UDC 004.8:811.134.3'36=111
Original scientific paper
Accepted for publication on 20. 9. 2024.
https://doi.org/10.29162/jez.2024.6

Vedran Juričić
Daliborka Sarić
Faculty of Humanities and Social Sciences
University of Zagreb

Evaluation of AI-based Grammar Correction for Portuguese

Correction of grammatical errors is today integrated into the most widely used text processing tools and is accessible online. However, these tools are primarily half-automatic, merely suggesting possible corrections and variations, and require interaction with a user, which can be a tedious task when used on lengthy texts. Recent advancements in the field of artificial intelligence and natural language processing offer a more efficient strategy. This paper analyzes a possibility of using ChatGPT for correcting grammar in Portuguese texts written by native speakers of Croatian. The texts were corrected by a native speaker of European Portuguese and by ChatGPT. The authors analyzed error detection and correction at various linguistic levels and accompanied it with examples. Due to class imbalance, the system's performance was evaluated using the F-measure. The calculation of false positives and true negatives was adjusted because of special cases of improper correction. Taking that into consideration, the $F_{0.5}$ score was 0.805. Nevertheless, it should be noted that the results would likely vary if the input corpus had different structure and proficiency level.

Keywords: Grammar correction, ChatGPT, Evaluation, System's performance

1. Introduction

There are several non-AI-based tools designed to offer suggestions for spelling, grammar and style in written Portuguese. However, their effectiveness, from the perspective of non-native speakers, appears to be limited. For instance, FliP *Correc-*

tor ortográfico e sintáctico (Priberam 2022), a tool based on pre-defined patterns, besides lacking a higher degree of context-awareness, provides numerous options for substituting words, which means that its successful use presupposes an already proficient command of Portuguese. The first automated grammatical error corrections were largely based on manually defined rules and sentence parsers, but their performance was not satisfactory enough. For example, CyWrite's F-score on articles was 0.65 and on run-on sentences 0.41. This is a relatively low score, but the system still shows clearer results and better performance compared to a popular commercial tool at the time, Criterion, which had F-score of 0.59 on articles and 0.07 on run-on sentences (Feng et al. 2016; Li et al. 2014).

In recent years, significant advancements have been achieved in the realm of grammatical error correction (GEC).[1] Different rule-based methods, classifiers, machine translation methods and neural machine translation systems have been developed and extensively studied (Bryant et al. 2023). Transformer models are nowadays widely used deep neural network architecture in natural language processing that can handle long-range dependencies (Vaswani 2017). Large language models (LLMs) are a type of artificial intelligence designed for understanding and generating human language, utilizing transformer architectures and trained on billions of parameters from extensive text datasets, including web articles, social networks, books, and other online sources.

One of the important LLMs is Google's T5, Text-to-Text Transfer Transformer, with 11 billion parameters (Raffel et al. 2020). Brown et al. (2020) trained GPT-3 (Generative Pre-trained Transformer) with 175 billion parameters, and its successor GPT-4, developed by OpenAI, was trained with 170 trillion parameters (OpenAI 2023a). Due to their heterogenous training data and architecture, these models have a deep understanding of the overall grammatical and semantic structure of natural language (Waisberg et al. 2023). ChatGPT (OpenAI 2023b) is a chatbot web application relying on GPT3.5 or GPT4 and optimized for interaction with humans. It enables writing prompts, filters content and generates output.

Considering ChatGPT's capability of understanding and producing text in various languages, we examine its performance as a *grammar checker* for Portuguese or, more specifically, a tool that can be used to detect and correct errors[2] made by

[1] In the context of correction tools, *grammatical* error is used metonymically, extending beyond traditional notions of *grammar* to encompass all linguistic levels.

[2] The term *error* is a relative concept in this context, and we use it in the following text for the sake of simplicity. Errors made by non-native speakers are, from the perspective of the interlanguage they create and use, features of their system, part of the interlanguage grammar (Corder 1967, Gass & Selinker 2008). The actual task of ChatGPT is to translate from *interlanguage Portuguese* to Portuguese.

non-native speakers.[3] In the following chapters we describe our experiment and present the results and conclusions.

