Linguistic Diversity of AI-Generated Texts: A Cross-Linguistic Comparison of Slovenian and English

Luka Terčon
University of Ljubljana, Slovenia
luka.tercon@ff.uni-lj.si

Kaja Dobrovoljc Zor
University of Ljubljana, Slovenia
Jožef Stefan Institute, Ljubljana, Slovenia
kaja.dobrovoljc@ff.uni-lj.si

Relevant UniDive working groups: WG4, WG3

1 Introduction

In recent years, texts produced by generative AI tools have established a large presence in various professional fields (Gasparini et al., 2025; Kasneci et al., 2023). Prior research has found that large language models (LLMs) generate text with a unique linguistic structure that is often different from the linguistic makeup of texts written by humans (Terčon and Dobrovoljc, 2025b; André et al., 2023; Herbold et al., 2023; Liu et al., 2023; Muñoz-Ortiz et al., 2024). A key question connected to this topic relates to the level of linguistic diversity within AI-generated texts (AIGT) compared to human-written texts (HWT). Several studies have noted that AIGT contains a lower linguistic richness compared to HWT and that this discrepancy may affect human linguistic expression in the future (Guo et al., 2025; Jon and Bojar, 2026). It is therefore becoming integral to study the extent to which LLMs produce less diverse text compared to human authors.

Several previous studies have already addressed this topic, often reporting that the level of lexical diversity is generally lower in AIGT compared to HWT (André et al., 2023; Liu et al., 2023; Muñoz-Ortiz et al., 2024; Liao et al., 2023—refer to Terčon and Dobrovoljc, 2025b for a broader overview) with a similar trend holding true also for syntactic and semantic diversity (Guo et al., 2025). While the extant research has most often been conducted for AIGT generated in English, there is a significant lack of research that looks at AIGT in other less-studied languages. Given this, the present article extends this line of research to Slovenian and English, examining whether the same patterns of linguistic diversity arise in both languages.

In what follows, Section 2 introduces the methods employed to conduct the experiments, Section 3 presents the experiment results, and Section 4 concludes the article.

2 Methodology

The central research question addressed in our study is **what is the level of linguistic diversity in AI-generated essays compared to human-written essays, and to what extent does this pattern hold cross-linguistically?**

The human-written portion of the data used for our analysis consists of a sample of essays from the Šolar corpus of Slovenian student essays (Arhar Holdt et al., 2022) and a sample of essays from the LOCNESS corpus of English native speaker student essays (Granger, 1998).[1] While the two corpora were published more than 20 years apart, both were produced in similar classroom settings, which is why we consider them as comparable sources of native speaker essay compositions. In order to ensure that the essays in the final selection comprise a comparable level of linguistic proficiency, only the subset of fourth year secondary school student essays was taken from the Šolar corpus and only the subset of essays composed by British A level students was taken from LOCNESS. The texts were taken from each corpus in their original transcribed forms, without applying any additional text normalization steps such as spelling and punctuation error correction. This was done in order to preserve as much information about the source text as possible, as even seemingly superficial phenomena such as spelling errors and punctuation distribution can reflect important tendencies in the text.

[1]The corpus data was provided by the Centre for English Corpus inguistics (CECL), University of Louvain, Belgium.

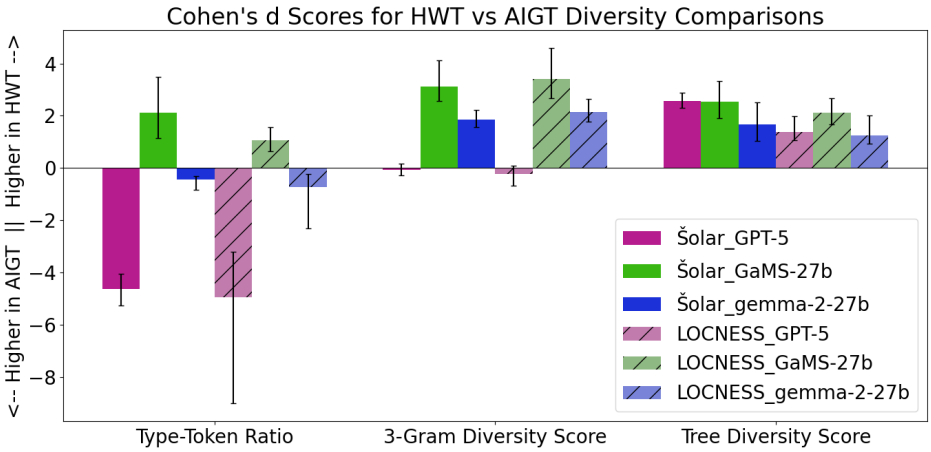

Figure 1: Cohen's d scores for each comparison. Positive values mean the diversity score is higher in HWT, while negative values mean the score is higher in AIGT. Labels starting with *Šolar_* and *LOCNESS_* refer to groups of essays generated by each model in Slovenian and English respectively. The error bars show 95% confidence intervals obtained through bootstrapping with n=10 000 resamples. After a Mann Whitney U test all comparisons except *Šolar_GPT-5* for the 3-gram diversity score returned a statistically significant value ($p < 0.05$).

