Issue 3/2025
Str. 162–176

**Matej Klemen, Martin Božič,
Špela Arhar Holdt and Marko Robnik-Šikonja**

Grammatical error correction of Slovenian school essays using large language models

Abstract: Grammatical error correction (GEC) is the task of automatically detecting and correcting grammatical errors in text. Large language models have enabled the development of accurate automated methods for detecting and correcting certain types of errors. In the educational domain, the aim of GEC is to aid teachers in correcting student errors. Excessive paraphrasing is a property of Generative Pre-trained Transformer-based models and is undesirable in the language education context. To avoid this, we develop multiple Slovenian models for correcting errors in spelling, word case (capitalization), word form, and word order. We describe the training data construction, training process, and model evaluation approach using the Šolar-Eval 1.0 corpus of school essays authored by primary and secondary school students. Our quantitative evaluation shows that the developed models have reasonably high accuracy levels, and our qualitative evaluation highlights the strengths and weaknesses of the models and the evaluation process. The analysis reveals multiple challenges and promising future directions for improving both model development and the evaluation process.

Keywords: large language models, grammatical error correction, educational domain, synthetic data construction

UDC: 371

https://doi.org/10.63384/sptB53z793a

Scientific article

Matej Klemen,Tch. Asst., University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia; e-mail: matej.klemen@fri.uni-lj.si;

Martin Božič, University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia; e-mail: martin.bozic@fri.uni-lj.si

Špela Arhar Holdt, PhD, Res. Assoc., University of Ljubljana, Faculty of Arts, Aškerčeva cesta 2, SI-1000 Ljubljana, Slovenia; e-mail: spela.arharholdt@ff.uni-lj.si;

Marko Robnik-Šikonja, PhD, full professor, University of Ljubljana, Faculty of Computer and Information Science, Večna pot 113, SI-1000 Ljubljana, Slovenia; e-mail: marko.robniksikonja@fri.uni-lj.si;

Introduction

Grammatical error correction (GEC) refers to the automatic detection and correction of grammatical errors in written text and is a key area of natural language processing (NLP). Grammar checkers – programs that verify grammatical correctness, flag potential language issues, and suggest corrections – are among the most fundamental language technologies and provide important digital infrastructure for modern languages, including Slovene (Krek 2023). Traditional grammar checkers relied on predefined rule sets, whereas modern systems leverage large language models (LLMs), particularly transformer-based architectures (Vaswani et al. 2017). These AI-driven models have the ability to consider the broad syntactic and semantic contexts of written texts and have thus become essential for advanced language technologies, significantly outperforming previous methods. Automated writing evaluation (AWE) is a key application of GEC; it relies on accurate error detection and correction to provide effective feedback on students' writing.

AWE systems provide significant benefits for both students and teachers, as they increase students' motivation to write and enhance their self-efficacy (Warschauer and Grimes 2008; Wilson and Czik 2016; Wilson and Roscoe 2020; Zupanc and Bosnić 2017). These tools also reduce the time required for providing feedback significantly, thereby allowing teachers to focus on the higher-level aspects of writing, such as content and organization (Cotos 2014; Connor et al. 2014; Warschauer and Grimes 2008). However, many issues need to be considered, such as the tool's technical capabilities, its alignment with pedagogical goals, and the precision of feedback it provides, which varies depending on the type of error (Ranalli 2016, 2018; Woodworth and Barkaoui 2020; Wilson et al. 2024). Without a solid GEC backbone and careful implementation, AWE tools risk providing inconsistent or even misleading feedback.

Given the role of GEC in language technologies and its impact on automated writing evaluation, we developed novel Slovenian GEC models for various levels of error identification and correction. Specifically, these models can correct spelling, morphology, and selected orthography (capitalization of letters) and syntax (word order) issues. Figure 1 presents an example of a model and its outputs.

We decided to build multiple models for different error types (instead of one general model) because of the challenging nature of constructing the training dataset. To recognize human-made grammatical errors, the models had to be trained on a dataset of general texts containing a realistic amount of errors as well as realistic types of errors. Such datasets are not publicly available for Slovene, so we resorted to constructing synthetic training datasets. While we could successfully synthesize various error types independently, capturing a realistic frequency of different error types within the same text proved to be a more challenging problem, which we decided not to tackle in this study. Of the constructed models, the spelling and word case models exhibited the most promising results; the other two require further work before they can be applied.