2. Methodology

The corpus of analyzed texts consisted of eight free compositions written in a classroom setting by non-native speakers of Portuguese (at the B2 proficiency level), third-year students of Portuguese language and literature (Faculty of Humanities and Social Sciences, University of Zagreb). The informants in this study were native speakers of Croatian. The texts underwent correction by a native speaker of European Portuguese, a contractual lecturer in the program. The original texts were inputted into ChatGPT (GPT 3.5), tasked with correcting errors according to the European variety of Portuguese. Both the native speaker and ChatGPT were provided with minimal instructions, solely directed to correct the text. The only distinction was the directive for ChatGPT to adhere to European Portuguese. The output from ChatGPT was then compared to the correct sentences and classified into proper and improper modifications. Throughout the remainder of this paper, the texts written by non-native speakers are referred to as the *source*, those corrected by the native speaker as the *reference*, and ChatGPT's output as the *target*.

In order to evaluate the system's performance, the error correction task was approached as a binary classification problem. Any language unit in source and target text, regardless of whether it is a character, a word or a sentence, can be compared with a certain unit in the reference text and classified as correct or incorrect. There are only two classes or labels for each unit, and thus, it is a binary classification.

The corpus consists of 69 sentences, 141 clauses and 762 words, with an average of 9.7 words per sentence. The technical definition of the term *word* relevant for this research encompasses all units separated by spaces or hyphens, meaning that enclitic pronouns and lexical components of compounds are treated as separate units. A pivotal decision in this study pertained to the choice of linguistic units for comparison. Sentences were excluded due to inherent issues that would negatively impact performance measurement. Primarily, the source text contains very few sentences that are completely correct. Consequently, a significant imbalance between correct and incorrect classes would arise, leading to evaluations primarily calculated on errors. Another reason is that some sentences contain multiple errors. Improper correction of only one error would label an entire sentence as incorrect, although the system properly corrected all other errors. Moreover, appropriately correcting a sentence with multiple errors would have the same impact on the system's performance as correcting a sentence with only one error. The characters

[3] Other available, though less popular, tools include Gemini, Bloom, and the open-source LLaMA 2.

were not considered relevant units as the performance would be influenced by word or sentence lengths. The correct words that are longer would benefit performance more than the shorter ones, and the same applies to the incorrect words. Furthermore, interpreting the results of such an analysis would prove challenging, i.e., what would be the significance of a performance of a system achieving 78% accuracy in correcting the characters?

ChatGPT's performance was evaluated using the F-score, a measure that combines precision and recall (or sensitivity) into a single number. In a binary classification problem, the classes are mostly labeled with ones and zeros, and referred to as positive and negative. *Precision* shows what ratio of positive predictions are correct, whereas *recall* shows what is the ratio of positive correct predictions in all the positive cases. Precision and recall are measures obtained from four numbers:

1. True Positive (TP) is a number of positive cases that a system correctly classified
2. True Negative (TN) is a number of negative cases that a system correctly classified
3. False Positive (FP) is a number of cases a system incorrectly classified as positives (they are actually negative)
4. False Negative (FN) is a number of cases a system incorrectly classified as negatives (they are actually positive)

Precision is then the ratio between true positives and the sum of true positives and false positives, that is TP / (TP + FP). Recall is the ratio between true positives and the sum of true positives and false negatives, that is TP / (TP + FN). The F-score has one parameter, usually noted by beta, which determines the importance of precision in respect to recall. If its value is 1, precision and recall are equally important. Values exceeding 1 favor recall and vice versa. The formula for the F-score is as follows:

$$F_\beta = (1+\beta^2)\cdot\frac{P\cdot R}{(\beta^2\cdot P)+R}$$

where P represents precision and R recall. In this paper we use value of 0.5, that is $F_{0.5}$.

In the context of an error correction system, a *positive* is a word in the source text that is incorrect, while a *negative* does not require any correction. Calculating the above numbers is trivial at first, but as it turned out, there are two cases that need to be considered separately.

Table 1 shows all the possible cases and their mapping to True Positive, True Negative, False Positive and False Negative. Special cases are marked with an asterisk. As shown in Table 1, if a source word is incorrect and ChatGPT made a proper correction, this is the case of a True Positive. If it did not detect an error, this is a False Negative. The special case is when ChatGPT detects an error but does not correct it properly. It should also be treated as a False Negative since the output word does not belong to the correct class. The same applies to the case when a source word is correct. If ChatGPT failed to correct it properly, it is a False Positive.

