Boglárka Vermeki

Modelling Language Proficiency with Puli-BERT-Large: A Case Study on CEFR Classification in Hungarian Learner Texts

1 Introduction

This study explores how well a culturally and linguistically tailored transformer model, the Puli-BERT-Large model (Yang et al., 2023) can classify Hungarian learner texts according to the Common European Framework of Reference for Languages (CEFR) proficiency levels. By combining transformer-based modelling with an annotated learner corpus, we assess how well Puli-BERT-Large can emulate human proficiency judgments and where they diverge. Beyond performance metrics, we analyse misclassifications to gain insight into model behaviour, learner language patterns, and the pedagogical assumptions embedded in both AI and CEFR frameworks.

The following research questions guide this investigation:

- **RQ1**: How accurately can a Hungarian-specific transformer model (Puli-BERT-Large) classify learner texts by CEFR level?
- **RQ2**: What kinds of linguistic features (syntactic, lexical, discourse-level) correlate with misclassification between adjacent CEFR levels (e.g., A2 - B1)?
- **RQ3**: How do surface-level linguistic features influence Puli-BERT-Large's proficiency predictions, and where does this lead to misclassification?

2 Dataset and Method

This study utilises two primary sources of learner language: the MID Learner Corpus (Baumann et al., 2020) and the MagyarOK Open Corpus (Szita, 2021). The MID Learner Corpus contains authentic written and transcribed spoken texts produced by learners of Hungarian as a second language, collected by the Corpus Linguistics and Language Pedagogy Research Group. The corpus currently includes texts produced by 347 learners, amounting more than 300,000 words. The written subcorpus forms the majority of the corpus. Texts were submitted both in handwritten and digital formats; handwritten texts were transcribed with attention to both content and form to preserve their analytical value. Each entry is annotated with a 13-variable metadata code encoding information such as institutional affiliation, course type, year, CEFR proficiency level (A1–B2), task type, and learner-specific variables including nationality, native language, gender, age, and instructional format (online or in-person). The corpus includes learners from 49 nationalities and 37 L1 backgrounds, with the most common being Arabic, Vietnamese, French, and various English varieties. Learner ages range from 19 to 77, with the majority being young adults. All data is anonymised, with learners identified by numerical codes (Baumann et al., 2020). The MagyarOK Open Corpus complements this dataset by providing texts from the four level coursebook. The used texts in the dataset were selected based on statistical keyness at each proficiency level. By bringing together the MID Learner Corpus and the MagyarOK Open Corpus, our aim was to create a dataset that blends real learner language with carefully designed teaching materials, creating a strong foundation for classifying texts by CEFR level.

2.1 Dataset Composition and Data Augmentation Strategies

The classification task focuses on four CEFR proficiency levels: A1, A2, B1, and B2. Each level is represented by 2,000 text samples, yielding a balanced dataset of 8,000 entries. The MID Learner Corpus was imbalanced, with the majority of texts concentrated at mostly the A1 and A2 levels. We implemented targeted data augmentation strategies to address this and ensure more representative training conditions. These augmentation methods were carefully designed to reflect the developmental language patterns typical of each CEFR level:

- **A1 and A2:** Limited data augmentation was applied, which involved *random word swaps* and *modifications to word order*.
- **B1 and B2:** Techniques included *synonym replacement* and *AI-assisted paraphrasing*, both of which were subsequently evaluated for their effectiveness and linguistic plausibility.

2.2 Model Selection

The main model used in this study is Puli-BERT-Large (Yang et al., 2023), a transformer-based language model specifically designed for Hungarian. Developed by the HUN-REN Hungarian Research Centre for Linguistics and available on Hugging Face. Puli-BERT-Large is particularly well-suited for tasks involving the Hungarian language. Unlike general multilingual models, it captures the unique linguistic and cultural features of Hungarian, offering a much more accurate representation. Its training on large-scale Hungarian text allows it to pick up on important morphological, syntactic, and lexical details (Yang et al., 2023), which are key elements for assessing language proficiency, especially in texts written by learners. This highlights the importance of developing and implementing language-specific models, as evidenced by studies on other morphologically rich languages such as Basque (Arriola et al., 2023) and Estonian (Vajjala & Lõo, 2014).

