sciendo
DOI 10.2478/jazcas-2025-0026

FROM RULE-BASED PROOFREADER BETA OPRAVIDLO TO AI-POWERED OPRAVIDLO 2.0

HANA ŽIŽKOVÁ[1] – ZUZANA NEVĚŘILOVÁ[2] – JAKUB MACHURA[3] – ALEŠ HORÁK[4] – DANA HLAVÁČKOVÁ[5] – PATRIK STANO[6]
[1]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6483-6603)
[2]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-7133-9269)
[3]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-6623-3064)
[4]Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic (ORCID: 0000-0001-6348-109X)
[5]Department of Czech Language, Faculty of Arts, Masaryk University, Brno, Czech Republic (ORCID: 0000-0002-9918-0958)
[6]Department of Machine Learning and Data Processing, Faculty of Informatics, Masaryk University, Brno, Czech Republic (ORCID: 0009-0001-8339-6084)

ŽIŽKOVÁ, Hana – NEVĚŘILOVÁ, Zuzana – MACHURA, Jakub – HORÁK, Aleš – HLAVÁČKOVÁ, Dana – STANO, Patrik: From Rule-based Proofreader Beta Opravidlo to AI-powered Opravidlo 2.0. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 290 – 299.

Abstract: The demand for accurate and error-free written communication in Czech has led to the development of automated proofreading tools. Beta Opravidlo, a rule-based online proofreader launched in 2022, demonstrated high precision and recall in correcting Czech texts. However, its reliance on predefined linguistic rules limited recall and processing speed. With advancements in machine learning and large language models (LLMs), a transition toward AI-powered proofreading became necessary. This article explores the evolution from Beta Opravidlo to Opravidlo 2.0, integrating deep neural networks to enhance correction capabilities. We discuss proofreading as a machine learning task, compare traditional rule-based and neural approaches, and challenges such as explainability, system integration or computational requirements. The most effective solution is a hybrid approach combining rule-based precision with AI-driven adaptability. Opravidlo 2.0 aims to improve recall, optimize inference time, and extend support to other Slavic languages. This interdisciplinary effort highlights the potential of AI-powered proofreading to set new standards in language correction and usability.

Keywords: Opravidlo, proofreader, Czech language, AI-powered proofreader

1 INTRODUCTION

The ability to express oneself in written and spoken language without linguistic errors is required and positively evaluated in the Czech environment. Readers often look at authors of texts who commit spelling mistakes with derision and distrust, and

texts with spelling mistakes reduce the author's credibility. Texts matter greatly for their correctness, even today are checked by human proofreaders. However, a human proofreader is not always convenient for various reasons: it is expensive and slow. Thus, in recent decades, all sorts of automatic tools have been developed to proofread text. This was not an easy task because Czech is an inflectional language with a lot of homonymy. One proofreader presented to the public is the rule-based online proofreader Opravidlo (Hlaváčková et al. 2022). Its beta version was released in mid-2022, and at that time, it was the proofreader with the highest precision and best recall. However, the situation changed with the advent of large language models (LLMs) and machine learning, whose concepts proved functional for correcting languages like Czech. The following article describes the starting point for Beta Opravidlo and the possibilities for AI-powered Opravidlo 2.0.

2 BETA OPRAVIDLO

Beta Opravidlo is an online proofreader freely available at www.opravidlo.cz. It was published in May 2022 and was created thanks to a project funded by TA ČR. The project was carried out in cooperation with three academic departments: Masaryk University, the Institute for Czech Language of the CAS, and Charles University. The team also included business partners Seznam.cz and Wikimedia ČR.

The project course was based on the experts' cooperation, knowledge-sharing, language data, software, hardware and some existing solutions to partial problems in developing the language proofreader. It also had the advantage of involving experienced experts (mostly linguists) and promising PhD students in computational linguistics in one project. The project's interdisciplinary nature brought together the knowledge of linguists, computational linguists and programmers, forming an ideal basis for developing the language proofreader.

