Analele Universității de Vest din Timișoara. Seria Științe Filologice
2024, Nr. 62/1, pp. 77-83
https://doi.org/10.35923/AUTFil.62-1.06

AN EXPERIMENTAL STUDY OF LEXICAL COMPLEXITY IN LITERARY TEXT ANALYSIS: CHATGPT'S PERFORMANCE IN ROMANIAN VS ENGLISH

Roxana ROGOBETE*

Mădălina CHITEZ**

*, **West University of Timisoara
*roxana.rogobete@e-uvt.ro
https://orcid.org/0000-0002-4155-3918
**madalina.chitez@e-uvt.ro
https://orcid.org/0000-0001-9005-3429

An experimental study of lexical complexity in literary text analysis: ChatGPT's performance in Romanian vs English

This study examines how ChatGPT generates literary analyses, with a particular focus on lexical complexity, coherence, and overall textual quality in Romanian and English. The research builds on previous studies that examine "natural" academic writing produced by students (see Tucan et al. 2020, Crașovan & Rogobete 2020) and aims to assess the extent to which ChatGPT aligns with established academic standards in literary criticism by applying a comparable set of assessment parameters to determine whether its lexical and structural choices align with academic conventions. We designed a three-stage experimental framework. In the first stage, ChatGPT was prompted to generate literary analyses in response to a given detailed task. In the second stage, we provided the model with a relevant academic article on the topic and then requested a revised response, allowing us to evaluate the extent to which its output could integrate discipline-specific discourse. The last stage involved a role-based prompt and information about the targeted audience. Our findings suggest that while ChatGPT can provide structured literary analyses, its responses often lack originality, fluency, and a complex understanding of figurative language – elements that are elementary in human-authored analyses. We evaluated the linguistic complexity of the selected Romanian datasets using the LEMI readability platform (Chitez *et al.*, 2024), and the English ones using ARI (Automated Readability Index) from https://readabilityformulas.com/, which provided a comparative assessment of lexical density, syntactic variation, and textual readability across both output sets. This study seeks to contribute to the broader discussion on AI-generated academic writing and the limitations of language models in academic writing and literary analysis.

KEYWORDS:

use of ChatGPT; LEMI readability formula for Romanian; academic writing; AI-generated texts; literary analysis.

The use of artificial intelligence (AI) in educational settings and, particularly, in the field of academic writing, has become an increasingly relevant topic in recent years. Language models such as OpenAI's ChatGPT, Anthropic's Claude, META AI's LLaMA or Google DeepMind's Gemini have demonstrated great potential in generating structured and coherent texts, assisting with tasks such as summarization, paraphrasing, and also argumentative writing (Yang *et al.* 2023; Gao *et al.*, 2023; Su, Lin, & Lai, 2023). However, their integration in academic contexts also raises concerns regarding ethics, originality, critical thinking, and the potential for dependency on algorithmically generated content (Floridi & Chiriatti 2020).

One of the dominant areas where AI-generated writing is being evaluated is in academic essay production, particularly in disciplines requiring complex textual analysis, like Humanities. Recent studies have shown that ChatGPT can produce essays that comply with basic academic conventions, yet often lack higher-order analytical reasoning and the ability to engage deeply with abstract concepts (Perkins, 2023; Rudolph *et al.*, 2023). The purpose of the present study is not to promote or normalize the illicit use of generative artificial intelligence (GenAI) tools in educational environments. Instead, it aims to examine the degree of linguistic and rhetorical competence displayed by these tools in response to several prompts. As many academics have underlined (Ripoll *et al.*, 2025; Bašić *et al.*, 2023; Floridi & Chiriatti, 2020; Stokel-Walker, 2023; Rudolph *et al.*, 2023), the pedagogical and epistemological issues raised by AI-generated writing need to be addressed critically. Therefore, our study adheres to an ethical framework that aims to analyze how language models perform across linguistic systems and genres, as well as the reasons behind some discrepancies.

