(*) Corresponding author
Peer review method: Double-blind
Original research article
DOI: 10.20544/HORIZONS.1.1.24.P07
Section A

A COMPARATIVE ANALYSIS OF FIRST-PERSON PRONOUN USE IN AI AND HUMAN ESSAYS

Elena Shalevska*
University "St. Kliment Ohridski" – Bitola
ORCID ID. https://orcid.org/0000-0002-3270-7137
Email: elena.shalevska@uklo.edu.mk

Abstract

This study comparatively analyzed the use of the personal pronoun "I" in AI-generated and human-written essays through an AntConc analysis of two self-made corpora of texts, to (dis)prove the hypothesis: AI-generated essays will use the first-person pronoun "I" significantly less frequently than human-written essays. The first corpus of essays in this study, all generated by ChatGPT, consisted of 100 essays totaling 29,442 words. The second corpus of 100 essays, written by humans, included texts with a total word count of 39,878. The frequency analysis revealed a significant disparity in the use of the pronoun "I" between the two corpora. In the AI-generated essays, "I" was used only 15 times, contrasting sharply with the human-written essays, where "I" appeared a total of 523 times. These findings showed the noteworthy differences in the narrative styles of AI and human writers, showing the potential limitations of free Generative AI models in replicating human-like self-reference and subjectivity in written discourse.

Keywords: Artificial Intelligence, ChatGPT, Academic Writing, Corpus Linguistics, Discourse analysis.

INTRODUCTION

In recent years, the development and use of artificial intelligence (AI) in generating written content has skyrocketed. And not just in the academic community, but worldwide, in general. Generative AI (GenAI) models, like OpenAI's ChatGPT, have been at the forefront of this AI-revolution.

GenAI is a branch of AI designed to create new content based on human instructions, utilizing the training data it has been exposed to. These models are specifically trained and programmed to generate original content—whether it be text, images, music, video, or other forms—that did not exist before (Singh 2023).

GenAI models nowadays are used in various areas, including education, content creation, and even automated customer service. They have been rapidly advancing and producing more and more content that is sometimes hard to distinguish from human-produces one. However, despite all of the advancements GenAI and natural language processing (NLP), questions remain about the extent to which AI can replicate human-like writing.

AI writing has been extensively studied in the past two to three years (Dillon et al. 2022; Fitria 2023; Mahama et al. 2023; Herbold et al. 2023; Shalevska 2023; Shalevska 2024 etc.) yet, one aspect the AI-produced texts – the use of the first-person pronoun "I" – has been under-researched.

The use of "I" in writing is more than just a grammatical choice; it reflects the writer's involvement, perspective, and subjectivity. In academic discourse, the use of "I" can be contentious, as it introduces a personal element that some scholars argue is at odds with the objective tone of scholarly writing. Others, however, assert that the strategic use of first-person pronouns can enhance clarity and reader engagement (Thonney 2013; Webb 2024).

This study aims to explore the differences in the use of the first-person pronoun "I" between AI-generated and human-written essays. By analyzing two corpora—one composed of essays generated by ChatGPT-3.5 and the other by human writers—this study seeks to shed light on how AI models approach personal engagement in text and how this compares to human writing.

Broader Context

The role of the first-person pronoun in academic writing has been the subject of significant research. The use of the pronoun is linked with self-references which can serve multiple purposes, including structuring the text, guiding the reader, and acknowledging funding sources (Chavez Munoz 2013).

In terms of self-reference and author visibility, Hyland (2002) asserts that the use of the first-person pronoun can increase the visibility of the author, thereby strengthening the argument being presented. Hyland also identifies different functions of personal pronouns in academic texts, such as: stating a purpose, explaining a procedure, stating results/claims, expressing self-benefits, elaborating an argument and giving acknowledgements.

Regardless of its function, Chung and Pennebaker (2007) argue that the use of "I" often reflects personal involvement and emotional intensity. Their research shows that this pronoun is frequently used in intimate and personal texts. Popescu (2007), similarly to Tang and John (1999), notes that in student essays, the first-person pronoun "I" is among the most frequently used words. Tang and John (1999) also find that the use of first-person pronouns can serve multiple functions, including expressing personal involvement, structuring arguments, and clarifying the writer's stance. Their study suggests that while first-person pronouns are often discouraged in academic writing, they can be strategically employed to enhance clarity and reader engagement.

Tang and John (1999) also propose a spectrum of authorial presence conveyed by the personal pronoun “I”, ranging from minimal to maximal authorial presence. The roles within this spectrum, from least-powerful to most-powerful authorial presence are: 1. “I” as representative; 2. “I” as the guide; 3. “I” as the architect; 4.“I” as the recounter of the research processes; 5. “I” as the opinion holder and 6. “I” as the originator, with “I” referring to all forms of both the first and second person singular and plural forms.