The freely accessible web interface allows users to type or insert Czech text, and the right-hand side contains suggestions for corrections. The user can decide whether or not to accept the correction. Some typographical corrections are made automatically without user intervention. The correction refers to an explanation of the phenomenon found in the Internet Language Reference Book (2025) for complex topics like agreement or punctuation.

Since May 2022, the public has used the proofreader, with more than 200,000 correction requests per month, however, the low recall is perceived as a drawback by its users.

2.1 How Beta Opravidlo works

The Beta Opravidlo works by first decomposing the inserted text into tokens. The Unitok tool is used for this. The MorphoDita tagger or majka tagger performs the subsequent morphological analysis. The choice of tagger depends on the

subsequent processing. Considering that Beta Opravidlo is a rule-based system, a total of 7500 rules are included in the tool, divided into several thematic modules. In most modules, the detection phase is performed by the SET parser (Syntax Elements of Text) (Kovář et al. 2011). This parser is primarily designed as a universal, language-independent syntactic parser. It processes the input text into a tree structure according to a specific error-detection grammar. This grammar is defined in a text file containing various rules written in a specific format. However, as the example below demonstrates, the rules created for SET are not limited to syntactic parsing alone. The following rule captures the situation where the word *"více"* '*more'* is, incorrectly, directly followed by a comparative adjective without the conjunction *"než" 'than',* unless this is remedied by a noun in the genitive plural as a third word in the prepositional phrase:

```
TMPL:(word více)(tag k2.*d2.*)
MARK 0 DEP 1 LABEL <komparativ-nok> PROB 100
TMPL:(word více)(tag k2.*d2.*)(tag kl.*nP.*c2.*)
MARK 0 DEP 1 LABEL <komparativ-ok> PROB 400
```

The first rule marks the word *"více"* '*more*' as redundant in following sentence: *"Bylo to více horší, než jsme čekali."* '*It was more worse than we expected.'* The second rule applies to grammatically correct sentences of the type: *"Dostal více těžších úkolů." 'He was given more difficult tasks.'*

The Beta Opravidlo contains the following modules:

- punctuation,
- non-grammatical structures,
- spelling,
- spelling in context,
- agreement,
- typography,
- dependent clauses,
- capital letters,
- preposition vocalisation,
- pronouns,
- other errors.

This modular system's advantage is that it can analyse the text in parallel. However, evaluating an error still takes several tens of seconds, which is a significant disadvantage.

The exact list of language phenomena that Beta Opravidlo can correct is available on this page https://www.opravidlo.cz/co-korektor-umi.html. The precision of the individual modules ranges from 91% to 96%, which we consider an excellent result. Regarding recall, given the nature of the Czech language, the lowest coverage is 40%,

and the best is 80%. We have found that the success rate of recall is highly dependent on the nature of the text and the number of errors in the text.

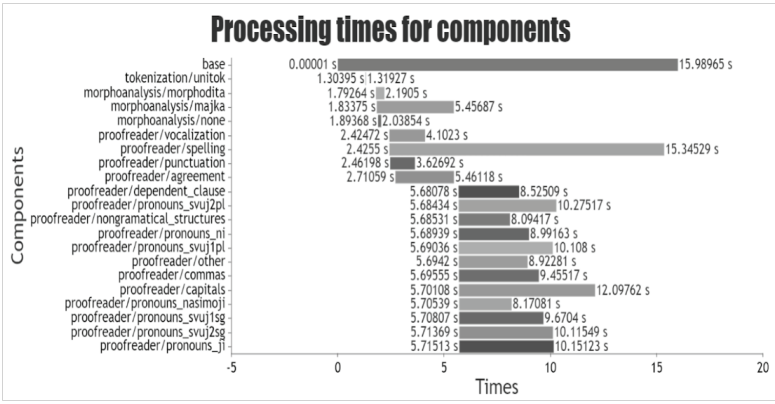

Fig. 1. Efficiency of Beta Opravidlo (Mrkývka 2024)

