



Ca' Foscari
University
of Venice

Master's Degree Programme in Language Sciences

Final Thesis

A Comparative and Diachronic Study of Linguistic Features in Human-Written and LLM-Generated Theses

Following the Linguistic Traces of Generative AI in Ca' Foscari University
Master's Theses

Supervisor

Ch. Prof. Gianluca Lebani

Co-supervisor

Dr. Claudia Cusano

Degree Candidate

Sevilnur Bahadır
907390

Academic Year

2025/2026

Acknowledgments

I firstly would like to thank my supervisor, Gianluca Lebani, who helped me considerably at all stages of my thesis, and was here to share his ideas and feedback whenever I needed. I truly appreciate his help throughout the whole thesis, and his reassurances every time that I experienced challenges. I also would like to thank Claudia Cusano for her kindness and her continuous support on my thesis, even during the busiest periods of her academic career.

And to my dear family, who made it possible for me to fulfill my dream of continuing with a master's degree in Linguistics. I thank them with all my heart for always appreciating how passionate I am about the things I want, and making me feel that I am cared for very deeply. Without them, I could have never arrived this point in my academic life.

I would like to give the most special thanks to my sister Selvihan, who is my favorite person on this earth, and is the best sister someone could ever ask for. Thank you for being with me in my darkest and brightest days, thank you for being right by my side at every milestone, and thank you for showing me how it feels to love and to be loved by someone so deeply. And another special place goes to my beautiful niece İdil, who came into this world only a year ago, yet filled my heart and my world with wonderful feelings in such a short time, and being the joy of my life with her beautiful smile. The thought of seeing you at the end of these difficult days gave me all the strength that I needed, and I know that your existence will be my motivator to always do better.

I also would like to thank my best friend Hilal. Our friendship started in our childhood, and it will last a lifetime. We shared such long days of hard work, where we were almost always at the brink of losing all hope and motivation. But, you always cheered me up and made me believe in myself. I know that you will be a great lawyer, and there will be a huge smile on our

faces when we remember these difficult days full of study sessions and videocalls. Know that my heart is with you no matter how far we are.

And my dear Elena, every time I felt weak and close to giving up, you showed up like an angel and supported me. I can never forget the times you wished me luck right before all my exams, always keeping in mind a lot of details about me, and being with me during all the difficulties I endured here. Thank you for reviving my hope that there are still genuinely kind, loving and thoughtful people on this world.

Table of Contents

1. Introduction	1
2. Literature Review	4
2.1. How Discernible are AI-Generated Texts from Human-Written Texts?	4
2.2. The Most Prominent Characteristics of AI-Written Text	7
2.3. LLM-Generated Texts and the Influence of LLMs on Human-Written Texts in the Academic Sphere	13
2.4. The Literature on the Methodology and Linguistic Features Used for Comparing Human-Written and Synthetic Texts	18
2.5. The Description of the Research Gap and the Importance of the Current Study	21
3. Methodology	22
3.1. Data Collection	23
3.1.1. Human-Written Theses	23
3.1.2. AI-Generated Theses	26
3.2. Corpus	29
3.3. Linguistic Features	30
3.3.1. Lexical Features	30
3.3.2. Syntactic Features	32
3.3.3. Readability Metrics	33
3.3.4. Stylistic Features	34
3.4. The Implementation of the Linguistic Features	36
3.5. The Quantitative Analysis of Variables	37
3.6. Statistical Analysis	37
4. Results	39
4.1. Lexical Features	40
4.1.1. Average Word Length	40
4.1.2. TTR	43
4.1.3. Hapax Legomena Ratio	45
4.1.4. Rare Word Ratio	47
4.1.5. Lexical Density	49
4.1.6. Adjective, Adverb, Noun and Verb Ratio	52
4.1.7. Part-of-Speech Diversity	61
4.1.8. LLM-Associated Word Frequency	62
4.2. Syntactic Features	66
4.2.1. Sentence Counts	68
4.2.2. Average Sentence Length	70
4.2.3. Average Dependency Length	72
4.2.4. Average Tree Depth	74
4.2.5. Average Branching Factor	76
4.3. Readability Features	78