In an interesting comparison to do with the use of “I”, Newman et al. (2008) indicate that women tend to use the first-person pronoun more often than men, suggesting possible gender differences in writing styles. Regardless of the gender, Tausczik and Pennebaker (2010) analyze the psychological significance of words, revealing that frequent use of personal pronouns like "I" is associated with more intimate and personal texts.

In a different context, Harwood (2007) finds that even politicians use first-person pronouns to make their statements sound more convincing. This widespread usage underscores the pronoun's importance in various forms of communication, including academic and political discourse.

Despite the extensive research on first-person pronoun use in human writing, there has been little to no study on how this pronoun is used in AI-generated texts. The current study fills this gap by examining the frequency and role of "I" in essays generated by an AI language model and comparing it with human-written essays.

RESEARCH METHODOLOGY

This study employs a corpus linguistics approach, using the AntConc software to analyze the frequency and distribution of the first-person pronoun "I" in AI-generated and human-written essays. A separate corpora was created for the two types of texts upon collection. A corpus can be defined as a collection of machine-readable authentic texts (including transcripts of spoken data) that is sampled to be representative of a particular natural language or language variety (McEnery et al. 2006, 5). Due to the unavailability of an existing corpus of AI-generated essays as of July 2024, this study relies on a self-compiled corpus, consisting of texts produced by ChatGPT-3.5, a freely accessible language model developed by the company OpenAI.

Population and Sample

The population for this study are academic texts produced by Generative AI models, and ones written by humans. To study the population, purposive sampling was employed and a set of two corpora of essays was created.

The AI-generated corpus consists of 100 essays generated by ChatGPT-3.5, totaling 29,442 words, while the human-written corpus also includes 100 essays with a total of 39,878 words. It is important to note that the human-written essays are all publicly available, were all written for the TOEFL exam of English proficiency, and collected and published by Dethi.com.

The AI-generated essays were produced by the author, by inputting a set of standardized prompts / essay topics into ChatGPT-3.5. All of the topics were publicly available at the website 5StarsEssays.com. The prompts i.e. topics cover a broad range of areas, ensuring a diverse sample reflective of typical academic writing. Similarly, the human-written essays were selected from a larger pool of TOEFL exam essays, ensuring they represented a range of scores and writing styles across different topics.

Following the guidelines established by McEnery, Xiao, and Tono (2006), the collected essays were then compiled into uniform digital corpora, with each text being carefully formatted to ensure consistency in the analysis. The primary focus of the analysis was on identifying and quantifying the use of the pronoun "I" in each corpus, with the results being compared to determine any significant differences.

Analysis

The analysis was conducted using AntConc's keyword and concordance tools. Keyword analysis was employed to identify the prominence of "I" in each corpus, while concordance analysis provided insights into the context in which "I" was used.

Hypothesis

The main hypothesis of this study is that: *AI-generated essays will use the first-person pronoun "I" significantly less frequently than human-written essays.* This difference is expected to reflect the model's tendency to produce more objective and impersonal text, in contrast to the subjective and personal nature of human writing.

Limitations

It’s important to acknowledge this study’s limitations, to ensure it can be redone in the future, considering all the ways in which it was limited. Firstly, the study includes a limited, self-made corpus of only 200 essays with a total of 69.320 words. Secondly, the AI corpus consists of essays written exclusively by the free version of the ChatGPT model – ChatGPT 3.5. Additionally, the human-written corpus is about 35.4% larger in word quantity than the AI one. Despite this, the study is considered to provide valuable insights into a topic that has not been empirically researched as of yet.

Ethical Considerations

The study adhered to ethical guidelines by ensuring that all data used was either publicly available or generated using ChatGPT within the bounds of OpenAI's terms of service. No personal or sensitive data was included in the analysis.

Conflict of Interest

The author hereby declares no conflict of interest.

RESULTS AND DISCUSSION

General Remarks

The analysis of the 100 AI-generated essays reveals a total of 29,442 words (referred to as "running tokens" in AntConc) and 4,126 distinct words in the first corpus. In contrast, the 100 human-written essays in the second corpus, while longer with a total of 39,878 words, contain only 3,877 distinct words. Despite being longer, human-written essays have a lower number of unique words compared to AI-generated essays. This suggests that AI-generated texts may use a more diverse vocabulary.