The AI-generated portion of the data was generated using three different LLMs—GPT-5 (OpenAI, n.d.),[2] gemma-2-27b (Gemma Team, 2024), and GaMS-27b (Vreš et al., n.d.)—where GaMS-27b represents a language-specific model based on the gemma-2-27b base model trained on a large amount of high-quality Slovenian data. As a starting point, we used a simple default prompt instructing the model to produce an essay with the same topic and length, approximating typical real-world usage. Examples are given in Appendix A.[3]

For each group of essays generated by each model in each language, we calculate three related linguistic diversity measures using the CompareTree tool (Terčon and Dobrovoljc, 2025a) and compare the results against the corresponding group of human-written essays. The tool computes the segmental type-token ratio, a measure used for assessing the level of lexical diversity calculated by first splitting the data into fixed-length segments of 1,000 words and then calculating the ratio between the number of unique lemmas and the total number of lemmas in the segment. The tool also supports calculation of two related measures for assessing the level of n-gram diversity and syntactic diversity. These are referred to as the segmental n-gram diversity score and the segmental tree diversity score and are obtained in the same way as the segmental type-token ratio, but this time taking n-gram[4] and syntactic tree[5] frequencies as the basis instead of word lemmas. The degree of difference between each AIGT and HWT group is then assessed using Cohen's *d* effect size.

3 Results

Figure 1 shows the results in a unified bar plot, where positive values indicate higher linguistic diversity in HWT, and negative values indicate higher diversity in AIGT. A full overview of numerical values is given in Appendix B.

Looking at the results, we observe a remarkable level of **cross-linguistic consistency**: Cohen's d values point in the same direction across all models and metrics in both languages, suggesting that the observed diversity patterns are robust for the combination of English and Slovenian. At the same time, the results reveal meaningful variation depending on both the model and the type of diversity examined, discussed below.

With respect to **lexical diversity**, differences between models are most pronounced. Diversity is highest for GPT-5, followed by gemma-2-27b, in line with the known tendency for lexical diversity

[2]The exact model snapshot used bears the name *gpt-5-2025-08-07* in the OpenAI API.

[3]The data used to conduct our study will be made available at the following address: https://www.clarin.si/repository/xmlui/handle/11356/2210

[4]We find that similar results emerge when testing for different values of n from n=3 to n=6. Here we report the values for n=3, hence the measure is referred to as the segmental 3-gram diversity score

[5]We rely on dependency relation (sub)trees extracted using the STARK tool (Krsnik and Dobrovoljc, 2025).

to increase with model size (Guo et al., 2025). In contrast, GaMS-27b shows lower lexical diversity than gemma-2-27b despite having the same number of parameters, suggesting that its Slovenian-specific fine-tuning may have restricted its lexical repertoire.

With respect to **n-gram diversity**, the results show a largely consistent pattern for GaMS-27b and gemma-2-27b, where human-written texts are more diverse than AI-generated texts, suggesting a higher degree of formulaicity in these models. GPT-5, however, shows comparable or slightly higher diversity in AIGT, likely reflecting its higher lexical diversity in general.

With respect to **syntactic diversity**, the pattern is highly consistent: all models in both languages produce less syntactically diverse text than human writers, suggesting that reduced structural variety may be a general property of LLM-generated text.

4 Conclusion

Our experiments show a clear gap in linguistic diversity between LLM-generated and human-written texts, with diversity generally lower in AI-generated texts, except at the lexical level, which appears to be model-sensitive. Importantly, these patterns are consistent across both Slovenian and English, suggesting that the observed impact on linguistic diversity might generalize across some languages. Future studies should inspect whether these patterns reveal a broader trend that holds true for other less-studied languages as well.

Our ongoing work will extend this analysis to other genres and examine the impact of different generation settings and prompt design.

Acknowledgements

This work was supported by the Slovenian Research and Innovation Agency through the Young researchers program, the SPOT project (A Treebank-Driven Approach to the Study of Spoken Slovenian, Z6-4617), the Language Resources and Technologies for Slovene research program (P6-0411), the PoVeJMo research program (Adaptive Natural Language Processing with the Help of Large Language Models), and the Large Language Models for Digital Humanities (LLM4DH, GC-0002) research program. We also thank Nives Hüll for the manual identification of titles and other metadata in the Šolar corpus.

References

Christopher MJ André, Helene FL Eriksen, Emil J Jakobsen, Luca CB Mingolla, and Nicolai B Thomsen. 2023. Detecting ai authorship: Analyzing descriptive features for ai detection. In *Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023)*.

Špela Arhar Holdt, Tadeja Rozman, Mojca Stritar Kučuk, Simon Krek, Irena Krapš Vodopivec, Marko Stabej, Eva Pori, Teja Goli, Polona Lavrič, Cyprian Laskowski, Polonca Kocjančič, Bojan Klemenc, Luka Krsnik, and Iztok Kosem. 2022. Developmental corpus šolar 3.0. Slovenian language resource repository CLARIN.SI.