4.3.1.	Flesch Reading Ease	78
4.3.2.	Gunning Fog Index	81
4.4.	Stylistic Features	83
4.4.1.	Stop Word Ratio	84
4.4.2.	Capitalized Word Ratio	86
4.4.3.	Punctuation Count	88
4.4.4.	Hedge Word Ratio	90
4.4.5.	Passive Voice Ratio	92
4.4.6.	Biber Features	97
5.	Discussion	105
5.1.	Lexical Features	105
5.2.	Syntactic Features	112
5.3.	Readability Metrics	114
5.4.	Stylistic Features	116
6.	Conclusion	120
6.1.	Limitations	122
6.2.	Practical Implications of the Current Thesis	123
References		125
Appendix		134

Abstract

The advent of Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI) has enabled students to receive easily accessible assistance in completing challenging tasks. Although diverse types of academic texts were analyzed in terms of LLM-associated linguistic features, there is no previous study focusing on comparing human- and AI-written master's theses. Thus, the first aim of this thesis is to analyze the linguistic features of master's theses written in Ca' Foscari University of Venice between 2019 and 2024 to make a diachronic comparison between the ones published before and after the launch of ChatGPT in November 2022. For this comparison, the years were categorized into pre-ChatGPT (2019-2022) and ChatGPT eras (2023-2024), and 161 theses from Economics and Finance, Environmental Sciences and Language Sciences areas were compiled from the archive accordingly. The second aim of the thesis is to make a systematic comparison between human-written and AI-generated theses. For this reason, 161 theses were generated through gpt-5-mini-2025-08-07 in August 2026, each of which thematically matches the archived theses. The final corpus included 161 human-authored theses and 161 AI-generated counterparts. The linguistic and statistical analyses yielded considerable differences in terms of average word length, proportion of content words, stop words, hedge words, adjectives, passive voice, readability, and Biber features. The findings for these measures indicate a distinction between human-written master's theses in the pre-ChatGPT and ChatGPT era, and between LLM-generated and human-authored master's theses. Finally, these specific features in ChatGPT era human-written theses were shown to have closer values to LLM-generated theses, which, by itself, is still not sufficient evidence of the use of AI in human-written master's theses.

Keywords: AI-generated, LLM-generated, human-written, linguistic features, master's theses, pre-ChatGPT, ChatGPT

1. Introduction

The advent of Large Language Models (LLMs) and Generative Artificial Intelligence (GenAI), which can provide answers to questions that are addressed in the form of prompts, has enabled people from a variety of fields to find relief in completing challenging tasks. Nowadays, they have become almost indispensable in people's lives since they contain a great amount of data and can generate complex outputs in a short span of time, surpassing humans in terms of efficiency (Lin & Liu, 2026). However, along with their ever-growing popularity due to their efficiency, comes a serious concern: LLMs have been given such a great quantity of human data from diverse sources (e.g. Internet, books, articles; Zindela, 2023) that they are able to write linguistic content and interact in a similar way to humans (Singh et al., 2024; Steiss et al., 2024, cited in Zhang & Crosthwaite, 2026). As indicated by Geng and Poibeau (2025), traces of AI-generated texts are attested in many areas such as student essays, speaking content and online texts, which shows that LLM-generated texts are becoming more and more interwoven into human-authored texts.

Terčon & Dobrovoljc (2025) maintained that AI- and human-written content are getting so blended that it could lead to the questioning of whether there is still any genuine human-authored content. According to Geng and Poibeau (2025), the blend between the AI- and human-authored texts has reached the extent, where, through the frequent use of GenAI tools, human authors could become influenced by the language used by LLM-generated texts, starting to adapt the models' linguistic style (De Cesare, 2026). Besides, they may want to adjust their writing style to avoid the LLM detection tools, which may lead them to write texts that could gradually mingle with AI-generated ones. This study assumes that LLMs could be used even more frequently as assistants for tasks like grammar correction, paraphrasing, literature

summarization, writing abstracts (Lin & Liu, 2026), and general post-editing (Altmäe et al., 2023; Gao et al., 2023; Lozić & Štular, 2023, cited in Bao et al., 2025), which makes the tracing of linguistic characteristics and differences between AI- and human-authored texts even more pressing.

