna Björk Nikulásdóttir, Þórunn Arnardóttir, Starkaður Barkarson, Jón Guðnason, Þorsteinn Daði

Gunnarsson, Anton Karl Ingason, Haukur Páll Jónsson, Hrafn Loftsson, Hulda Óladóttir, Eiríkur Rögnvaldsson, et al. 2022. Help yourself from the buffet: National language technology infrastructure initiative on clarin-is. In *CLARIN Annual Conference*, pages 109–125.

Anna Björk Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Eiríkur Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language technology programme for icelandic 2019-2023. In *Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020)*, pages 3414–3422.

Ildikó Pilán and Elena Volodina. 2018. Investigating the importance of linguistic complexity features across different datasets related to language learning. In *Proceedings of the Workshop on Linguistic Complexity and Natural Language Processing*, pages 49–58, Santa Fe, New-Mexico. Association for Computational Linguistics.

Robert Reynolds. 2016. Insights from Russian second language readability classification: complexity-dependent training requirements, and feature evaluation of multiple categories. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 289–300, San Diego, CA. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus - a recipe for good language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Vésteinn Snæbjarnarson, Haukur Simonarson, Pétur Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjálmur Þorsteinsson, and Hafsteinn Einarsson. 2022. A warm start and a clean crawled corpus – a recipe for good language models.

Steinþór Steingrímsson, Örvar Kárason, and Hrafn Loftsson. 2019. Augmenting a BiLSTM tagger with a morphological lexicon and a lexical category identification step. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1161–1168, Varna, Bulgaria. INCOMA Ltd.

Sowmya Vajjala and Kaidi Lõo. 2014. Automatic CEFR level prediction for Estonian learner text. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 113–127, Uppsala, Sweden. LiU Electronic Press.

Zarah Weiss, Xiaobin Chen, and Detmar Meurers. 2021. Using broad linguistic complexity modeling for cross-lingual readability assessment. In *Proceedings of the 10th Workshop on NLP for Computer Assisted Language Learning*, pages 38–54, Online. LiU Electronic Press.

Aryan Yekrangi. 2022. "leveraging simple features and machine learning approaches for assessing the cefr level of english texts". Master's thesis, "University of Eastern Finland".