Can We Detect ChatGPT-generated Texts in Czech and Slovak Languages?

Petr Šigut and Tomáš Foltýnek

Faculty of Informatics, Masaryk University,
Brno, Czech Republic
514530@mail.muni.cz, foltynek@fi.muni.cz

Abstract. The wide availability of generative AI exacerbates existing threats to society. It would not be easy even for linguists to tell whether the text we are reading was generated by a Large Language Model (LLM) or written by a human.[1]. Researchers have started developing tools that detect AI-generated content [2]. This paper tested how two of these tools, Compilatio [3] and GPT-2 Output Detector [4], performed with Czech, Slovak and English texts. There was only one tool somewhat capable of detecting AI-generated texts: Compilatio. Other tools were designed to work only with English texts. Hence, we also tested whether automatically translating the Czech and Slovak texts to English before uploading them to the detectors would have given any promising results. Ultimately, we showed that the texts generated by ChatGPT4 were less detectable than the texts generated by ChatGPT3.5.

Keywords: ChatGPT, AI-detection, Czech, Slovak

1 Introduction

The launch of ChatGPT in November 2022 impacted many areas of human activity. For example, universities have been concerned with detecting unauthorised content generation [5], researchers are worried about the influx of AI-generated papers and the impact of the use of AI in the medical field [6], others are worried by the rise of AI-generated fake news. These threats create the need for reliable AI detection tools. This is particularly relevant regarding texts generated by AI in languages other than English, as most existing tools are trained to work with English texts primarily.

Several studies compare the performance of AI detection tools [9,8,7]. The study by Chaka [7] performed a test with generated documents from ChatGPT, YouChat and Chatsonic, which were subsequently translated into German, French, Southern Sotho, isiZulu and Spanish. Their test consisted of five AI content detectors: Open AI Text Classifier, Writer, GPTZero, Copyleaks and Giant Language Model Test Room. Still, their findings showcased that only one tool (Copyleaks) could detect some AI-generated documents in German, French and Spanish languages. Overall, the paper concluded that none of the tools are fully ready to detect AI-generated texts in different languages accurately.

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2023, pp. 35–43, 2023.

Considering the above-mentioned results, this paper seeks to examine whether the state of AI detection has changed, particularly when dealing with languages other than English. It builds upon previous research conducted by Weber-Wulff et al.[9], who tested the performance of 14 tools for detecting AI-generated text in English. This paper examines how accurately the detectors can recognise AI-generated text in Czech and Slovak languages compared to English. It tests two of the best currently publicly available tools: Compilatio and GPT-2 Output detector. The paper focuses on both the texts in their original languages (Czech and Slovak) and investigates whether their subsequent translation to English might deliver different results. Lastly, the paper also examines the extent to which ChatGPT4 and ChatGPT3.5 versions differ in the detectability of the content they produce.

2 Methodology

2.1 Selection of suitable AI-detectors

Firstly, it was necessary to select the most suitable tools with the ability to detect AI-generated content in Czech and Slovak languages. Hence, the Internet was searched on the 10th of October, for all publicly available tools that could detect such content. Publicly available tools found in this round of searching were, in turn, combined with the detectors that were researched in the initial study [9]. Overall, this approach yielded 22 AI detection tools: Compilatio, Duplichecker, Crossplag, GPT-2 output detector, Go winston, Gptzero.me, Zerogpt.com, Zerogpt.cc, ContentAtScale, Contentdetector, Copyleaks, Smodin.io, Plagiarismdetector, Scribbr, Undetectable.ai, Writer, CheckforAI, DetectGPT, OpenAI classifier, PlagiarismCheck, Writeful gpt detector and Sapling.ai.

To make our paper feasible and achieve meaningful results, we wanted to filter out low-quality tools. Therefore, we uploaded 4 AI-generated documents (2 in Czech, 2 in Slovak) to each tool. If the tool had correctly identified at least one, we would have included it in our paper. This approach was chosen because our main research question was detecting AI-generated text in Czech and Slovak documents. We decided to test the tools with multiple documents in both languages in case a lower accuracy of the tool could have caused the incorrect result.

The only tool that correctly recognised at least one of these texts as AI-generated was Compilatio, with a success rate of 50%. Four tools (Go winston, Copyleaks, PlagiarismCheck and Writeful gpt detector) did not recognise the document's language and thus refused to process it. Two tools (Undetectable.ai and Writer) had lagging web pages and could not provide any results. Three tools subject to testing in the previous study [9] (CheckforAI, DetectGPT, and OpenAI Classifier) were no longer operating. The rest of the tools incorrectly evaluated all four documents as human-written.