When it comes to academia, the increasing use and integration of LLMs into texts becomes even more alarming, because it could decrease the originality of ideas, argument structures, and word choices that are peculiar to each discipline, which implies the possible loss of variety in writing styles (Bao et al., 2025). It is widely known that university students are gradually relying more on ChatGPT while performing tasks such as academic essays, assignments and online exams (Zindela, 2023; Guo et al., 2025). Furthermore, it was revealed that between 6.5% and 16.9% of the peer reviews which were submitted to the AI conferences ICLR 2024, NeurIPS 2023, CoRL 2023 and EMNLP 2023 may have been altered considerably via LLMs (Liang et al. 2024). A gradual increase in the percentage of LLM usage in academic texts that may continue to be observed in the future poses a threat to authorship attributions (Bao et al., 2025) and the credibility and the quality of scholarly work. Therefore, the issue of the detectability of AI-generated work from human-written work becomes especially important for academic honesty (Pan, Oliveira & Mello, 2025).

Considering the steps on the path of becoming a scholar, the initial works that are submitted are bachelor's and master's degree theses. These primary academic texts have been subject to plagiarism checks via the application of online tools, among which the most prominent one is Turnitin¹ (Batane, 2010). However, after the introduction of GenAI tools, the issue of academic integrity has become more complicated in the sense where academic writers were able

¹ <https://www.turnitin.com/>

to use LLM outputs without being detected for plagiarism, considering that making use of LLM-generated responses in academic texts were not easily detectable, and were not like using another scholar's work without giving credit.

Nowadays, as the natural outcome of the quick development and adoption of GenAI tools, texts that are generated by AI are getting gradually easier to detect via computational tools such as OpenAI detector, ChatGPT detector, Academic detector (cf. Guo et al., 2025), though not all of them work with 100% success (Liu et al. 2023, cited in Zindela, 2023). It was also argued that the efficiency of AI-detection tools is still debated, because they have been shown to return false positives and especially lower performance in detecting AI-written parts in texts that were written with AI-human collaboration (Pan, Oliveira & Mello, 2025). Besides, Guo et al. (2025) mentioned that the distinguishability of such texts by current detectors decrease even more when the amount of information provided by humans in the prompts (i.e. human involvement) increases. The existing detectors tend to classify such texts as human-written, although their features clearly differ from a real human-authored text (Guo et al., 2025).

There also exist applications that sidestep AI-detectors, which are called “AI humanizers” (Giray, 2026). These tools may alter the content and hide the algorithmic origin of a text, as a result of which that text may pass the AI detection checks (Giray, 2026). Due to such tools, the theses written by university students may be accepted, despite the possible existence of AI-generated pieces of texts in those works. Therefore, this thesis assumes that theses submitted after the introduction of ChatGPT, which was in November 2022², might have been written with the help of ChatGPT and other GenAI chatbots without being noticed by detectors. The aim of the current thesis is not to detect such theses; but to test and find the linguistic indicators of

² <https://openai.com/index/chatgpt/>

possible LLM usage in human-written master's theses, demonstrate the year-by-year change in the proportion of the occurrence of LLM-associated indicators, and discover the linguistic characteristics of AI-generated master's theses.

2. Literature Review

2.1.How Discernible are AI-Generated Texts from Human-Written Texts?

It has been found that humans can become able to differentiate between AI-generated and human-written content and identify a number of features that are related to machine-generated pieces of texts after being exposed to and familiarizing themselves with such content, though this ability is still not very comprehensive (Zindela, 2023). For that reason, there are plenty of studies that intend to address the issue of the distinguishability of LLM-generated texts from human-authored texts. For instance, Reinhart et al. (2025) carried out a comparative study to evaluate the rhetorical styles that are employed by LLMs, by using the Biber tag set provided by Biber (1995; 1988; 2019). They utilized OpenAI's GPT-4o and GPT-4o Mini, and four variants of Meta Llama 3 to generate data that are comparable to a human-written corpus of English. As a result, the researchers discovered that LLMs are unable to reach the stylistic variation that is attested in human-authored texts, showing a preference limited to certain grammatical structures (e.g. participial clauses, nominalization, "that" clauses as subjects; Reinhart et al., 2025, p.3) and vocabulary (such as "tapestry", "intricate", "amidst"; p.4). The research by Reinhart et al. (2025), therefore, has shown that although LLMs appear to have human-like language ability, they still have limitations as to the stylistic, grammatical and lexical choices they make depending on different genres and contextual factors, unlike human authors. The lack of variety in AI-generated texts that has been demonstrated by Reinhart et al. (2025) was further confirmed by the results of Zeleke, Soni and Manikonda (2025), where the researchers also found that

human annotators converged upon the observation that AI-generated texts have more structured responses while human-authored texts feature variations in responses and response structures based on topic. Therefore, it could be implied that there are still evident stylistic, lexical and grammatical differences between human-written and AI-generated texts, which requires a detailed, systematic analysis.