3 FROM BETA OPRAVIDLO TO OPRAVIDLO 2.0

We received valuable feedback from users while testing the Beta Opravidlo before its release. On the one hand, it was very positively evaluated that we follow the rule-based path in developing the proofreader. On the other hand, the question of incorporating machine learning and using neural networks was discussed, which was not planned in the project. When we started developing the Beta Opravidlo in 2019, it was still unclear what progress machine learning would make, yet we considered incorporating machine learning into some modules. Experimentally, we performed comparisons of punctuation insertion, and it turned out that neural networks showed higher recall and, on some texts, higher precission (Machura et al. 2022). In mid-2022, we believed the issue stemmed from specific errors that neural networks failed to correct, assuming they simply required retraining. In the following years, and thanks to experimental studies, it turned out that neural networks could be very effective in correcting Czech texts (Machura et al. 2023; Medková and Horák 2022).

All these findings have resulted in the creation of a new interdisciplinary team, working on integrating deep neural networks into the Beta Opravidlo, with the aim of increasing its ability to correct language errors first in the Czech language, subsequently in other Slavic languages.

4 PROOFREADING AS A MACHINE LEARNING TASK

Proofreading is a complex task combining several subtasks; from the functional point-of-view, proofreading consists of:

- error detection,
- suggestions for error correction,
- explanation of the error.

From the language point-of-view, proofreading can be seen as a combination of:

- spellchecking,
- grammar correction,
- typography correction,
- style improvement.

The evolution of proofreading consists of changes in methods, coverage of text phenomena, and performance metrics used for comparison.

In the 1980s, natural language processing (NLP) addressed the "ill-formed input", where parsers struggled to output a parse tree. With the rise of machine learning, grammar correction was formulated as an NLP task: grammar error correction (GEC). In (Wang et al. 2020), GEC is described as "Errors that violate rules of English and expectation usage of English native speakers in morphological, lexical, syntactic and semantic forms are all treated as a target to be corrected." GEC systems usually get an ungrammatical sentence as input and output the corrected sentence. This approach is a sister task of machine translation. Therefore, GEC systems followed a similar path in their methods: early approaches used statistical methods, and later, neural architectures such as long short-term memory (LSTM), and convolutional neural networks (CNN). Current studies predominantly examine transformer-based models, including BERT variants and LLMs like PaLM2-XS (Liu et al. 2024).

Early proofreading systems focused only on spellchecking, solving the task with a "dictionary" – a simple wordlist for a particular language. Later, GEC systems focused on fewer errors, such as correcting prepositions (Prokofyev et al. 2014). Recent GEC systems attempt to perform comprehensive error correction. The coverage of the text phenomena is connected with the used methods: for tasks more related to language rules, rule-based systems or n-gram statistics perform well, for punctuation or fluency-related issues, a larger context is needed, and therefore, transformer-based methods yield better results.

Evaluation metrics also changed from accuracy measures to metrics for fluency and overall text quality. Particularly, GEC systems commonly use F0.5 score, GLEU, BLEU, METEOR, precision, recall, and F1 score as performance metrics, with recent work incorporating broader evaluation frameworks.

The majority of GEC systems are trained and evaluated in English. More recently, with the emergence of multilingual models, GEC for non-English texts is achieving plausible results. Multilingual transformer models, particularly mT5, show promise in handling non-English languages due to their pre-training on multilingual datasets. Successful GEC systems are developed e.g. for Arabic, a more recent work for a Slavic language is (Kholodna and Vysotska 2023). A recent and comprehensive survey on GEC can be found in (Bryant 2023).

4.1 Challenges in transition to machine learning system

While the rule-based method is effective in targeted error correction (e.g. the correct form of pronouns), it struggles with long-range dependencies or syntactic ambiguity. Transformer-based approaches are better at text understanding so that they can handle the text in a more comprehensive way, potentially with much higher recall. The targeted error correction has a strong advantage we would like to keep and develop: the errors are classified and can be explained easily. For example, if the rule-based system detects an incorrect pronoun form, it can label the error, provide a correct form, and explain the rules for pronominal inflection. With a simple machine translation-like approach, we would lose the system's ability to explain errors and teach the language users.