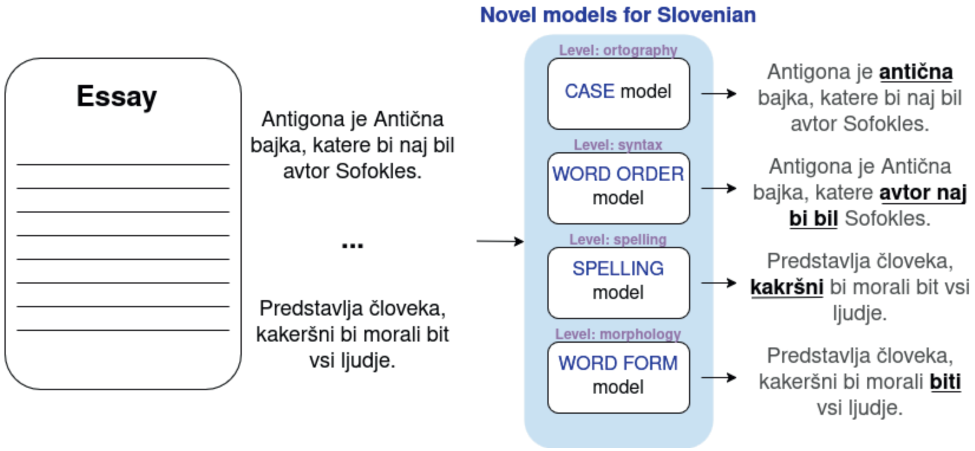
Novel models for Slovenian
Essay
Antigona je Antična bajka, katere bi naj bil avtor Sofokles.
...
Predstavlja človeka, kakeršni bi morali bit vsi ljudje.
Level: ortography
CASE model
Level: syntax
WORD ORDER model
Level: spelling
SPELLING model
Level: morphology
WORD FORM model
Antigona je antična bajka, katere bi naj bil avtor Sofokles.
Antigona je Antična bajka, katere avtor naj bi bil Sofokles.
Predstavlja človeka, kakršni bi morali bit vsi ljudje.
Predstavlja človeka, kakeršni bi morali biti vsi ljudje.

Figure 1: High-level overview of the problem and our contributions. We introduced multiple models for grammatical error correction in Slovenian.

This paper presents the development process, the effectiveness of the models in addressing common linguistic errors, particularly in educational contexts, and the current applicability of the models to writing support systems.

Related Work on Grammatical Error Correction

GEC systems are commonly developed for high-resource languages such as English and less so for smaller languages such as Slovenian. English systems such as Grammarly have reached mass adoption due to their ability to detect errors and suggest corrections with high accuracy. For Slovenian, the with the broadest error correction coverage is Amebis Besana, a commercial product based on hand-crafted linguistic rules.

Some recent works have involved the development of neural network systems for specialized problems. Božič (2020) developed an accurate Bidirectional Encoder Representations from Transformers-based (BERT) system for automatic comma placement and released the model for public use through an online interface. Klemen et al. (2024) developed an accurate BERT-based spelling error detection model using synthetically generated training data. Spelling errors were synthesized using various heuristics based on intuition, such as the insertion of random characters and commonly misspelled words.

It has become increasingly common to apply large language models such as ChatGPT to GEC tasks, with the models instructed to correct the input text based on a prompt (Lin 2024). However, such systems are prone to significant input modifications, making them unsuitable in educational contexts. Considering this, we developed and evaluated multiple novel models for the detection and correction of spelling, word form, word order, and word case errors. We analysed their predictions in an educational context using the Šolar-Eval 1.0 dataset (Arhar Holdt et al. 2023) and thus obtained insights into their strengths and weaknesses as well as the challenges of applying such systems in education.

Grammatical Error Correction Using Language Models

In this section, we describe the models developed for GEC. We discuss the underlying language models used as the base and then describe our customizations, such as the input and output formats, and the synthetic training data construction approach.

Base Language Models

We built our detection and correction models by fine-tuning pretrained transformer models BERT (Bidirectional Encoder Representations from Transformers) and T5 (Text-to-Text Transfer Transformer) using synthetically constructed training data. The base models were pretrained on large text corpora, with the aim of capturing general language properties. Additional fine-tuning allowed the models to learn specific downstream tasks.

BERT (Devlin et al. 2019) is an encoder-only language model pretrained using masked language modelling and next sentence prediction tasks. Its architecture makes it particularly effective for discriminative tasks, such as the detection of grammatical errors. However, it is less suitable for text generation tasks, so we did not use it for our correction models. We specifically used the BERT-like model SloBERTa (Ulčar and Robnik-Šikonja 2021) in our work, pretrained on multiple Slovenian corpora.