Table 1. Possible cases in error correction problem

SOURCE WORD	MODIFICATION	MEASURE
INCORRECT	None	False Negative
	Proper correction	True Positive
	Improper correction	False Negative*
CORRECT	None	True Negative
	Proper correction	False Positive
	Improper correction	False Positive*

3. Results

The output texts were compared with both the source and the reference texts. Table 2 shows the confusion matrix, a two-dimensional matrix that shows all combinations of actual (reference) and predicted (output) values.

Table 2. Confusion matrix

		Output values	
		1	**0**
Reference values	**1**	111	16
	0	29	606

As can be seen, ChatGPT had 111 correct modifications. For 606 words it correctly detected they contained no errors and did not change them. Its output contained 717 correct words (111+606), and 45 incorrect words (16+29). Based on these numbers, we calculated precision, recall and F-score, which are shown in Table 3.

Table 3. Performance metrics and values

Measure	**Value**
P	0.790
R	0.874
F	0.805

Precision of 0.79 means that 22% (1 – P) of modifications were not necessary and resulted in a correct or incorrect word. On the other hand, recall of 0.874 means that ChatGPT failed to detect or improperly corrected 12.6% (1 – R) of words. Their weighted harmonic mean gives a F-score of 0.805, which is generally considered a very good value.

Based on the values in Table 2, we can also calculate an alternative measure, accuracy. Accuracy is the ratio of correctly classified cases in a total number of cases. Accuracy is 94.23% but should not be considered as a measure in this case because of a great class imbalance, i.e., the ratio of positive and negative cases in the source text is almost 1:5.

Below, we present some examples of True Positive cases, i.e., cases where reference and target sentences are equal. ChatGPT identifies errors at various linguistic levels. Among syntactic errors, it successfully corrects the usage of, for example, prepositions and articles, through elimination (the preposition *de* in ex. 1 and the article *a* in ex. 2 are omitted), addition (in ex. 3 the preposition *de* is added, thereby forming a contraction with the pronoun; in ex. 4, the article *o* is added) and substitution (in ex. 5 the preposition *a* as a component of the contraction with the article (à) in ex. 5 is replaced with the preposition *para*).

1. Source Depois, decidiu **de** ir para centro de cidade para beber o café com as minhas amigas.
   Target Depois, decidi ir para o centro da cidade para tomar café com as minhas amigas.
   Reference Depois, decidi ir para o centro da cidade para tomar café com as minhas amigas.
   ‘Afterwards, I decided to go to the city center to have coffee with my friends.’

2. Source Voltei a casa às 13 horas porque tenho as aulas **do** francês às 13.
   Target Voltei a casa às 13 horas porque tenho aulas **de** francês às 13.
   Reference Voltei a casa às 13 horas porque tinha aulas **de** francês às 13.
   ’I returned home at 1:00 PM because I have/had French classes at 1:00.’

3. Source Depois isso, arranjei o apartamento e comi um pouco.
   Target Depois **disso**, arranjei o apartamento e comi um pouco.
   Reference Depois **disso**, arrumei o apartamento e comi um pouco.
   ‘After that, I tidied up the apartment and ate something.’

4. Source Depois lavei cabelo e tomei o pequeno-almoço.
   Target Depois lavei **o** cabelo e tomei o pequeno-almoço.
   Reference Depois lavei **o** cabelo e tomei o pequeno-almoço.
   ‘Afterwards, I washed my hair and had breakfast.’

5. Source Na sexta-feira fui **à** cama às 8 horas da manhã [...]
   Target Na sexta-feira fui **para** a cama às 8 horas da manhã [...]
   Reference Na sexta-feira fui **para** a cama às 8 horas da manhã [...]
   ‘On Friday, I went to bed at 8 o’clock in the morning [...]’

The excessive use of first-person singular pronouns in the subject[4] position has also been corrected:

6. Source Antes de jogar um jogo e falar com a minha irmã, **eu** jantei.
Target Antes de jogar um jogo e falar com a minha irmã, jantei.
Reference Antes de jogar um jogo e falar com a minha irmã, jantei.
'Before playing a game and talking to my sister, I had dinner.'