As we did not want to base our research on a single tool, we decided to include the GPT-2 Output Detector for comparison despite excluding it in the

previous step. We chose this tool because it was evaluated as the second-best publicly available tool in the initial study [9]. We did so because, according to the study, the best publicly available tool – Compilatio – was already part of our paper.

2.2 Test set

When creating the test set of the documents, we decided to unify as many parameters as possible to minimise the differences between the various tests across different languages. All human-written texts (in all languages – English, Czech and Slovak) were created by the paper's first author. They were written within an upper limit of 500 characters, and all sentences were complete. It was essential to use documents that were not publicly available on the Internet and thus could not have been a part of the training set for ChatGPT (3.5 or 4) or one of the selected detectors.

Subsequently, the documents that were generated by ChatGPT had the same prompts in English, Czech and Slovak. ChatGPT generated all documents in the same language as the given prompt.

We used an online translation tool, DeepL [10], to translate Czech and Slovak documents into English. These documents were then used to test the AI detection tools to examine the effects of translation on the detectability of AI.

Overall, we had 15 categories. Each category consisted of 9 documents, so in total, 135 documents were subject to this paper. The categories were as follows: Written by human:

- in the Czech language
- in the Slovak language
- in the English language
- in Czech language and translated to English
- in Slovak language and translated to English

AI-generated:

- in the Czech language generated by ChatGPT3.5
- in the Slovak language generated by ChatGPT3.5
- in the English language generated by ChatGPT3.5
- in the Czech language generated by ChatGPT4
- in the Slovak language generated by ChatGPT4
- in the English language generated by ChatGPT4
- in Czech language generated by ChatGPT3.5 and translated to English
- in Slovak language generated by ChatGPT3.5 and translated to English
- in Czech language generated by ChatGPT4 and translated to English
- in Slovak language generated by ChatGPT4 and translated to English

2.3 Testing

The process of testing was done during two weeks. As the landscape of generative AI is evolving quickly [5,9], the period had to be as short as possible to ensure fair conditions for all tested tools. We uploaded every document one by one from the test set to both of the detectors and processed it. Consequently, the detector's score was recorded, and to ensure the integrity of the data, a screenshot of the result was taken.

Table 1: Division of documents based on the score given by an AI-detector.

Human-written documents (NEGATIVE):		
[100 - 80%) AI	False positive	FP
[80 - 60%) AI	Partially false positive	PFP
[60 - 40%) AI	Unclear	UNC
[40 - 20%) AI	Partially true negative	PTN
[20 - 0%] AI	True negative	TN
Documents generated by AI (POSITIVE):		
[100 - 80%] AI	True positive	TP
[80 - 60%) AI	Partially true positive	PTP
[60 - 40%) AI	Unclear	UNC
[40 - 20%) AI	Partially false negative	PFN
[20 - 0%] AI	False negative	FN

Both detectors provided a score on a scale from 0 – 100%, which indicated how confident they were that the document was AI-generated. To measure how correct the results from the detectors were, we divided them into ten categories according to Table 1, taken from [9].

We tested the detectors with a document set in the appropriate language and consisted of eighteen documents: nine human-written texts + nine ChatGPT-generated documents. In one case, we compared the human-written documents to those generated by ChatGPT3.5, and in the other case, we compared the same human-written documents to those generated by ChatGPT4.

When testing the detectors with the translated texts, we used the same human-written texts in Czech and Slovak and AI-generated texts in Czech and Slovak as in the previous tests. This time, all AI-generated and human-written documents were translated to English through DeepL to be processed by GPT-2 Output Detector, which only worked with English.

2.4 Relevant metrics

We decided to use accuracy, sensitivity and specificity as the relevant metrics for the paper. In this paper, we counted accuracy as accuracy_semibin as defined by Weber-Wulff et al. [9]. Sensitivity and specificity are commonly used for evaluating the efficiency of classifying tests [11].