T5 (Raffel et al. 2020) is an encoder-decoder language model pretrained using independent, identically distributed denoising and span corruption tasks. Its architecture enables the mapping of variable-length input sequences to variable-length output sequences, which makes it useful for generative tasks – in our case, the correction of grammatical errors. We used the T5-sl-small (Ulčar and Robnik-Šikonja 2023) model pretrained on multiple Slovenian corpora in our work.

Model Customization for Slovenian

We fine-tuned the base language models using data synthetically generated from Gigafida 2.0 (Krek et al. 2020) corpus. In this section, we describe the training modifications we made, including input and output modifications, and the strategies used to synthetically generate data. We fine-tuned all models for two epochs and used a learning rate of $2 \cdot 10^{-5}$. Table 1 presents an overview of the developed models.

Model	**Problem**	**Type**
T5-slo-word-spelling-annotator	Spelling	Detection
T5-slo-word-spelling-corrector	Spelling	Correction
T5-slo-word-form-corrector	Word form	Correction
T5-slo-word-order-corrector	Word order	Correction
SloBERTa-word-case-classification-multilabel	Word case	Detection*

*Table 1: Overview of the newly developed models. The model type reflects the main goal of the model. The correction models could also flag suggested corrections as errors without modifying the words and could thus be used for error detection. * = the model is specialized for detection, but the predicted classes also enable error correction. We have released the models publicly on the CJVT HuggingFace repository.*

To build *spelling error detection* and *spelling error correction* models, we applied the strategies described in our previous work (Klemen et al. 2024) to obtain a training dataset and then used it with a generative model (T5) instead of a discriminative one (BERT). The aim was to observe the accuracy of a system that treats spelling detection and correction as a text generation problem rather than as a classification problem. The strategies included random word separation, neighbouring word concatenation, replacement of commonly misspelled word and character pairs, and random insertion or deletion of characters. To transform the spelling detection problem into a generative task, a special text label was added after each incorrectly spelled word, whereas the correctly spelled words were left unchanged. The task for the spelling correction model was to map an incorrectly spelled input text to a correctly spelled one, so no task transformation was needed.

To build a *word form correction* model, we fine-tuned a T5 model using a synthetic training set with partially replaced word forms. Each word was a word form replacement candidate with probability $p_{wordform} = 0.2$. Each candidate was then replaced with a random inflected form with the same lemma. This was done using the Sloleks 2.0 lexicon, which contains all word forms for more than 100,000 lemmas.

To build a *word order correction* model, we fine-tuned a T5 model using a synthetic training set with locally shuffled word orders. An input sentence was segmented into parts based on conjunctions and commas. Then, $N_{reorder}$ words are randomly shuffled in each part, with the highest weight assigned to $N_{reorder} = 1$ (indicating no shuffle), and the weight linearly decreased to the number of words in each segment. This allowed the model to learn to correct shorter word shuffles as well as larger sentence part rearrangements.

To build a *word case correction model*, we trained a BERT model to determine the case of each word. The model received lowercased text as the input, and its goal was to determine whether the word was uppercase, lowercase, or all-uppercase. To obtain the ground truth labels for our training set, we assumed that the initial cases in the Gigafida 2.0 corpus were placed correctly. We thus limited our training data to a subset of documents produced by publishers with language editors.

Evaluation

In this section, we present our analysis of the proposed models' performance. We start by describing the evaluation procedure, including the data and quality metrics used. We then discuss our quantitative and qualitative analyses of the models.

Evaluation Procedure

To evaluate the models in an educational setting, we used the Šolar-Eval dataset (Arhar Holdt et al. 2023), which contains 109 essays authored by Slovene primary and secondary school students. The dataset was sampled from the Šolar 3.0 corpus and reannotated for optimized correction consistency, homogeneity, and minimal language intervention. It was annotated using a three-level error categorization scheme, in accordance with publicly released online guidelines[1]. For model evaluation, we extracted data subsets containing specific types of errors. For example, to evaluate the word order correction model, we only considered errors of the type 'S/BR', which indicated a syntax-level error.

We quantitatively evaluated the models using general machine learning metrics, namely precision, recall, and the F0.5 score, an aggregate metric that

[1] https://wiki.cjvt.si/attachments/52

combines precision and recall of a system, assigning a higher weight to precision than to recall. This choice was supported by intuition, considering that many false positives can be frustrating in a practical setting, as well as the literature on measuring the correlation between automated metrics and human judgement (Grundkiewicz et al. 2015). The F0.5 score is defined as follows:

$$F_{0.5} = \frac{1.25 \cdot P \cdot R}{0.25 \cdot P + R} = \frac{1.25 \cdot TP}{1.25 \cdot TP + 0.25 \cdot FN + FP}$$

Here, P is the system precision, R is the system recall, TP is the true positive count, FN is the false negative count, and FP is the false positive count.