To further illustrate this difference, a type-token ratio (TTR) was calculated for each of the two corpora. The TTR1 for the AI-generated essays is 14.00%, calculated by dividing the number of distinct words by the total word count [(4,126 / 29,442) * 100]. The human-written essays have a TTR2 of 9.71% [(3,877 / 39,878) * 100]. The percentage difference between these TTRs is approximately 30.64%, which shows a notable disparity.

This indicates that, although human-written essays are approximately 35% longer in total word count, they exhibit a significantly lower type-token ratio compared to AI-generated essays. This may suggest that human writers tend to focus on lenght at the expense of vocabulary variety.

The use of the personal pronoun “I” in AI-generated essays

The AntConc analysis for the keyword “I” reveals that the first-person singular pronoun in this corpus is almost entirely nonexistent. In fact, the only instances of “I” occur in so-called narrative essays, and even there, the pronoun is used very sparingly—just 15 times in a corpus of 29,442 words. This results in a frequency of use of about 0.051%.

When analyzing the frequency list, generated by AntConc, it’s interesting to note that the personal pronoun “I” does not even appear among the top 100 most frequent words in the corpus.

It is important to note that "I" appears in the corpus 13 more times, but not as a pronoun but as the number 1, represented by a Roman letter, in the phrase "World War I".

The use of the personal pronoun "I" in human-written essays

The obtained results from the second corpus contrast sharply with the results for AI-authored essays. In this corpus, the pronoun appears in every essay without exception, sometimes only once, but often – multiple times per essay. Specifically, in the 39,878-word corpus, the pronoun "I" is used 523 times, which means the word has a frequency of about 1.311%.

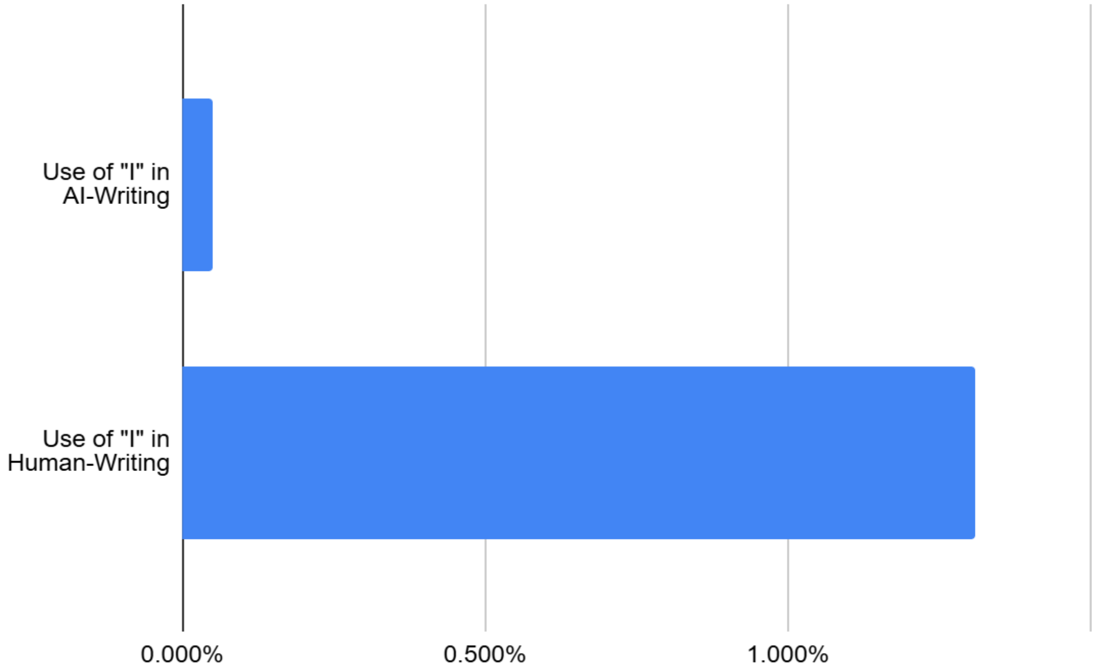
In this corpus, the pronoun in question ranks as the 9th most frequent word.

DISCUSSION

These findings suggest that AI-generated texts are more objective and neutral, focusing on presenting information rather than conveying personal opinions, attitudes, and experiences. Human-written texts, on the other hand, are "littered" with instances of "I". This higher usage of the pronoun reflects the subjectivity inherent in human writing.

The marked difference in the use of "I" not only indicates a lack of subjectivity in AI-generated texts but can also serve as a tool for distinguishing between human-authored and machine-generated content. Simply put, the more "I"s one can see, the more likely the text was written by an actual human.