We designed a three-stage framework to analyze the results generated by ChatGPT, based on six Romanian and six English literary texts. The Romanian texts delivered to ChatGPT span the 19th and 21st centuries and include works by I. L. Caragiale, Mihai Eminescu, Carolina Vozian, Cătălina Stanislav, and Andrei Dósa. The English-language texts encompass a range of modern and contemporary poetry, featuring authors such as William Carlos Williams (*Metric Figure*), Allen Ginsberg (*A Supermarket in California*), Wallace Stevens (*Of Modern Poetry*), Emily Dickinson (*Death*), Wilfred Owen (*Dulce et decorum est*), and Robert Frost (*Wild Grapes*). We provided ChatGPT with each poem separately and gave it three different prompts:

Stage 1	Independent generation, detailed task provided
Stage 2	Supported generation: Detailed task and useful resource provided in order to be cited and incorporated (citation)
Stage 3	Contextual, supported generation: Target audience (university professor) and role (student) provided in order to refine stage 2

In the first stage, ChatGPT was given the literary text and a detailed prompt, in order to describe, analyze and interpret the given text. The instrument was instructed to consider a specific assessment rubric, that included the following: (1) description of the text and its particular features; (2) analysis of specific elements, such as construction techniques, use of figurative language, and the configuration of the text's semantic coherence; (3) well-argued interpretation, based on the preceding analyses; and (4) overall coherence, correctness, and the use of appropriate theoretical language. For the second stage, we provided ChatGPT with an additional valid resource: a pre-existing analysis of the text – usually available online or extracted from printed resources. This source was to be used as a reference, asking ChatGPT to integrate the specific scholarly perspectives into its response and to include the source properly in the reference list. The revised response allowed us to evaluate the extent to which ChatGPT's output could integrate discipline-specific discourse. The third and final stage involved a role-based prompt and information about the targeted audience: ChatGPT was instructed to assume the role of a university student addressing an academic audience of university professors.

Although comparisons between human and artificial writing may appear overfamiliar or even outdated in current AI scholarship (Perkins, 2023; Yang *et al.*, 2023), the present study reframes this debate by shifting the focus toward cross-linguistic performance. The central comparison is not between human and machine production *per se*, but between the outputs generated in two distinct linguistic contexts – Romanian and English – under identical constraints.

In the first stage of the experiment, the Romanian outputs demonstrate formal compliance to the task's structural requirements, but still reflect superficial analysis. The responses focus on the text's formal organization, identifying its structural components and noting patterns of symmetry, opposition, or contrast. The model does not follow a deep analysis, and all answers use the same bullet-point structure. According to the LEMI readability formula developed for Romanian (see Chitez *et al.*, 2023; Chitez *et al.*, 2024), the texts generated by ChatGPT were assessed as suitable for readers above the 12th-grade level, having a high amount of complex words (>50%). The responses proved to have syntactic sophistication (average of 20 words per sentence); however, the lexical complexity did not show a corresponding increase. In other words, while the texts were long, morpho-syntactically complex, the vocabulary remained relatively repetitive, schematic across analyses. The lexical diversity across these outputs was approximately 57%, indicating that while sentence structures reached higher-level standards, the range of distinct lexical items was still limited. This suggests that the model is able to produce formally complex texts without necessarily enhancing lexical richness in Romanian. The English texts generated by ChatGPT follow a similar structural pattern to the Romanian ones and also exceed the 12th-grade level, as indicated by an average ARI index of 16. In terms of lexical diversity, the English texts are comparable to the Romanian outputs (54%), with an average of 24 words per sentence. However, the results show a relatively higher presence of academically oriented vocabulary, with coverage of the Academic Word List (AWL) at 23%. Thus, the English texts integrate formal academic vocabulary slightly better than the Romanian ones.