Muñoz-Ortiz et al. (2024) made a linguistic comparison between human-written and AI-generated news texts, where they made use of six different LLMs belonging to three different LLM families. They found that human-written news texts exhibit a higher range in sentence lengths and higher sentence diversity while AI-generated news texts show sentences with less variation and more uniform lengths (Muñoz-Ortiz et al., 2024). Furthermore, they observed that LLMs do not reflect the area-specific vocabulary that is required of journalistic texts, thereby having more homogenous outputs that do not have the stylistic elements that are peculiar to different types of texts. The calculation of standardized type-token ratio and measure of textual lexical diversity (MTLD) features further displayed that human authors employ richer vocabulary choices and higher lexical diversity scores than LLMs. Regarding part-of-speech and the use of punctuation, the researchers found that human-written news texts included a specific preference for using nouns, adjectives and punctuation symbols, while LLMs were shown to include more symbols and numbers, which was interpreted as the models' attempt to appear more credible. Moreover, as to dependency types, LLMs have been shown to feature more numeric modifier, aux, copula and nominal subject dependencies, which aligns with the part-of-speech frequency of each of these elements in LLM-generated news texts. And again, aligning with the lower frequency of adjectives in AI-generated texts, there were fewer subtrees with adjective modifiers. Finally, the constituent analysis that they further conducted exhibited that

human-written news texts consist of more noun, adjective and prepositional phrases, whereas LLM-generated texts have more auxiliary verbs and verb phrases. Overall, the results by Muñoz-Ortiz et al. (2024) demonstrated that the linguistic features of LLM texts differ significantly from human-authored writing, adding to the distinguishability of these two text sources.

Pan, Oliveira and Mello (2025) conducted a study on the stylistic distinguishability of AI- and human-written academic texts. For their study, they instructed GPT-4 to edit and rewrite student-written essays, as a result of which they achieved texts that are written with AI-human collaboration. The researchers used the BERT model for a classification test to see whether AI-human collaboration can be recognized by machine learning. They found that the words in LLM-generated texts are rarer, longer and more complex, which also decrease the readability ease (further confirmed by Zou et al. 2026). In addition, the BERT model that they used for the classification task showed great success, returning only one false positive result. These findings by Pan, Oliveira and Mello (2025) clearly demonstrate that the writing styles of GPT-4 and students were distinct enough to be recognized by machine learning methods and the analysis of stylistic features, thereby touching upon how useful these techniques could be especially on AI-human collaborated text detection.

Another important research was carried out by Zhang and Crosthwaite (2026) on student-written and AI-generated argumentative essays. The researchers conducted a frequency analysis, collocation analysis and keywords analysis on their corpus. The results of their frequency analysis demonstrated that while the essays written by students included vocabulary that are related to more personal, familiar and human-centered topics (e.g. “sleep”, “therapy”, “children”, “human”, p.12), ChatGPT-generated essays involved more frequent employment of vocabulary that concentrates on more societal and technological themes (e.g. “challenges”, “impact”,

“energy”, p.13). The collocation analysis showed that student writers favor simpler and shorter collocations whereas LLM-generated essays employ longer and more complex collocations, likely originating from the great amount of data that they are trained with, which provide them with a plethora of high frequency collocations attested in professional academic writing (Zhang & Crosthwaite, 2026; Wu, 2025). Furthermore, the comparative keywords analysis revealed that vocabulary such as “youtube”, “autism”, “children” and “employees” were more commonly associated with the student-written texts, which denotes a distinction between AI- and human-generated content, the latter involving words that are linked to personal experiences, authentic context and emotions (Zhang & Crosthwaite, 2026, p.18). Their results confirm Ma et al. (2023), who found that LLM-generated texts do not have the emotional and stylistic depth that characterizes human writing, which originate from personal experiences (cited in Culda et al., 2025).