A further justification for employing BERT is that recent studies have shown that BERT-based models outperform traditional, feature-engineered methods in predicting CEFR levels. Research by Schmalz and Brutti (2021) and for example, found that BERT embeddings provide significantly better accuracy in CEFR classification than models relying on hand-crafted linguistic features, partic- ularly when large datasets are used. These findings are echoed in more recent work by Volodina et al. (2024), who reported strong performance from BERT-based models across several languages, even in low-resource scenarios, making them a good fit for Hungarian as well. Likewise, Uhrström (2023) found that transformer models not only offer strong predictive performance but are also more resilient in contexts where data is limited or language variation is high, common challenges in learner corpora.

2.3 Fine-Tuning and Evaluation Procedure

The dataset was divided into training and test sets using an 80/20 split. Following the standard fine-tuning pipeline for BERT-based models (Devlin et al., 2019), the model was trained for three epochs on the training set. Model performance was evaluated using classification accuracy and confusion matrix analysis, with particular attention paid to common misclassification patterns between adjacent CEFR levels (e.g., A2 vs. B1).

3 Results

3.1 Classification Performance

The fine-tuned Puli-BERT-Large model achieved an overall accuracy of 81.5% on the test set, with a cross-entropy loss of 0.72. A closer inspection of the confusion matrix (Figure 1) reveals distinct patterns of misclassification, largely concentrated around transitional boundaries between CEFR levels. The model performs quite reliably at the two ends of the CEFR scale (A1 and B2) with relatively few classification errors in these categories. However, the middle levels, A2 and B1, are more challenging, and the model makes more mistakes there.

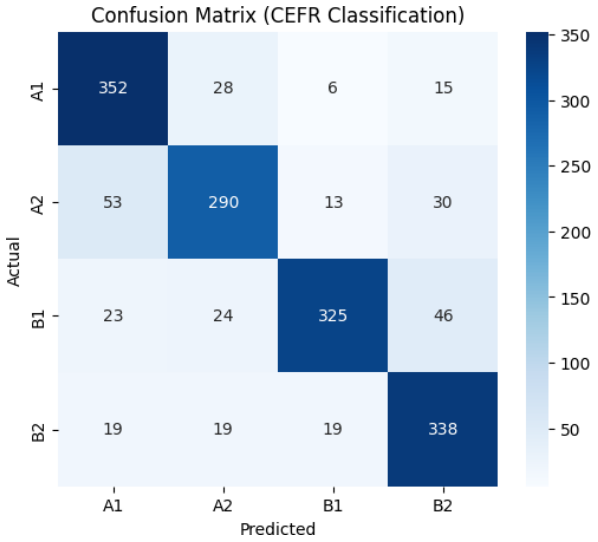

Figure 1: Confusion Matrix of CEFR Classification Task

3.2 Linguistic Analysis of Model Behavior

To explore the underlying factors influencing the model's decisions, the study includes a qualitative analysis of both correctly and incorrectly classified texts. These findings are also presented in detail as part of the research.

Example of correct classification (true: B2, predicted: B2):

(1) Nehezen szoktam meg, hogy a történelem órán mást tanítanak.
difficultly get.used-PST-1SG PRTCL that the history class-LOC other-ACC teach-3PL
"I had a hard time getting used to the fact that they teach something different in history class"

This B2-level sentence reflects advanced language skills through its use of subordinate clauses, temporal and cognitive verbs, and cohesive devices that signal contrast or elaboration. These features are typical of upper-intermediate proficiency and likely align with how the model interprets this level.

Example of incorrect classification (true: B2, predicted: A2):

(2) Az országomban sokan álmodnak arról, hogy külföldre költözzenek, de sokan közülük nem tudják elhagyni az országot, mert nincs pénzük vagy útlevelük.
the country-POSS.1SG-INE many-PL dream-3PL about.that that abroad-ILL move-SBJV.3PL but many-PL among-3PL not can-3PL PRFX-leave-INF the country-ACC because not.have.3SG money-3PL.POSS or passport-3PL.POSS
"In my country, many people dream of moving abroad, but many of them cannot leave the country because they don't have money or a passport"

Even though the sentence had a clear structure, used complex modal expressions, and included embedded subordinate clauses, the model still underestimated the writer's proficiency. This suggests that the model tends to overlook subtle meanings and real-world context, focusing too much on surface-level syntax.