Loretta Gasparini, Nitya Phillipson, Daniel Capurro, Revital Rosenberg, Jim Buttery, Jayne Howley, Sarath Ranganathan, Catherine Quinlan, Niloufer Selvadurai, Michael Wildenauer, and et al. 2025. A survey of large language model use in a hospital, research, and teaching campus. *Data & Policy*, 7:e78.

Gemma Team. 2024. Gemma. `https://www.kaggle.com/m/3301`.

Sylviane Granger. 1998. The computer learner corpus: a versatile new source of data for SLA research. In Sylviane Granger, editor, *Learner English on computer*, pages 3–18. Routledge.

Yanzhu Guo, Guokan Shang, and Chloé Clavel. 2025. Benchmarking linguistic diversity of large language models. *Transactions of the Association for Computational Linguistics*, 13:1507–1526.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific Reports*, 13(1):18617.

Josef Jon and Ondřej Bojar. 2026. Thesis proposal: Are we losing textual diversity to natural language processing? In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 188–206, Rabat, Morocco. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. 2023. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Luka Krsnik and Kaja Dobrovoljc. 2025. STARK: A toolkit for dependency (sub)tree extraction and analysis. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 44–51, Ljubljana, Slovenia. Association for Computational Linguistics.

Wenxiong Liao, Zhengliang Liu, Haixing Dai, Shaochen Xu, Zihao Wu, Yiyang Zhang, Xiaoke Huang, Dajiang Zhu, Hongmin Cai, Quanzheng Li, Tianming Liu, and Xiang Li. 2023. Differentiating chatgpt-generated and human-written medical texts: Quantitative study. *JMIR Med Educ*, 9:e48904.

Yikang Liu, Ziyin Zhang, Wanyang Zhang, Shisen Yue, Xiaojing Zhao, Xinyuan Cheng, Yiwen Zhang, and Hai Hu. 2023. Argugpt: evaluating, understanding and identifying argumentative essays generated by gpt models.

Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting linguistic patterns in human and llm-generated news text. *Artificial Intelligence Review*, 57(10):265.

OpenAI. n.d. Introducing GPT-5. https://openai.com/index/introducing-gpt-5. Accessed: 2026-04-02.

Luka Terčon and Kaja Dobrovoljc. 2025a. ComparaTree: A multi-level comparative treebank analysis tool. In *Proceedings of the 23rd International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2025)*, pages 129–139, Ljubljana, Slovenia. Association for Computational Linguistics.

Luka Terčon and Kaja Dobrovoljc. 2025b. Linguistic characteristics of ai-generated text: A survey. *arXiv preprint arXiv:2510.05136*.

Domen Vreš, Iztok Lebar Bajec, Tjaša Arčon, Gašper Jelovčan, and Marko Robnik-Šikonja. n.d. GaMS-27B-Instruct. https://huggingface.co/cjvt/GaMS-27B-Instruct. Accessed: 2026-04-02.

A Text Generation Prompts

The AI-generated essays are generated using a special type of prompt in each language. These prompts include different degrees of essay metadata, depending on the availability of this metadata in the two source human-written corpora. Examples are provided below:

A.1 Slovenian Prompt

```
Napiši esej z naslovom {essay_title} in
    podnaslovom {essay_subtitle}, ki se
    nanaša na literarno delo {
    referenced_literary_work}, s približ
    no {length_of_corresponding_HWT}
    besedami. Odgovori samo z esejem
    brez spremnega besedila.
```

A.2 English Prompt

```
Write an essay using approximately {
    length_of_corresponding_HWT} words
    addressing the following topic: {
    topic_of_corresponding_HWT}
Provide only the essay without any
    additional accompanying text.
```

B Numerical Overview of Results

Table 1 presents the Cohen's d effect size values for all AIGT vs HWT comparisons.

Language	Essay Group	TTR	3GD	TDS
Slovenian	GPT-5	**-4.62**	-0.06	**2.57**
	GaMS-27b	**2.11**	**3.13**	**2.55**
	gemma-2-27b	**-0.43**	**1.85**	**1.66**
English	GPT-5	**-4.94**	**-0.24**	**1.37**
	GaMS-27b	**1.07**	**3.4**	**2.11**
	gemma-2-27b	**-0.72**	**2.13**	**1.24**

Table 1: Overview for the Cohen's d effect size scores showing AIGT vs HWT comparisons of diversity scores for various essay groups. The score for each essay group represents the Cohen's d effect size calculation for the comparison between the diversity score of a group of essays generated by a specific LLM in one of the two included languages and the diversity score of the corresponding group of human-written essays. The diversity score acronyms are as follows: TTR – Type-Token Ratio, 3GD – 3-gram diversity score, TDS – Tree Diversity Score. Bolded values represent comparisons for which a Mann-Whitney U test returns a statistically significant value ($p < 0.05$).