For our qualitative analysis, we manually observed a random sample of true positive, false positive, and false negative predictions by the models and discussed the commonly occurring patterns. We focused on evaluating the strengths and weaknesses of the newly developed models in the educational domain rather than a detailed comparison with existing (and scarcely available) systems. Previous work (Klemen et al. 2024) has shown that untuned commercial large language models such as ChatGPT[2] perform poorly at the problems we tackled in Slovenian, which further motivated their exclusion as baseline models. We plan to perform a thorough baseline comparison in our future work, with the release of the next iteration of models built using our method.

Quantitative Evaluation

Model	**Problem**	**Precision**	**Recall**	**F0.5**
T5-slo-word-spelling-annotator	Spelling	0.587	0.716	0.609
T5-slo-word-spelling-corrector	Spelling	0.663	0.847	0.693
T5-slo-word-form-corrector	Word form	0.581	0.258	0.465
T5-slo-word-order-corrector	Word order	0.571	0.148	0.364
SloBERTa-word-case-classification-multilabel	Word case	0.640	0.898	0.679

Table 2: Precision, recall, and F0.5 scores of the models for an error identification task.

Table 2 presents the precision, recall, and F0.5 scores achieved by the newly developed models for an error identification task. We observed whether an annotated error was identified correctly but not necessarily corrected accurately. The spelling detection, spelling correction, and word case correction models achieved moderately high F0.5 scores (between 0.609 and 0.679), and their recall scores were higher than their precision scores, indicating that a large portion of annotated errors were detected but a portion of them were falsely flagged. We observe that

[2] https://chatgpt.com/

the T5 model trained to detect and correct errors (T5-slo-word-spelling-corrector) achieved higher metric scores on the detection task than the model only trained to detect errors (T5-slo-word-spelling-annotator). This indicates the benefit of multi-task learning: the model that learned to also correct errors was better at detecting them. In contrast, the word form and word order correction models obtained poor F0.5 scores of 0.465 and 0.364, respectively. These models achieved higher precision than recall, which indicates that only a few annotated errors were detected by the models, but these were relatively accurate.

The automatic metric scores were particularly low for the word order correction model. This was partly due to the model missing many errors. However, the estimate obtained was also overly pessimistic because of the relatively flexible word orders of the Slovenian language. In certain cases, multiple different word orders were valid, but only one was annotated as correct in the evaluation dataset.

Model	**Problem**	**TP Correction Accuracy**
T5-slo-word-spelling-corrector	Spelling	0.903
T5-slo-word-form-corrector	Word form	0.821
T5-slo-word-order-corrector	Word order	0.250
SloBERTa-word-case-classification-multilabel	Word case	0.946

Table 3: Accuracy of corrections made on the subset of true positive identification model predictions.

Table 3 presents the accuracy levels of the corrections generated for true positives in the detection task. This shows how many of the correctly identified errors were also accurately corrected. We found that the correction models achieved very high accuracy levels in most cases, between 0.821 and 0.946. The word order correction model was an exception; it achieved an accuracy of 0.250. This was partly due to suboptimal correction performance and partly a consequence of the relatively free word order in Slovenian and the existence of multiple valid orders. To obtain further insights into the model predictions, we analysed them qualitatively. We present these conclusions next.

General Qualitative Performance Observations

We found that the correction models built on top of the generative T5 models were prone to leaving out text, particularly dependent clauses. This is undesirable, as such negations may change the meaning of the text or remove essential information. In terms of practical usability, this is as undesirable as overcorrection and presents an obstacle for the adoption of fully automatic grammatical error correction models. In certain cases, the limited context provided to the models gave rise to potential ambiguity regarding the correct outcome. Either because of the evaluation setting or the model's limited context window, there were multiple valid outcomes, despite only one being annotated as correct in the dataset.

We also found that the models faced issues when dealing with proper nouns. This was likely due to the inclusion of literary works in the dataset; the names found in literary works tend to be uncommon in general use. We plan to tackle this issue by injecting additional named entity knowledge into the models in the future. In the next section, we focus on our qualitative analysis of the word case model's predictions, as it performed well in the quantitative analysis.