Currently, we perform experiments in several streams:

- filling in the punctuation,
- edit-based approaches,
- grammar error explanation methods.

4.2 Filling in the punctuation

Since missing punctuation is one of the most common errors that is also difficult to capture by rules, we focus on punctuation errors in the first phase, similarly to (Machura et al. 2023). A Czech variant of the RoBERTa model has been modified by the addition of a classifier head and fine-tuned to classify tokens based on the presence or absence of a comma. This approach can be extended to include other punctuation marks (e.g. a full stop, a question mark, an exclamation point). Other grammatical phenomena, such as casing, can be resolved similarly, possibly by the same model, by including a second classification head and fine-tuning both tasks. This approach simplifies the grammar correction task, enhancing the achieved precision and recall. However, it lacks explainability, which is critical for a reliable grammar correction service, as it provides credibility. Consequently, a classifier model is a useful secondary option in addition to an approach that provides explanations. The confidence of classification can be utilised to assess its credibility, and the confidence score could be displayed to the end user, enabling them to decide whether to accept the suggestion to add punctuation or not. This approach is to be tested and evaluated.

4.3 Edit-based approaches

In (Omelianchuk et al. 2020), the authors generate a sequence of token-level edits to perform grammatical error corrections. The advantages of such an approach are: 1) minimum intervention in users' text, and 2) explainability of the errors.

Our experimental setup is a sequence to edit architecture, where each token of the input sequence gets labels such as KEEP, DELETE, APPEND, REPLACE, and TRANSFORM. The last label is enriched with the type of transformation. Each label can be enriched by the explanation.

So far, we have used synthetic data. We introduced errors into the Czech part of the WMT dataset (Bojar et al. 2014) – we removed punctuation, added punctuation after random tokens, and converted capitalised tokens to lowercase in various combinations.

Gold standard	Pamatujte: kdo rychle dává, dvakrát dává.
Input	pamatujte: kdo, rychle dává dvakrát dává
Output labels	$MAKE_CASE_UPPER $EXTRA_COMMA $KEEP $MISSING_PUNCT_, $KEEP $MISSING_PUNCT_.

Tab. 1. Example of the training data

For the task, we fine-tuned the RobeCzech-base (Straka et al. 2021) model with 953,620 sentences, for evaluation, we used 2,999 sentences. The system achieved F1=96.7. We know punctuation and capitalisation are only a small part of the proofreading task; however, the results seem very promising, and we plan to continue with this approach.

4.4 Grammar error explanation methods

In Song et al. (2024) the authors performed a series of experiments with ChatGPT-4 to explain grammar errors in natural language. They developed a two-step pipeline that leverages fine-tuned and prompted LLMs to perform structured atomic token edit extraction, followed by prompting GPT-4 to explain each edit.

We do not plan to use generative LLMs in the production version. The main reasons are deployment costs and prediction time. However, using generative LLMs for comparison and evaluation is desirable.

4.5 A hybrid approach to proofreading

Currently, a hybrid approach seems to be beneficial. We plan to keep the rule-based approach for phenomena well covered by the rules (high precision and high recall) and at the same time, the inference time is not longer than that of a neural model. In the later development phase of the neural approaches, we will decide whether the neural model outperformed the rule-based system in some aspects. Also, an ensemble model could be made, possibly including even multiple models.

While the rule-based system is not scalable, the hardware can influence the prediction time of neural models. Currently, we plan to deploy the proofreading service on GPU servers at the Faculty of Informatics at Masaryk University. It depends on the possibilities of large research infrastructures such as CLARIN whether a future GPU deployment would be possible.