In stage 2, we observed differences between the outputs in Romanian and English after ChatGPT was provided with an additional source to embed into its response. In the Romanian texts, there was a slight decrease in lexical diversity from stage 1 (54%), indicating an even more limited use of vocabulary. In addition, the model's referencing style was limited and mostly incorrect in Romanian: the model paraphrased the additional source but did it in a shallow manner: it omitted any precise bibliographic reference. In most instances, it only referred to the title of the additional source, and did not mention important elements of academic citation such as authorship, date of publication, or page numbers. This action resulted in a summary that was concise, but incomplete and lacked the rigor expected of academic writing. The English outputs demonstrated a small increase in lexical diversity from stage 1 (+5%) and displayed stronger referencing strategies, as the model employed direct quotations more frequently than paraphrase, while usually noting the publishing date in parentheses. Although still imperfect, this emphasizes a greater ability to synthesize information and to acknowledge references in English. This second step shows that, although ChatGPT can use source material in both languages, the English outputs tend to better balance lexical diversity and accurate referencing, whereas the Romanian outputs show tendencies toward simplification and minimal, often inaccurate citation. For instance, the model often provides a references list:

> **Works Cited**
>
> Mane, Dr. Prabhanjan. "Allen Ginsberg's Vision of America in *A Supermarket in California*." *Literary Endeavour*, vol. X, no. 2, Apr. 2019, pp. 1–5.

In the case of the Romanian texts, the behavior described by Rudolph *et al.* (2023) as "bullshit spewer" becomes evident in the model's handling of sources and references. ChatGPT offers a synthesis that is incomplete and occasionally misleading. For instance, one of the responses included "complex language and rich in symbols (CUM NEGUSTORII DIN CONS…)", while the article provided as additional resource did not include this statement. The model generates a very short commentary that mimics scholarly analysis, but hallucinates and incorporates inaccuracies or unsupported claims.

In stage 3, we optimized the guided prompt, which was complemented with information related to the role and the audience. We did not offer other external constraints, such as character/word limit and kept the need for referencing. Quantitative analysis indicates that the English content now covers approximately 33% of the Academic Writing List (results provided by Text Inspector), an increase of 10% in comparison with stage 2. Lexical diversity had increased to 81%, reflecting a broader range of word choices, while lexical density reached 50%. According to Text Inspector, lexical sophistication, when compared to the BNC corpus, achieves 25%, showing a substantial presence of less frequent words. Readability analysis indicates an Automated Readability Index (ARI) of 14, which corresponds to a level suitable for upper undergraduate or early graduate readers. The table below provides three excerpts from the English results: the first output remains mainly descriptive, the second one simulates a more analytical approach, without textual evidence, while the third one engages with contextual references ("Beat poetics"), although concise and perhaps limited in terms of analysis:

Output 1	Allen Ginsberg's *A Supermarket in California* is a free-verse poem that juxtaposes a mundane setting with profound existential and cultural questions. The poem is characterized by its conversational tone, vivid imagery, and its invocation of Walt Whitman, a 19th-century poet renowned for his democratic ideals and celebration of individuality. The poem meanders through the speaker's thoughts as he imagines encountering Whitman in a neon-lit supermarket, blending the real with the surreal.
Output 2	Allen Ginsberg's *A Supermarket in California* is a seminal work from his 1956 anthology *Howl and Other Poems*. This free-verse poem is set in a surreal, dreamlike supermarket and features a direct engagement with the ghost of Walt Whitman. Through vivid imagery and an introspective tone, Ginsberg reflects on America's shift from Whitman's idealized vision of unity, abundance, and

	democracy to a modern consumerist society. The poem functions as both homage and critique, using Whitman as a symbol of the poet's longing for a more authentic, spiritual connection in a world increasingly dominated by materialism and alienation
Output 3	Written in 1955, *"A Supermarket in California"* reflects Allen Ginsberg's Beat poetics—confessional, expansive, and intertextually layered. The poem opens with a direct apostrophe to Walt Whitman, whom Ginsberg addresses as both poetic ancestor and ghostly interlocutor during a nocturnal walk through a modern supermarket.