Linguistic Qualitative Evaluation of the Word Case Model

For our detailed qualitative evaluation, we chose the word case model because it was one of the most promising models and the example cases are most intuitive for explanations in English. The evaluation involved 386 sentences from the Šolar-Eval 1.0 corpus, each containing isolated word case corrections such as *Ko smo končali trgatev, smo se odšli kopat v vipavo > Vipavo.* ('When we finished the grape harvest, we went swimming in the vipava > Vipava.) The exported data included the original sentence, the ID of the source essay from Šolar-Eval 1.0, the proposed computational correction, the human correction, and an indication of whether the machine and human decisions aligned. We manually analysed all the data and grouped the false positives (FPs) and false negatives (FNs) into bottom-up categories based on the type of the problem. The results are presented in Table 4.

Category and Problem Type	Number of Issues	Percentage of Issues
True positives	229	59.33%
False positives	131	33.94%
Proper vs. common noun ambiguity	67	17.36%
Sentence boundary misidentification	35	9.07%
Uninterpretable errors and unknown causes	15	3.89%
Residual capitalization issues	12	3.11%
Unnecessary human correction	2	0.52%
False negatives	26	6.74%
Proper vs. common noun ambiguity	15	3.89%
Sentence boundary misidentification	6	1.55%
Unnecessary human correction	3	0.78%
Uninterpretable errors and unknown causes	2	0.52%
Total	386	100.00%

Table 4: The qualitative evaluation results of the word case model include the number and percentage of true positives, false positives, and false negatives, along with manually assigned problem types.

Šolar-Eval 1.0 predominantly consists of descriptive or analytical essays on literary works that students have read. A significant portion of the identified issues with the word case model could be attributed to proper vs. common noun ambiguity, as common nouns are frequently used as proper nouns in the literary works. For example, *Uvod, Krst, Prilika* ('Introduction, Baptism, Parable') are used as section titles and thus require capitalization. Similarly, character names and titles that typically function as common nouns appear as proper nouns, such as *Doktor, Paž, Baron* ('Doctor, Page, Baron'). In some cases, capitalization depends on the meaning conveyed, such as *(krščanski) Bog* vs. *bog* ('Christian God' vs. 'a god'). Challenges also arose with geographic names, as capitalization depends on whether a place is inhabited, such as *Vrh* vs. *vrh* ('Peak' vs. 'peak'). These led to difficulties for the students as well as the system, and they led to both false negatives (15 out of 26) and false positives (67 out of 131). The examples below illustrate these challenges. The first example shows how the system erroneously corrected the proper name 'Baron' to a common noun. The second shows how it failed to recognize 'Introduction' as a book section title that should be capitalized.

– Na gradu sta prebivala Matiček kot vrtnar, ki ga Baron > **b**aron poviša v prvega služabnika, in Nežka kot hišna. ('In the castle lived Matiček as the gardener, whom the Baron > **b**aron promoted to first servant, and Nežka as the housemaid.')
– Za uvod > **u**vod h Krstu pri Savici bi lahko rekli, da je ep v malem. ('For the Introduction > **i**ntroduction to Baptism at the Savica, one could say that it is an epic in miniature.')

The second major group of cases involved *sentence boundary misidentification errors* – that is, errors in recognizing the beginning of a new sentence or segment. This group accounted for 35 false positives and 6 false negatives. A significant number of these cases stemmed from inconsistencies in corpus structure, specifically because the boundaries between titles, subtitles, and bullet points were not clearly marked. Direct speech or quoted text within a sentence was another common source of errors, as the model struggled to determine whether capitalization was required. Additionally, some cases were influenced by ellipses; the capitalization of the word following an ellipsis depends on how the sentence break is interpreted. The first example below illustrates issues arising from corpus text segmentation, where titles were not properly separated from the start of the essay. The second example shows how the system failed to correct the capitalization of the second quote, and the third how its interpretation of the ellipsis was different from the human correction. It is important to note that cases like these are often ambiguous, and even human corrections can vary.

– Šolska naloga Pričakovanja > **p**ričakovanja in resničnost Človek bi pričakoval, da bi ljudje, vsi enaki, držali skupaj. ('School essay Expectations > **e**xpectations and reality One might expect that people, all equal, would stand together.')
– Ne stori komu tega, česar ne želiš, da bi on tebi in kdor > **k**dor je brez greha, naj prvi vrže kamen vanj. ('Do not do unto others what you do not want

them to do unto you' and 'whoever > **w**hoever is without sin, let them cast the first stone.')