Furthermore, morphological (incorrect verb root in 7) and morphosyntactic (concordance in 8) errors are successfully corrected:

7. Source Então o meu namorado também chegou e à noite **fuimos** todos a sair.
Target Então, o meu namorado também chegou e à noite **fomos** todos sair.
Reference Então, o meu namorado também chegou e à noite **fomos** todos sair.
'Then, my boyfriend also arrived, and in the evening, we all went out.'

8. Source Festejamos **todo** a noite.
Target Festejamos **toda** a noite.
Reference Festejámos **toda** a noite.
'We celebrated all night.'

There are also some examples of appropriate lexical substitutions (the Target's option is different from the Reference's but it is equally suitable):

9. Source Depos disso, tomei um pequeno almoço e, enquanto almoçava, **guardava** uma série no Netflix.
Target Depois disso, fiz um pequeno-almoço e, enquanto comia, **assisti** a um episódio de uma série na Netflix.
Reference Depois disso, tomei um pequeno-almoço e, enquanto almoçava, **vi** uma série na Netflix.
'After that, I had breakfast and, while I was having lunch, I watched a series on Netflix.'

A good mastery of temporal-aspectual semantics can also be observed. For example, temporally delimited states and events in the past are systematically shifted

[4] The analysed texts exhibit excessive use of subject pronouns from the point of view of European Portuguese, a typical pro-drop language. Brazilian Portuguese, which was not tested in this work, has a somewhat different status regarding this parameter.

from the imperfect to the preterite tense (10), as well as successive episodic situations (11):

10. Source Quando cheguei à casa, **dormia** por uma hora porque estava muito cansada.
Target Quando cheguei a casa, **dormi** por uma hora porque estava muito cansada.
Reference Quando cheguei a casa, **dormi** durante uma hora porque estava muito cansada.
‘When I got home, I slept for an hour because I was very tired.’

11. Source Depois de escrever o TPC **ia** para o supermercado com a minha mãe.
Target Após fazer os trabalhos de casa **fui** para o supermercado com a minha mãe.
Reference Depois de escrever o TPC **fui** ao supermercado com a minha mãe.
‘After doing my homework, I went to the supermarket with my mom.’

The proper correction of already correct source units is not uncommon (False Positive cases). Intervening on correct words by replacing them with their synonyms results in grammatically and semantically correct words and sentences but they may carry a different stylistic nuance and change the register to a more formal variant (as is evident in the selection of a lexeme in 12, the use of the preposition *após* instead of maintaining *depois de* in 13, or opting for the conditional form of the verb instead of the more informal imperfect).

12. Source Logo depois **voltei a**os estudos e fui dormir às 3 horas de manhã.
Target Logo depois, **retomei** os estudos e fui dormir às 3 horas da manhã.
Reference Logo depois **voltei a**os estudos e fui dormir às 3 horas da manhã.
‘Right after that, I returned to studying and went to bed at 3 o’clock in the morning.’

13. Source **Depois d**a aula gastei um tempo falando com a minha amiga.
Target **Após** a aula, passei um tempo conversando com a minha amiga.
Reference **Depois d**a aula estive um tempo falando com a minha amiga.
‘After class, I spent some time chatting with my friend.’

14. Source Estava a estudar a noite inteira e já sabia que **ia** dormir muito tarde.
Target Passei a noite inteira estudando e já sabia que **iria** dormir muito tarde.
Reference Estive a estudar anoite inteira e já sabia que **ia** dormir muito tarde.
‘I studied all night long and I already knew I was going to sleep very late.’

We consider as False Positive* some improper corrections (over-corrections) of correct source words. Such interventions result in grammatically correct sentences with, possibly, a slightly different interpretation, as demonstrated by the unnecessary addition in example 15 (*um episódio de uma série* 'an episode of a series' instead of the original *uma série* 'a series'), or a completely different meaning, as in the case of replacing the verb *tomar* in the phrase *tomar o pequeno-almoço* 'to have a breakfast' with *fazer* which means 'to make (a breakfast)' in example 16:

15. Source Depos disso, tomei um pequeno almoço e, enquanto almoçava, guardava **uma série** no Netflix.
Target Depois disso, fiz um pequeno-almoço e, enquanto comia, assisti a **um episódio de uma série** na Netflix.
Reference Depois disso, tomei um pequeno-almoço e, enquanto almoçava, vi **uma série** na Netflix.
'After that, I had breakfast and, while I was having lunch, I watched a series on Netflix.'