According to LEMI readability scores, the Romanian texts exceed once again the typical complexity expected of 12th-grade materials, having an average lexical diversity of 63% (+9% in comparison with stage 2). However, references are mostly subject to hallucinations: ChatGPT invented bibliographic entries, going beyond the references provided in stage 2: „Lavinia Stan (2022) argumentează că „literatura confessională contemporană românească se dezvoltă în zona hibridă dintre realism și poetică digitală, explorând modul în care emoțiile sunt mediate de tehnologie și memorie" (Stan, 2022, p. 134)."; „Simona Popa (2021) argumentează că „literatura feministă românească recentă utilizează poemul ca spațiu de contestare a normelor și constrângerilor istorice, politice și sociale, transformând corpul în teren al experienței critice" (Popa, 2021, p. 89)."; „Nicolae Manolescu interpretează Telegrame ca pe „o prefigurare a literaturii limbajului" și notează că „absența oricărui punct de vedere subiectiv echivalează cu o poetică a neutralității ironice" (*Arca lui Noe*, 1980, p. 384)."; „Criticul Florin Băluț subliniază că „poezia contemporană românească tinde să confrunte lumea virtuală cu cea reală, de multe ori printr-un eu poetic fragmentar, cu trăiri contradictorii și hiper-analitice" (Băluț, 2018, p. 76).".

To evaluate the semantic coherence of the texts, we used SpaCy, a natural language processing library in Python, within Google Colab. The program processed the texts and extracted semantic similarity scores (between adjacent sentences), which are shown in the following table. An interesting fact is that, for Romanian texts, the scores show little variation, suggesting that, regardless of the quality of the prompt, ChatGPT is quite rigid and offers limited room for improvement, at least for now. In contrast, we identified small variations for the English model:

Output	**Semantic coherence**
RO-1	0,40
RO-2	0,40
RO-3	0,38
EN-1	0,41
EN-2	0,46
EN-3	0,40

Interestingly, many of the patterns identified in ChatGPT's outputs – such as formulaic language, repetitive vocabulary, and schematic structure – mirror those found in human-authored responses (see Tucan *et al.*, 2020 for an analysis on automated use of conventional forms in university writing). These outcomes can reveal both the limitations of LLMs and a sort of "mimicry effect" (in the sense that AI reproduces human writing habits and styles). However, this creates a potential feedback loop: AI models learn in part based on human artifacts, which themselves may contain biases or stereotypical interpretations. Then, the AI could replicate these issues in new outputs, reinforcing conventional analytical approaches, amplifying errors and oversimplifying interpretation. This effect is corroborated by the fact that the model's performance in different target languages is constrained by training data and digital repositories. This describes an epistemic inequality in AI, since ChatGPT will perform differently in languages with less corpus representation. English-language users may benefit from exposure to a richer lexical model, while Romanian users are at risk of plateauing in terms of vocabulary that might be considered rather moderate in variation. Additionally, the limited resources available in Romanian bring the risk of hallucination, where we have identified a resistance from the LLMs to to incorporate additional resources. Furthermore, evaluation literacy becomes even more relevant, since users may overestimate the difficulty of AI-generated Romanian texts because syntactic sophistication can mask lexical simplicity or lack of content substance.

Our study does not encourage undeclared use of GenAI tools, but analysed the level of proficiency for these instruments, mainly ChatGPT. Moreover, the aim was not to discern human vs AI generated samples, but evaluating how generative models behave in academic writing contexts and addressing gaps in NLP research. By analyzing outputs generated from successive prompts – first using an open-ended evaluative task, secondly providing a more or less scholarly reference to be incorporated into the response, and finally offering the academic context – we explored how *prompt design* and linguistic setting interact to modelate and alter AI performance. The observed differences between Romanian and English outputs (even though applied to a limited number of texts) indicate that language models, while apparently universal, still encode structural biases linked to training data distribution and linguistic typology. Thus, the value of the comparison lies not in its replication of the human–machine dichotomy, but in exposing the *linguistic asymmetry of AI-generated competence* across languages.