– Usoda?? Pišemo jo sami, čeprav … So > **S**o stvari, dogodki, ki jih nikoli ne bomo mogli doumeti, in prav je tako. ('Fate?? We write it ourselves, even though … There > **T**here are events we will never be able to comprehend, and that is as it should be.')

In some cases (2 false positives and 3 false negatives), the suggested language correction could have been handled differently or was unnecessary, which means the issue was not an actual program error. The example below illustrates an *unnecessary human correction,* which involved the model interpreting 'story' as a common noun – a valid decision in this case.

– V zgodbi > **z**godbi o Kajnu in Abelu bi bilo lahko pravično tudi, če bi bog sodil umor z umorom in Kajna. ('In the story > **s**tory of Cain and Abel, it could also have been fair if God had judged murder with murder and Cain.')

Among the false positives were several notable *residual capitalization issues,* including unrecognized proper names, misinterpretation of common nouns as proper nouns, and errors influenced by a foreign language or borrowed vocabulary features (a total of 12 cases). The first example below shows the model's application of the general rule that adjectives ending with -ski (ajdovski) are written with a lowercase letter, failing to recognize that in this case (Ajdovski gradec), it is part of a geographical proper name. The second example demonstrates a misinterpretation of a common noun as a proper noun, while the third involves the correction of a German borrowing, which remains capitalized in the literary text referenced.

– Črtomir se zelo spremeni na prehodu iz Uvoda v Krst, ker po porazu pri Ajdovskemu > **a**jdovskemu gradu spozna, da se je utopično boril za versko samostojnost slovencev. ('Črtomir undergoes significant change in the transition from Introduction to Baptism, as after the defeat at Ajdovski > **a**jdovski Castle, he realizes that his struggle for religious independence was futile.')
– Po vrnitvi z bega > **B**ega so me starši zelo nadrli, še leto po tem sem imela vse prepovedano, npr. televizor, računalnik ('After returning from the Run > **r**un, my parents scolded me severely; even a year later, I was forbidden from watching TV, using the computer, etc.')
– Ko so prišli do železnice, kjer naj bi potekal pregled za odločitev za Volksdojčerja > **v**olksdojčerja ali za pravega arijskega nemca. ('When they arrived at the railway station, where an inspection was to be conducted to decide between becoming a Volksdeutscher > **v**olksdeutscher or a true Aryan German.')

The most challenging issues affecting further development and implementation are those that cannot be clearly interpreted, making it impossible to determine why the program made an error. *Uninterpretable errors and unknown causes* accounted for 15 false positives and 2 false negatives in our analysis. The first example below shows that the model incorrectly adjusted the capitalization after a full stop, with no clear justification for the decision. A similar issue is present in the second example; the model incorrectly interpreted the name ‘Nežka’ as a common noun, even though its interpretations were correct throughout the rest of the text.

- Baron sumi, da ima baronica v sobi skritega tujega moškega, saj so bila vrata njene sobe zaklenjena. Dreza > **d**reza vanjo, dokler mu ta ne prizna, da je v sobi Tonček. (‘The baron suspects that the baroness is hiding another man in her room, as the door was locked. He prods > **h**e prods her until she admits that Tonček is inside.’)
- Oslepljeni baron dvori svoji ženi v prepričanju, da je Nežka. Na koncu spozna v svoji ”Nežki > **n**ežki” ženo, ki mu vse oprosti – baron tretjič naleti. (‘The blind baron courts his wife, believing her to be Nežka. In the end, he recognizes his “Nežka > **n**ežka” as his wife, who forgives him – this is the third time the baron has been fooled.’)

Despite the identified challenges, the model successfully handled several cases across different categories and correctly applied capitalization rules where needed: 59.33% of the outcomes were *true positives*. The first example below shows the model’s correct capitalization of a regional name upon distinguishing it from a general geographical reference. The second example shows that it recognized the beginning of a new sentence and applied capitalization correctly. The third example demonstrates the model’s ability to recognize demonyms and apply capitalization accordingly.

- Veliko sem tudi prejokala, saj sem se spominjala najinih potovanj po primorskem > **P**rimorskem. (‘I also cried a lot, as I remembered our travels through primorska > **P**rimorska.’).
- Takrat mu Laert pove za zaroto in Hamlet ubije še Klavdija. tako > **T**ako vsi umrejo. (‘At that moment, Laertes reveals the conspiracy, and Hamlet kills Claudius as well. so > *So* they all die.’)
- Njihov največji problem je bil ta, da so bili nemci > **N**emci. (‘Their biggest problem was that they were germans > **G**ermans.’)