16. Source Depos disso, **tomei** um pequeno almoço e, enquanto almoçava, guardava uma série no Netflix.
Target Depois disso, **fiz** um pequeno-almoço e, enquanto comia, assisti a um episódio de uma série na Netflix.
Reference Depois disso, **tomei** o pequeno-almoço e, enquanto almoçava, vi uma série na Netflix.
'After that, I had breakfast and, while I was having lunch, I watched a series on Netflix.'

Although ChatGPT generally distinguishes between the Brazilian and European varieties of Portuguese and makes corrections, as requested, to European Portuguese, typically Brazilian elements occasionally appear in corrected texts, for example, the gerund form in the progressive construction instead of *a + infinitive*:

17. Source Estava **a estudar** a noite inteira e já sabia que ia dormir muito tarde.
Target Passei a noite inteira **estudando** e já sabia que iria dormir muito tarde.
Reference Estive **a estudar** a noite inteira e já sabia que ia dormir muito tarde.
'I studied all night long and I already knew I was going to sleep very late.'

We have identified 16 instances requiring correction (False Negatives), encompassing various linguistic levels, including the usage of prepositions (18, 19), tense sequencing (20), noun gender (21), orthography (22).

18. Source Depois disso, fui ao bar a **encontrar** a minha amiga Nusha.
Target Depois disso, fui ao bar **encontrar** a minha amiga Nusha.
Reference Fui ao bar **encontrar**-**me com** a minha amiga.
'I went to the bar to meet up with my friend.'

19. Source Depois do almoço estudei um pocadinho e **em** 2 horas foi encontrar-se com as minhas amigas no parque Maksimir.
Target Após o almoço, estudei um pouco e, **em** duas horas, encontrei-me com as minhas amigas no Parque Maksimir.
Reference Depois do almoço, estudei um pouco e às 2 horas fui encontrar-me com as minhas amigas no parque Maksimir.
'After lunch, I studied for a bit, and at 2 o'clock, I went to meet my friends at Maksimir Park.'

20. Source Voltei a casa às 13 horas porque **tenho** as aulas do francês às 13.
Target Voltei a casa às 13 horas porque **tenho** as aulas do francês às 13.
Reference Voltei para casa às 13 horas porque **tinha** aulas de francês às 13.
'I returned home at 1:00 PM because I have/had French classes at 1:00.'

Although the sentence featuring the present tense in the subordinate clause is accurate, the native speaker (Reference) inferred, from a broader contextual understanding, which included a proficient grasp of the student's interlanguage, that the intended meaning here was episodic rather than habitual, prompting the application of tense sequencing rules.

Likely owing to a deficiency in contextual information and familiarity with the particular interlanguage accustomed by the lecturer (Reference), ChatGPT occasionally fails to detect certain covert errors, such as the misattribution of gender to the noun *bolo* 'cake'. Although the noun *bola* primarily means 'ball', the sentence is acceptable because it could possibly refer to the name of a culinary item. However, the nuanced comprehension of the student interlanguage, which the professor naturally masters better than ChatGPT, comes to the forefront again.

21. Source Preparámos **uma bola** muito simples.
Target Preparamos **uma bola** muito simples.
Reference Preparámos **um bolo** muito simples.
'We made a very simple cake.'

According to the reformed spelling, the accent on preterite forms is optionally used, but in European Portuguese it continues to appear systematically in texts. However, the accent is missed in corrected Target texts:

22. Source **Festejamos** todo a noite.
Target **Festejamos** todo a noite
Reference **Festejámos** toda a noite.
'We celebrated all night.'

The data did not reveal any False Negatives*.