5 PREPARATION OF TESTING DATA

For the purpose of training models for automatic comma insertion, a comprehensive analysis of comma distribution was conducted using the SYN2020 corpus (Křen et al. 2020). This corpus contains just over 8 million commas. The study focused on the following aspects:

a. The most common lexical contexts, specifically:

a) the most frequent expressions that appear immediately after commas,
b) the most common particles and interjections that occur before commas, and
c) the most frequent vocative phrases, which must be separated by a comma as they lie outside the syntactic structure of the sentence.

The aim of this part of the analysis was to determine the actual distribution of commas in relation to specific lexical expressions—specifically, how frequently a given expression appears with a preceding comma and in what proportion it occurs without one. For instance, the word "že" 'that' appears 951,302 times in the SYN2020 corpus, with 902,351 of these instances (94.85%) following a comma. The remaining 5.15% occur without a preceding comma.

The goal is to construct a testing dataset that accurately reflects these proportions. Hypothetically, if the dataset contained 100,000 commas, 11.25% of those commas would be followed by the expression *"že" 'that'*, and additionally, 5.15% of the total occurrences of *"že" 'that'* would appear without a preceding comma. This approach ensures realistic representation of the expression *"že" 'that'* and allows the model to learn both its typical usage with commas and less frequent instances without them.

b. Proportion of sentences of a given token length with and without commas

This parameter examines the ratio of sentences of a specific length (in number of tokens) that contain at least one comma versus those of the same length that contain none. The analysis is based on the SYN2020 corpus, which includes 6,791,880 sentences ending with a period, exclamation mark, or question mark.

The most frequent sentence length in the corpus is 9 tokens (identified using the CQL query: `<s> [word!="\.|\!|\?"]{9} [word="\.|\!|\?"] </s>)`, totaling 323,369 sentences—representing 4.76% of all sentences. Of these, 129,130 sentences contain at least one comma (1.90% of all sentences), while 194,239 contain no commas (2.86%).

To ensure the testing dataset reflects these proportions, the same ratios are applied. Hypothetically, if the testing dataset includes 100,000 sentences, 1.90% should be 9-token sentences that include a comma, and 2.86% should be 9-token sentences without any comma. These ratios were similarly calculated for all sentence lengths ranging from 1 to 45 tokens.

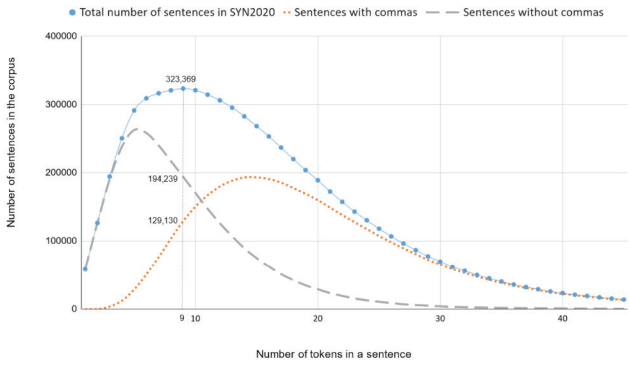

Fig. 2. Proportion of sentences of a given token length with and without commas

6 CONCLUSION

The development of Beta Opravidlo has demonstrated the strengths of rule-based proofreading, particularly in precision and linguistic transparency. However, the advent of machine learning and neural networks offers significant potential to improve recall and overall correction effectiveness. By integrating AI-driven approaches, Opravidlo 2.0 aims to enhance recall, provide high precision, and expand its capabilities beyond Czech to other Slavic languages.

Despite challenges such as data availability, explainability, and system integration, a hybrid approach combining rule-based and machine-learning methods appears to be the most promising solution. Future research and development efforts will focus on refining this hybrid model, optimising deployment infrastructure, and ensuring a seamless user experience. With continued interdisciplinary collaboration, Opravidlo 2.0 has the potential to set a new standard in automated proofreading for complex languages like Czech.