Example of incorrect classification (true: A2, predicted: B1):

(3) Dr. Antonio Agostinho Neto az elso″ angolai elnök volt. Orvos és költo″ is volt, és fontos szerepet játszott az ország függetlenségi harcában.

Dr. Antonio Agostinho Neto the first Angolan president was doctor and poet also was and important role-ACC play-PST.3SG the country independence struggle-POSS-INE
"Dr. Antonio Agostinho Neto was the first president of Angola. He was also a doctor and a poet, and he played an important role in the country's struggle for independence."

This excerpt uses rich vocabulary and tells a clear, coherent story, but its structure is fairly simple. The misclassification may be due to the model mistakenly linking the presence of named entities and biographical content with a higher proficiency level.

4 Conclusion

The current study shows that Puli-BERT-Large can reliably classify Hungarian learner texts according to CEFR proficiency levels. More than just assigning labels correctly, the model reveals clear and interpretable patterns in its decision-making process. By leveraging deep learning on a varied set of learner texts, the project contributes both practical tools for assessing learner language and insights into how language models perceive and process it. This analysis looks at more than just accuracy. It also considers how well the model's classifications match or differ from the teaching expectations based on the CEFR.

References

Arriola, J. M., Alkorta, J., Arrieta, E., & Iruskieta, M. (2023). Towards automatic essay scoring of basque language texts from a rule-based approach based on curriculum-aware systems. *Proceedings of the NoDaLiDa 2023 Workshop on Constraint Grammar - Methods, Tools and Applications*, 20–28. https://aclanthology.org/2023.nodalida-cgmta.4

Baumann, T., Majoros, J., Pelcz, K., Schmidt, I., Szita, S., & Vermeki, B. (2020). Bemutatkozik a korpusznyelvészeti és szakmódszertani munkacsoport. *Hungarológiai Évkönyv*, *21*(1-2), 32–41. https://epa.oszk.hu/02200/02287/00021/pdf/EPA02287_hungarologiai_evkonyv_2020_01_032-041.pdf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Schmalz, V. J., & Brutti, A. (2021). Automatic assessment of english cefr levels using bert embeddings. *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2021)*. https://ceur-ws.org/Vol-3033/paper14.pdf

Szita, S. (2021). A magyarok nyílt korpusz használatáról. *Hungarológiai Évkönyv*, *22*(1-2), 72–88. https://epa.oszk.hu/02200/02287/00022/pdf/EPA02287_hungarologiai_evkonyv_2021_01_072-088.pdf

Uhrström, E. (2023). *Exploring cefr classification using transformer-based models* [Master's thesis, University of Gothenburg]. https://gupea.ub.gu.se/handle/2077/77332

Vajjala, S., & Lõo, K. (2014). Automatic cefr level prediction for estonian learner text. *Proceedings of the Third Workshop on NLP for Computer-Assisted Language Learning*, *22*, 113–127. https://aclanthology.org/W14-3509

Volodina, E., Pilán, I., Alfter, D., & Lundkvist, P. (2024). Automatic cefr classification for learner writing in multiple languages: Combining generic and language-specific components. *Proceedings of the 13th Workshop on NLP for Computer-Assisted Language Learning (NLP4CALL 2024)*, 92–101. https://aclanthology.org/2024.nlp4call-1.11

Yang, Z. G., Dodé, R., Ferenczi, G., Héja, E., Jelencsik-Mátyus, K., Kőrös, Á., Laki, L. J., Ligeti-Nagy, N., Vadász, N., & Váradi, T. (2023). Jönnek a nagyok! bert-large, gpt-2 és gpt-3 nyelvmodellek magyar nyelvre. *XIX. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2023)*, 247–262. https://acta.bibl.u-szeged.hu/78417/1/msznykonf_019_247-262..pdf