4. Discussion

Based on the corpus used, the value of the $F_{0.5}$-score was determined to be 0.805. This is somewhat larger than the value reported by Penteado & Perez (2023). They conducted a similar study in Brazilian Portuguese, but their corpus comprised texts produced by native speakers. The sentences were divided into categories (grammar, spelling, fast typing and internet language), with incorrect and correct pairs provided[5]. The mean value of $F_{0.5}$ score for ChatGPT utilizing GPT3.5 was 0.737. This was outperformed by Google Docs with the mean $F_{0.5}$ score of 0.818. The same analysis was made with GPT-4, which revealed superior performance, achieving the mean $F_{0.5}$-score of 0.91. Penteado & Perez (2023) determined that GPTs have lower precision as they prioritize fluency over grammatical accuracy, which results in unnecessary modifications in the text and an increase in false positives. It is worth noting that this paper is the only study we found that explores error correction in Portuguese.

Wu et al. (2023) evaluated multiple systems on an English dataset specially created for grammatical error correction. It is composed of short paragraphs written by non-native speakers of English and is a part of the CoNLL2014 task, which aims to improve error correction and provides annotated corpus and automated scoring. 14 teams participated and the system with the best performance had $F_{0.5}$-score of 0.373 (Ng et al. 2014). In their study, Wu et al. (2023) randomly selected 100 sentences and evaluated three systems: ChatGPT, GECToR and Grammarly. ChatGPT performed worse than commercial products with the $F_{0.5}$-score of 0.531, compared with 0.608 for GECToR and 0.633 for Grammarly. However, ChatGPT's performance was slightly better when used on short sentences, achieving $F_{0.5}$-score of 0.6.

Fang et all. (2023) conducted a comprehensive analysis of ChatGPT's performance on multiple datasets in English, German and Chinese languages, comprising more than 10 thousand sentences. The best $F_{0.5}$-score values obtained were 0.532 for English, 0.384 for German and 0.635 for Japanese datasets.

[5] From a methodological perspective, it is not appropriate to discuss the *errors* of native speakers without proper qualification, since non-standard production does not equate to incorrect production.

As expected, ChatGPT's performance varies with different datasets and languages. To examine its effectiveness in Portuguese more thoroughly, our further research includes evaluations using essays written by high year students, which may contain more subtle and profound errors. Additionally, since ChatGPT does not always produce the same output, we intend to evaluate its performance by applying multiple grammar corrections to the same corpus and analysing their consistency.

5. Conclusion

This paper examined ChatGPT's performance and potential usage in grammatical error correction for text written in non-native Portuguese. We used our own corpus, which consisted of 69 sentences, 141 clauses and 762 words, written by third-year students of Portuguese language and literature. Its $F_{0.5}$-score reached 0.805, which can generally be considered acceptable. Besides that, the score is similar to those obtained in other authors' research, which was concisely described in the paper. However, it should be noted that the scores are not directly comparable, as studies slightly differ in methodology and the corpora analysed.

It has been suggested that ChatGPT's performance is primarily affected by unnecessary modifications. In other words, despite the correctness of a word in the source text, ChatGPT may still opt to alter it, resulting in sentences with different stylistic nuances or even unintended interpretations. The majority of these errors stem from the system's lack of training on the specific interlanguage dataset and the resulting deficiency in context comprehension. The texts written by Croatian undergraduate students highlight the complexity of knowledge possessed by a human (e.g., a teacher) who proved to be more adept at correcting informants' texts, not merely on a surface level but also proficiently identifying covert errors (grammatically correct sentences containing unintended forms or meanings). The failure to detect an error, referred to as a False Negative, is a less frequent type of the system's mistake.

By analysing the system's behaviour in our study and other relevant research, we have identified its significant educational potential, since it can be used as a reliable writing assistant with an advantage of having a highly interactive interface. Therefore, we have outlined directions for further research, which include evaluating ChatGPT with different corpora written by non-native speakers of Portuguese and assessing the effectiveness of different prompts applied to the same corpus.

Acknowledgement: We thank Maria José Homem for assistance in interpreting numerous examples.

References

Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877–1901.

Bryant, C., Zheng, Y., Muhammad Reza Qorib, Cao, H., Hwee Tou Ng, & Briscoe, T. (2023). Grammatical Error Correction: A Survey of the State of the Art. *Computational Linguistics*, 1–59. https://doi.org/10.1162/coli_a_00478

Corder, S. P. (1967) The Significance of Learners' Errors. International Review of Applied Linguistics in Language Teaching, 5, 161–170. (reprinted in Corder S. P. (1981). Error Analysis and Interlanguage. Oxford University Press).