Accuracy shows how likely the detector is to make the correct decision. Partially true decisions reward accuracy with half a point; other results count as incorrect.

$$Accuracy = \frac{TP + TN + 0.5 * (PTP + PTN)}{\text{n of all documents}} * 100$$

Sensitivity is the probability that the detector correctly detects AI-generated content among the AI-generated documents.

$$Sensitivity_{original} = \frac{TP}{TP + FN} * 100$$

Subsequently, we made a slight change to the original formula of sensitivity. Since the results from the detectors were of ten categories, and this formula is designed for binary classification, it would not account for partial results. So, we decided to reward partially correct results with a lower weight, and in the denominator of the formula, we included the number of all AI-generated documents.

$$Sensitivity = \frac{TP + 0.5 * PTP}{\text{n of AI-generated documents}} * 100$$

Specificity is the probability that the detector correctly evaluates human-written text as human-written. High specificity minimises the portion of falsely labelled human-written texts as AI-generated.

$$Specificity_{original} = \frac{TN}{TN + FP} * 100$$

As with sensitivity, we updated the original formula of specificity. We did this to reward partially correct human classifications and include the number of all human-written documents in the denominator. Specificity for this paper was computed with this formula.

$$Specificity = \frac{TN + 0.5 * PTN}{\text{n of human-written documents}} * 100$$

3 Results

The results of this paper interestingly show that the accuracy of both tools, Compilatio and GPT-2 Output detector, significantly dropped compared to the results from the initial research [9]. English texts generated by ChatGPT4 were not recognised by either of the tools, and all passed as human-written.

The results of our testing revealed that the GPT-2 Output detector was incapable of correctly classifying Czech and Slovak documents and classified all of them as human-written. Compilatio, on the other hand, could process both Czech and Slovak documents. It was more accurate (67%) with Slovak

documents than Czech documents (56%). Furthermore, it was more likely to detect AI-generated documents while keeping its accuracy slightly lower than English texts.

In almost all categories, we could see that text generated by ChatGPT4 was less detectable. In the one case where Compilatio seemed more accurate and sensitive towards detecting Czech content from ChatGPT4, there was an insignificant difference of one correctly classified document.

Overall, the tools were unreliable in delivering correct answers and thus their judgement should be taken cautiously. Nevertheless, Compilatio performed well with the English texts, managing to have no false positives while still being able to detect ChatGPT3.5.

Table 2: Results for ChatGPT3.5.

GPT3.5	**English**		**Czech**		**Slovak**	
	Compilatio	GPT-2 O.D.	Compilatio	GPT-2 O.D.	Compilatio	GPT-2 O.D.
Specificity	100%	60%	56%	100%	61%	100%
Sensitivity	22%	28%	56%	0%	72%	0%
Accuracy	61%	45%	56%	50%	67%	50%

Table 3: Results for ChatGPT4.

GPT4	**English**		**Czech**		**Slovak**	
	Compilatio	GPT-2 O.D.	Compilatio	GPT-2 O.D.	Compilatio	GPT-2 O.D.
Specificity	100%	60%	56%	100%	61%	100%
Sensitivity	0%	0%	67%	0%	67%	0%
Accuracy	50%	32%	61%	50%	64%	50%

3.1 Effects of translation

In the following test, we sought to examine how useful the detectors were when presented with translated documents. Compilatio, when presented with translated texts in Czech, performed with higher specificity (78% compared to 56%) and a bit lower accuracy (44% compared to 56%) but a much lower sensitivity (56% compared to 11%); hence its ability to detect AI-generated text was considerably worse. The same applied to documents in the Slovak language; here, the sensitivity dropped to the bare minimum (0% for ChatGPT3.5 and 11% for ChatGPT 4); hence, the detector was nearly unable to detect AI-generated text.

GPT-2 Output Detector was a very different case; here, we had a tool that could not operate on Czech and Slovak documents and thus had zero sensitivity. How-

Table 4: Results for Compilatio.

Compilatio	**Czech Documents**				**Slovak Documents**			
	Original		Translated to EN		Original		Translated to EN	
	GPT3.5	GPT4	GPT3.5	GPT4	GPT3.5	GPT4	GPT3.5	GPT4
Specificity	56%	56%	78%	78%	61%	61%	89%	89%
Sensitivity	56%	67%	11%	22%	72%	67%	0%	11%
Accuracy	56%	61%	44%	50%	67%	64%	47%	50%

ever, when the texts were translated into English, the tool's performance dramatically increased. When compared to Compilatio, it had a little higher specificity (61% compared to 56%), lower sensitivity (33% compared to 56%) and a bit lower accuracy (47% compared to 56%) with translated Czech documents. With Slovak documents, it also performed surprisingly well; it had a decent specificity (83%), so it did not generate too many false positives and it had a notable sensitivity (44%) and accuracy (64%). Nevertheless, it was less decisive than Compilatio and more often it gave partial or unclear results. Compilatio made definitive results (TP, TN, FP, FN) in 96% cases, compared to GPT-2 Output Detector's 75

Table 5: Results for GTP-2 Output Detector.