Conclusion

In our work, we developed multiple novel models for Slovenian GEC and evaluated them quantitatively and qualitatively in an educational context. The spelling and word case models achieved moderately high automated metric scores, whereas the word order and word form models performed poorly.

The qualitative evaluation of the word case model with Šolar-Eval 1.0 highlighted several key challenges related to capitalization. The most prevalent issues stemmed from proper vs. common noun ambiguity, sentence boundary misidentification, residual capitalization issues, and uninterpretable errors. Our evaluation showed that although there were high numbers of false negatives and false positives, many of these stemmed from the challenges presented by literary works and the complexity of capitalization rules in Slovenian. Evaluations using other datasets would likely yield more promising results. Efforts to improve this model could focus on refining its handling of quotation marks, ellipses, and sentence segmentation and on using different pretrained base models. Additionally, future work could involve the development of context-aware disambiguation strategies, particularly for words that can function as both proper and common nouns. Evaluations using a more diverse dataset that goes beyond literary analysis essays would also provide valuable insights into the model's applicability and potential refinements.

The unpredictability and variability of certain decisions affect the consistency and reliability of error correction and thus pose the most significant challenges to the application of GEC models based on new approaches. While modern transformer-based models have substantially improved grammatical error detection and correction by leveraging broad syntactic and semantic contexts, their decision-making processes remain opaque. The resulting unpredictability is particularly problematic in the context of automated writing evaluation, as inconsistent or unclear corrections may mislead users and compromise the system's pedagogical effectiveness. As GEC models become more widely implemented in educational and professional settings, addressing these inconsistencies is essential for ensuring their accuracy, transparency, and practical applicability.

The work was supported by the Slovenian Research and Innovation Agency (ARIS) via the projects J7-3159, GC-0002, and L2-50070, the core research programme P6-0411, and the young researchers grant. The work was also supported by EU through ERA Chair – grant no. 101186647 (AI4DH) – and the cofinancing of research innovation projects in support of green transition and digitalization (project PoVeJMo, no. C3.K8.IB). The computational resources were provided by SLING through project S24O01-42.

References

Arhar Holdt, Š., Gantar, P., Bon, M., Gapsa, M., Lavrič, P. and Klemen, M. (2023). *Dataset for evaluation of Slovene spell- and grammar-checking tools* Šolar-*Eval 1.0. Slovenian language resource repository CLARIN.SI*. Retrieved from: http://hdl.handle.net/11356/1902 (accessed on 5. 6. 2025)

Arhar Holdt, Š. and Kosem, I. (2024). Šolar, the developmental corpus of Slovene. *Language Resources and Evaluation*, pp. 1–27.

Božič, M. (2020). *Globoke nevronske mreže za postavljanje vejic v slovenskem jeziku* (Bachelor thesis). Ljubljana: Univerza v Ljubljani, Faculty of Computer and Information Science.

Connor, C. M., Goldman, S. R. and Fishman, B. (2014). Technologies that support students' literacy development. In: *Handbook of research on educational communications and technology.* New York: Springer, pp. 591–604.

Cotos, E. (2014). Automated writing evaluation. In: *Genre-Based automated writing evaluation for L2 research writing*. London: Palgrave Macmillan, pp. 43–79.

Devlin, J., Chang M., Lee, K. and Toutanova K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1 (Long and Short Papers), pp. 4171–4186.

Gantar, P., Bon, M., Gapsa, M. and Arhar Holdt, Š. (2023). Šolar-Eval: Evalvacijska množica za strojno popravljanje jezikovnih napak v slovenskih besedilih. *Jezik in slovstvo*, 68, issue 4, pp. 89–108.

Klemen, M., Božič, M., Arhar Holdt, Š. and Robnik-Šikonja, M. (2024). Neural spell-checker: beyond words with synthetic data generation. In: *Text, speech, and dialogue: 27th International Conference,* TSD 2024, Brno, Czech Republic, September 9–13, 2024, Proceedings, Part I, pp. 85–96.

Krek, S. (2023). Language report Slovenian. In: *European language equality: A Strategic agenda for digital language equality*. Cham: Springer International Publishing, pp. 211–214.

Raffel, C., Shazeer N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research,* 21, issue 140, pp. 1–67.

Ranalli, J., Link, S. and Chukharev-Hudilainen, E. (2016). Automated writing evaluation for formative assessment of second language writing: Investigating the accuracy and usefulness of feedback as part of argument-based validation. *Educational Psychology,* 37, issue 1, pp. 8–25.

Ranalli, J. (2018). Automated written corrective feedback: how well can students make use of it? *Computer Assisted Language Learning,* 31, issue 7, pp. 653–674.