Feng, H.-H., Saricaoglu, A., & Chukharev-Hudilainen, E. (2016). Automated Error Detection for Developing Grammar Proficiency of ESL Learners. CALICO Journal, 33(1), 49–70. https://www.jstor.org/stable/calicojournal.33.1.49

Fang, T., Yang, S., Lan, K., Wong, D. F., Hu, J., Chao, L. S., & Zhang, Y. (2023). Is chatgpt a highly fluent grammatical error correction system? a comprehensive evaluation. arXiv preprint arXiv:2304.01746.

Gass, S. M., & Selinker, L. (2008). Second Language Acquisition: An Introductory Course (3rd ed.). Routledge.

Li, Z., Link, S., Ma, H., Yang, H., & Hegelheimer, V. (2014). The role of automated writing evaluation holistic scores in the ESL classroom. *System*, *44*, 66–78.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014, June). The CoNLL-2014 shared task on grammatical error correction. In Proceedings of the Eighteenth Conference on Computational Natural Language Learning: Shared Task (pp. 1–14).

OpenAI. (2023). Gpt-4 technical report. https://cdn.openai.com/papers/gpt-4.pdf.

OpenAI. (2023). ChatGPT. Openai.com. https://openai.com/chatgpt

Penteado, M. C., & Perez, F. (2023). Evaluating GPT-3.5 and GPT-4 on Grammatical Error Correction for Brazilian Portuguese. arXiv preprint arXiv:2306.15788.

Priberam. Ferramentas para a Língua Portuguesa. FLiP. (2022, November 8). https://www.flip.pt/

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. (2020). Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research, 21, 1–67. https://arxiv.org/pdf/1910.10683v4.pdf

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

Waisberg, E., Ong, J., Masalkhi, M., Kamran, S. A., Zaman, N., Sarker, P., Lee, A. G., & Alireza Tavakkoli. (2023). GPT-4: a new era of artificial intelligence in medicine. https://doi.org/10.1007/s11845-023-03377-8

Wu, H., Wang, W., Wan, Y., Jiao, W., & Lyu, M. (2023). ChatGPT or grammarly? evaluating ChatGPT on grammatical error correction benchmark. arXiv preprint arXiv:2303.13648.

PROCJENA KVALITETE ISPRAVLJANJA GRAMATIČKIH POGREŠAKA UTEMELJENOG NA UMJETNOJ INTELIGENCIJI NA PORTUGALSKOM

Korekcija gramatičkih pogrešaka danas je integrirana u najčešće korištene alate za obradu teksta i dostupna je putem interneta. Međutim, ti su alati uglavnom poluautomatizirani jer samo predlažu moguće ispravke i varijacije te zahtijevaju interakciju s korisnikom, što može biti zamorno, osobito pri obradi duljih tekstova. Nedavni napredak u području umjetne inteligencije i obrade prirodnog jezika nudi učinkovitije strategije. Ovaj rad analizira mogućnost korištenja ChatGPT-a za ispravljanje gramatike u portugalskim tekstovima koje su napisali izvorni govornici hrvatskog jezika. Tekstove su ispravljali izvorni govornik europskog portugalskog i ChatGPT. Autori su analizirali detekciju i ispravljanje pogrešaka na različitim jezičnim razinama te ih popratili primjerima. Zbog neravnoteže u razredima, učinkovitost sustava procijenjena je pomoću F-mjere. Izračun lažno pozitivnih i istinski negativnih rezultata prilagođen je zbog posebnih slučajeva nepravilnih ispravaka. Uzimajući to u obzir, F0.5 rezultat iznosio je 0,805. Ipak, treba napomenuti da bi se za ulazni korpus s drugačijom strukturom i razinom jezične kompetencije mogli dobiti drugačiji rezultati.

Ključne riječi: ispravak gramatike, ChatGPT, evaluacija, učinkovitost sustava

Authors' addresses:

Vedran Juričić
Faculty of Humanities and Social Sciences
HR – 10 000 Zagreb, Ivana Lučića 3
vjuricic@m.ffzg.hr

Daliborka Sarić
Faculty of Humanities and Social Sciences
HR – 10 000 Zagreb, Ivana Lučića 3
dsaric@ffzg.unizg.hr