ACKNOWLEDGEMENTS

The authors acknowledge that this work was supported by the OSCARS project, funded by the European Commission's Horizon Europe Research and Innovation programme (grant agreement No. 101129751), led by the five Science Clusters: ENVRI, ESCAPE, LS RI, PaNOSC, and SSHOC.

References

Bojar, O. et al. (2014). Findings of the 2014 Workshop on Statistical Machine Translation. In Proceedings of the Ninth Workshop on Statistical Machine Translation, Baltimore, Maryland, USA. ACL, pp. 12–58.

Bryant, Ch. et al. (2023). GEC: A Survey of the State of the Art. Computational Linguistics 2023; 49(3), pp. 643–701. Accessible at: https://doi.org/10.1162/coli_a_00478.

Internet Language Reference Book (2025). Praha: ÚJČ AV ČR.

Hlaváčková D. et al. (2022). Opravidlo.

Kholodna, N., and Vysotska, V. (2023). Technology for grammatical errors correction in Ukrainian text content based on machine learning methods. Radio Electronics, Computer Science, Control, (1), 114. Accessible at: https://doi.org/10.15588/1607-3274-2023-1-12.

Kovář, V. et al. (2011). Syntactic Analysis Using Finite Patterns: A New Parsing System for Czech. In Human Language Technology. Challenges for Computer Science and Linguistics. Berlin/Heidelberg: Springer, pp. 161–171. Accessible at: http://dx.doi.org/10.1007/978-3-642-20095-3_15.

Křen, M. et al. (2020). SYN2020: A representative corpus of written Czech. UCNK FF UK. Accessible at: http://www.korpus.cz.

Liu, R. et al. (2024). Proofread: Fixes All Errors with One Tap. Accessible at: arXiv preprint arXiv:2406.04523.

Machura, J. et al. (2022). Automatic Grammar Correction of Commas in Czech Written Texts. Online. In: P. Sojka et al. (eds): TSD 2022. Cham (CH): Springer, pp. 113–124. Accessible at: https://dx.doi.org/10.1007/978-3-031-16270-1_10.

Machura, J. et al. (2023). Is it Possible to Re-educate RoBERTa? Jazykovedný časopis, 74(1), pp. 357–368. Accessible at: https://dx.doi.org/10.2478/jazcas-2023-0052.

Medková, H., and A. Horák. (2022). Distinguishing the Types of Coordinated Verbs with a Shared Argument by means of New ZeugBERT Language Model and ZeugmaDataset. In: A. Dimou et al. (eds.): Towards a Knowledge-Aware AI: SEMANTiCS 2022. Amsterdam: IOS Press, pp. 206–218. Accessible at: https://dx.doi.org/10.3233/SSW220022.

Mrkývka, V. (2023). Webový korektor jako prostředek formalizace pravidel českého jazyka. PhD Thesis, Brno: MU.

Omelianchuk, K. et al. (2020). GECToR – Grammatical Error Correction: Tag, Not Rewrite. In Proceedings of the 15th Workshop on Innovative Use of NLP for Building Educational Applications, Seattle, WA, USA, pp. 163–170.

Prokofyev, R. et al. (2014). Correct Me If I'm Wrong: Fixing Grammatical Errors by Preposition Ranking. In Proceedings of CIKM'14. Association for Computing Machinery, New York, NY, USA, pp. 331–340. Accessible at: https://doi.org/10.1145/2661829.2661942.

Song, Y. et al. (2024). GEE! Grammar Error Explanation with Large Language Models. In Findings of the Association for Computational Linguistics: NAACL 2024, Mexico City, Mexico. ACL, pp. 754–781.

Straka, M. et al. (2021). RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model. In: K. Ekštein et al. (eds): TSD 2021. Lecture Notes in Computer Science, Vol. 12848. Springer, Cham. Accessible at: https://doi.org/10.1007/978-3-030-83527-9_17.

Wang, Y. et al. (2020). A comprehensive survey of grammar error correction. Accessible at: arXiv preprint arXiv:2005.06600.