Ulčar, M. and Robnik-Šikonja, M. (2021). SloBERTa: Slovene monolingual large pretrained masked language model. In: Data Mining and Data Warehouses - SiKDD, Information Society - IS 2021: proceedings of the 24th international multiconference, volume C, pp. 17–20.

Ulčar, M. and Robnik-Šikonja, M. (2023). Sequence to sequence pretraining for a less-resourced Slovenian language. *ArXiv cs.CL* 2207.13988.

Warschauer, M. and Grimes, D. (2008). Automated writing assessment in the classroom. *Pedagogies: An International Journal,* 3, issue 1, pp. 22–36.

Wilson, J. and Czik, A. (2016). Automated essay evaluation software in English language arts classrooms: Effects on teacher feedback, student motivation, and writing quality. *Computers & Education,* 100, pp. 94–109.

Wilson, J. and Roscoe, R. D. (2020). Automated writing evaluation and feedback: Multiple metrics of efficacy. *Journal of Educational Computing Research,* 58, pp. 87–125.

Wilson, J., Cruz Cordero, T., Potter, A., Myers, M., MacArthur, C. A., Beard, G., Fudge, E. A., Raiche, A. and Ahrendt, C. (2024). Recommendations for integrating automated writing evaluation with evidence-based instructional practices. *International Journal of Changes in Education*, 2, issue 1, pp. 46–54.

Woodworth, J. and Barkaoui, K. (2020). Perspectives on using automated writing evaluation systems to provide written corrective feedback in the ESL classroom. *TESL Canada Journal,* 37, issue 2, pp. 234–247.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems,* 30.

Krek, S., Arhar-Holdt, Š., Erjavec, T., Repar, A., Gantar, P., Ljubešić, N., Kosem, I. and Dobrovoljc, K. (2020). *Gigafida 2.0: The Reference Corpus of Written Standard Slovene.* Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 3340–3345.

Zupanc, K. and Bosnić, Z. (2017). Automated essay evaluation with semantic analysis. *Knowledge-Based Systems*, 120, pp. 118–132.

Lin, S. (2024). Evaluating LLMs' grammatical error correction performance in learner Chinese. *PLOS ONE,* 19, issue (10): e0312881.

Grundkiewicz, R., Junczys-Dowmunt, M. and Gillian, E, (2015). Human evaluation of grammatical error correction systems. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pp. 461–470.

Matej KLEMEN (Univerza v Ljubljani, Filozofska fakulteta in Fakulteta za računalništvo in informatiko, Slovenija)
Martin BOŽIČ (Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Slovenija)
Špela ARHAR HOLDT (Univerza v Ljubljani, Filozofska fakulteta, Slovenija)
Marko ROBNIK-ŠIKONJA (Univerza v Ljubljani, Fakulteta za računalništvo in informatiko, Slovenija)

POPRAVLJANJE SLOVNIČNIH NAPAK V SLOVENSKIH ESEJIH Z VELIKIMI JEZIKOVNIMI MODELI

Povzetek: Strojno popravljanje slovničnih napak je naloga, ki zajema samodejno zaznavanje in popravljanje slovničnih napak v besedilu. Na področju izobraževanja je cilj metod pomagati učiteljem pri popravljanju napak učencev. Veliki jezikovni modeli omogočajo razvoj natančnih avtomatskih metod za zaznavanje in popravljanje določenih vrst napak. Da bi se izognili pretiranemu parafraziranju, ki je značilno za modele tipa GPT, in je v kontekstu poučevanja jezika nezaželeno, predstavimo več razvitih slovenskih modelov tipa BERT in T5 za popravljanje različnih vrst napak. Te vključujejo črkovalne napake, napake v rabi velikih začetnic, besednih oblik in besednega reda. V članku opišemo postopek ustvarjanja učnih podatkov, postopek učenja ter postopek evalvacije modelov na korpusu Šolar-Eval 1.0, ki vsebuje šolske spise osnovnošolcev in srednješolcev. Avtomatska evalvacija kaže razmeroma visoko natančnost razvitih modelov, medtem ko ročna kvalitativna evalvacija razkrije prednosti in slabosti razvitih modelov ter evalvacijskega postopka. Analiza razkriva številne izzive in obetavne smeri za nadaljnje izboljšave tako pri razvoju modelov kot pri postopku evalvacije.

Ključne besede: veliki jezikovni modeli, popravljanje slovničnih napak, izobraževalna domena, sintetiziranje podatkov

Elektronski naslov: matej.klemen@fri.uni-lj.si