GPT-2 O.D.	**Czech Documents**				**Slovak Documents**			
	Original		Translated to EN		Original		Translated to EN	
	GPT3.5	GPT4	GPT3.5	GPT4	GPT3.5	GPT4	GPT3.5	GPT4
Specificity	100%	100%	61%	61%	100%	100%	83%	83%
Sensitivity	0%	0%	33%	33%	0%	0%	44%	39%
Accuracy	50%	50%	47%	47%	50%	50%	64%	61%

4 Discussion

At the time of writing of this paper, there was only one publicly available AI detector that was able to detect AI-generated content in the Czech and Slovak languages: Compilatio. Its performance with Czech and Slovak documents was not much better than deciding by flipping a coin. With both languages, the detector had nearly 60% specificity, the rest were false positives.

Our findings showcased that the accuracy with English documents and ChatGPT3.5 text of both Compilatio and GPT-2 Output Detector dropped from April 2023 when we conducted the initial study [9] till the conducting of this paper in October 2023. Compilatio went from 77% to 61%, and GPT-2 Output Detector dropped from 73% to 45%. Nevertheless, with English documents, Compilatio had a specificity of 100% and therefore had no false positives.

ChatGPT4 showed that it could generate English content that was less likely to be detected. Neither of the tools detected English documents from ChatGPT4. Interestingly, both tools detected the Slovak and Czech content from ChatGPT4 just as likely as the content from ChatGPT3.5. It showed that the premium version of ChatGPT could only generate better content in English.

5 Conclusions

All in all, this paper demonstrated that the current state of detection of AI-generated content in Czech and Slovak languages does not deliver satisfying results. It is thus advisable to avoid relying solely upon the results provided by such tools. Translating the documents to English and then uploading them to the detectors allowed us to use an English-only tool, GPT-2 Output Detector and get comparable results to Compilatio. This could imply that translation tools preserve the human characteristics of human-written text. Still, further research has to be done to confirm the actual reasons for such an outcome. Ultimately, this paper demonstrated that English texts generated by ChatGPT4 are generally less detectable than those generated by ChatGPT3.5. Such an outcome hints towards the rapid progress in AI-generated content, which in many regards remains faster than any efforts and tools targeted at AI detection.

References

1. Casal, J. & Kessler, M. Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods In Applied Linguistics*. **2**, 100068 (2023)
2. Tang, R., Chuang, Y. & Hu, X. The science of detecting llm-generated texts. *ArXiv Preprint ArXiv:2303.07205*. (2023)
3. Compilatio Anti-plagiarism Software | Plagiarism Prevention and Detection, `https://www.compilatio.net/en`. Last accessed 10 Oct 2023
4. GPT-2 Output Detector, `https://openai-openai-detector.hf.space/`. Last accessed 10 Oct 2023
5. Foltýnek, T., Bjelobaba, S., Glendinning, I., Khan, Z. R., Santos, R., Pavletic, P., Kravjar, J.: ENAI Recommendations on the ethical use of Artificial Intelligence in Education. International Journal for Educational Integrity **19**(1), 12 (2023)
6. De Angelis, L., Baglivo, F., Arzilli, G., Privitera, G., Ferragina, P., Tozzi, A. & Rizzo, C. ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health. *Frontiers In Public Health*. **11** pp. 1166120 (2023)
7. Chaka, C. Detecting AI content in responses generated by ChatGPT, YouChat, and Chatsonic: The case of five AI content detection tools. *Journal Of Applied Learning And Teaching*. **6** (2023)
8. Elkhatat, A., Elsaid, K. & Almeer, S. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal For Educational Integrity*. **19**, 17 (2023)
9. Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P. & Waddington, L. Testing of detection tools for AI-generated text. *ArXiv Preprint ArXiv:2306.15666*. (2023)

10. DeepL Translate: The world's most accurate translator — deepl.com, `https://www.deepl.com/translator`. Last accessed 20 Oct 2023
11. Parikh, R., Mathai, A., Parikh, S., Sekhar, G. & Thomas, R. Understanding and using sensitivity, specificity and predictive values. *Indian Journal Of Ophthalmology*. **56**, 45 (2008)