According to these findings, human writers are 25.7 times more likely to use the pronoun "I" compared to Generative AI models. Humans use the pronoun frequently, which is in line with previous findings (Tang & John 1999; Chung and Pennebaker 2007 etc.) This also aligns with Popescu's assertion (2007, 184) that the pronoun "I" is among the most frequently used words in her corpus of student essays. To better illustrate the frequency difference, the results are also presented in Graph 1 below:



Graph 1: Use of the first-person pronoun in AI and Human Essays

CONCLUSION

This study reveals a distinct divergence in the use of the first-person singular pronoun "I" between human-written and AI-generated essays. The AntConc analysis of the two corpora, shows that AI-generated texts demonstrate a marked absence of the pronoun "I," with a frequency of only 0.051%. Unlike them, humans seem to be using "I" much more frequently as their essays show a significantly higher frequency of the pronoun at 1.311%, ranking it as the 9th most frequent word in the corpus.

This pronounced difference in pronoun usage not only highlights the inherent subjectivity in human writing but also provides a clear metric for distinguishing between human and machine-generated content.

Future research could, of course, expand on this and investigate the usage of the pronoun "I" in AI-generated texts across various other free and subscription-based AI models to further enhance our understanding of AI-generated written discourse and the ways it is (still) limited.

BIBLIOGRAPHY

1. Chavez Munoz, M. 2013. "The 'I' in Interaction: Authorial Presence in Academic Writing." *Revista de Lingüística y Lenguas Aplicadas* 8. https://doi.org/10.4995/rlyla.2013.1162.
2. Dillon, S., and J. Schaffer-Goddard. 2022. "What AI Researchers Read: The Role of Literature in Artificial Intelligence Research." *Interdisciplinary Science Reviews*. https://doi.org/10.1080/03080188.2022.2079214.
3. Fitria, T. N. 2023. "Artificial Intelligence (AI) Technology in OpenAI ChatGPT Application: A Review of ChatGPT in Writing English Essay." *ELT Forum: Journal of English Language Teaching* 12(1): 44-58. https://doi.org/10.15294/elt.v12i1.64069.
4. Harwood, N. 2007. "Political Scientists on the Use of First-Person Pronouns in Academic Writing." *Journal of Academic Writing* 17(3): 245-259.
5. Herbold, S., A. Hautli-Janisz, U. Heuer, et al. 2023. "A Large-Scale Comparison of Human-Written versus ChatGPT-Generated Essays." *Scientific Reports* 13: 18617. https://doi.org/10.1038/s41598-023-45644-9.
6. Hyland, K. 2002. "Authority and Invisibility: Authorial Identity in Academic Writing." *Journal of Pragmatics* 34(8): 1091-1112. https://doi.org/10.1016/S0378-2166(02)00035-8.
7. Mahama, I., D. Baidoo-Anu, et al. 2023. "ChatGPT in Academic Writing: A Threat to Human Creativity and Academic Integrity? An Exploratory Study." *Indonesian Journal of Innovation and Applied Sciences (IJIAS)* 3(3): 228-239. https://doi.org/10.47540/ijias.v3i3.10053.
8. McEnery, A., R. Xiao, and Y. Tono. 2006. *Corpus-Based Language Studies: An Advanced Resource Book*. London: Routledge.
9. Popescu, T. 2007. "College Essay-Writing: A Corpus-Based Analysis." *Translation Studies: Retrospective and Prospective Views, 2nd Edition*. Accessed January 15, 2024. https://www.academia.edu/186628/COLLEGE_ESSAY_WRITING_A_CORPUS_BASED_ANALYSIS.
10. Shalevska, E. 2023. "AI Language Models, Standardized Tests, and Academic Integrity: A Chat (GPT)." *International Journal of Education Teacher* 26: 17-25.
11. Shalevska, E. 2024. "The Digital Laureate: Examining AI-Generated Poetry." *RATE Issues*. https://rate.org.ro/media/blogs/b/shalevska.pdf?mtime=1708596426 (August 10, 2024)

12. Singh, K. 2023. *Principles of Generative AI: A Technical Introduction*. Carnegie Mellon University. https://www.cmu.edu/intelligentbusiness/expertise/genai-principles.pdf. (July 12, 2024)
13. Tang, R., and S. John. 1999. "The 'I' in Identity: Exploring Writer Identity in Student Academic Writing through the First-Person Pronoun." *English for Specific Purposes* 18(1): 23-39. https://doi.org/10.1016/S0889-4906(97)00025-6.
14. Thonney, T. 2013. "Teaching the Conventions of Academic Discourse." *Across the Disciplines* 10(1). Retrieved from WAC Clearinghouse.
15. Webb, E. 2024. "The Power of First-Person Voice in Scientific Writing." *Proceedings of the National Academy of Sciences*.