As these generative writing tools become more widely accessible to the general public, they are likely to advance rapidly. Moreover, as they gain access to and integrate more sophisticated linguistic, semantic, and argumentation-generating capabilities, they are expected to have a greater influence on writing practices and education. It is to be researched in which way genAI will be responsible for "cheap production of good, semantic artefacts" (Floridi & Chiriatti, 2020), or a more balanced AI training data will bring cross-linguistic equity.

References:

Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). ChatGPT-3.5 as writing assistance in students' essays. *Humanities and social sciences communications, 10*(1), 1-5.

Chitez, M., Dascălu, M., Udrea, A. C., Strilețchi, C., Csürös, K., Rogobete, R., & Oravițan, A. (2024). Towards building the LEMI readability platform for children's literature in the Romanian language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 16450–16456). ACL Anthology. https://www.aclanthology.org/.

Chitez, M., Rogobete, R., Oravițan, A., & Csürös, K. (2023). Developing LEMI: A new corpus-based literacy support tool for schoolchildren. In *CALL for all Languages. EUROCALL 2023 Short Papers, 15–18 August 2023, University of Iceland, Reykjavik* (pp. 153–158). Editorial Universitat Politècnica de València. http://ocs.editorial.upv.es/index.php/EuroCALL/EuroCALL2023/paper/viewFile/16966/8304.

Floridi, L., Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds &*

Machines, *30*, 681-694. https://doi.org/10.1007/s11023-020-09548-1.

Gao, M., Ruan, J., Sun, R., Yin, X., Yang, S., & Wan, X. (2023). Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Perkins, M. (2023). Academic Integrity considerations of AI Large Language Models in the post-pandemic era: ChatGPT and beyond. *Journal of University Teaching and Learning Practice*, *20*(2), 1-24.

Ripoll Y Schmitz, L.M., Sonnleitner, P. (2025). Evaluating AI-generated vs. human-written reading comprehension passages: an expert SWOT analysis and comparative study for an educational large-scale assessment. *Large-scale Assess Educ* **13**, 20. https://doi.org/10.1186/s40536-025-00255-w.

Rudolph, J., Tan, S., & Tan, S. (2023). ChatGPT: Bullshit spewer or the end of traditional assessments in higher education? *Journal of Applied Learning and Teaching*, *6*(1). https://doi.org/10.37074/ jalt.2023.6.1.9.

Stokel-Walker C. (2023). ChatGPT listed as author on research papers: many scientists disapprove. *Nature*, *613*(7945), 620-621. doi: 10.1038/d41586-023-00107-z. PMID: 36653617.

Su, Y., Lin, Y., & Lai, C. (2023). Collaborating with ChatGPT in argumentative writing classrooms. *Assessing Writing*, *57*, 100752.

Tucan, D., Rogobete, R., Chitez, M., Radu-Pop, A.-M. (2020). Cât de pregătiți sunt elevii de liceu pentru scrierea academică de nivel universitar? Studiu didactic contrastiv bazat pe date de corpus lingvistic. *Annals of the West University of Timisoara. Humanities Series*, *58,* 69-92. https://analefilologie.uvt.ro/wp-content/uploads/2022/01/D-Tucan_R-Rogobete_M-Chitez_AM-R_Pop_Anale-Litere-2020.pdf.

Yang, X., Li, Y., Zhang, X., Chen, H., & Cheng, W. (2023). Exploring the limits of ChatGPT for query or aspect-based text summarization. *arXiv preprint arXiv:2302.08081*.