The results of Zhang and Crosthwaite (2026) are especially valuable because the aforementioned differences in lexical choices between human- and AI-generated argumentative essays were observed even though the LLM was prompted to write on the exact topic of the student-written essay. The preference of ChatGPT over words that are closely related to broader, more global and abstract themes (e.g. “sustainable”, “vehicles”, “renewable”, p. 19) even when it is prompted to write on the same topics with students demonstrate that LLMs make different lexical and thematic choices that could be distinguished from their human-written counterparts.

2.2. The Most Prominent Characteristics of AI-Written Text

Several studies have focused on the specification of the differences between AI-generated and human-written texts. For instance, it has been found that LLM-generated texts feature less diverse vocabulary, more syntactically complex structures, and more nominalization (André et

al., 2023; Herbold et al., 2023; Liu et al. 2023; Muñoz-Ortiz et al., 2024; Reviriego et al., 2023, cited in Terčon & Dobrovoljc, 2025 and Culda et al., 2025), which are in line with the results of Reinhart et al. (2025). In their paper, Terčon and Dobrovoljc (2025) did a survey of the methodology and characteristics of AI-generated texts that were the focus of the prominent studies in the literature. They found that rarely used vocabulary such as “knowledgeable” and “stand out feature”, modal verb “can”, the conjunction “and”, the prepositions “by”, “with”, and “for”, and the personal pronoun “their” are more attested in AI-generated texts than in human-written ones (Desaire et al., 2023; Kim and Desaire, 2024; Mitrovic et al., 2023; Fujiwara, 2024, cited in Terčon & Dobrovoljc, 2025, p. 9). In addition, it has also been found that LLM-generated outputs feature more repetitions than human-authored texts (Simon et al., 2023; Yanagita et al., 2024, as cited in Terčon & Dobrovoljc, 2025). Another linguistic aspect that was considered in the comparison of LLM-generated and human-authored texts is the word length, which led to more differing results than the other lexical criteria. While André et al. (2023) indicated that AI-generated texts have longer words than human-written ones, Yildiz Durak et al. (2025) concluded that there was no significant difference in terms of word length between the two types of source text (cited in Terčon & Dobrovoljc, 2025). It has also been reported that LLMs use fewer idioms (Simon et al., 2023), include words that do not exist (Yanagita et al., 2024), and mostly utilize commas and periods instead of a more diverse application of punctuation marks, unlike human-written texts (as cited in Terčon & Dobrovoljc, 2025).

According to Terčon and Dobrovoljc (2025), in terms of syntactic complexity measures, studies in the literature mostly focus on calculating the dependency distance, whose higher values denotes a higher syntactic complexity. Four studies out of five that were examined in their paper reported AI-generated texts to have higher syntactic complexity scores (Culda et al., 2025;

Herbold et al., 2023; Liu et al., 2023; Nkhobo and Chaka, 2023). Finally, as to discourse markers and style, the authors demonstrated that AI-generated texts feature less discourse items, which are used repetitively, and that the style of these texts are more formal and impersonal than human-written ones. A clear indicator to this in the LLM-generated texts was the fewer number of adjectives and adverbs that are attested, because these items would suggest a more personal involvement of the author in the text, the lack of which leads to a more impersonal tone (Terčon & Dobrovoljc, 2025). Markowitz et al. (2024), on the other hand, found that texts generated by GPT-3.5 were less readable, had more emotional items, and included more adjectives (as cited in Toney et al., 2026). The results of Markowitz et al. (2024), when compared with the outcomes of the aforementioned studies which found fewer use of adjectives and adverbs, could be originating from the specific type of texts. Since Markowitz et al. (2024) focused on generating positive hotel reviews, the language model might have overcompensated the use of emotionally loaded words, considering that they are computational agents that respond fully to tasks that are provided to them in the form of prompts. As De Cesare (2026) indicated in her book, AI has no intentionality and does not have sufficient knowledge as to which information to include more in detail, and which to leave out. Rather, it operates without intention, by acting solely upon the prompt that it is given, which could help explain the higher use of adjectives found in Markowitz et al. (2024) when specifically prompted to generate positive hotel reviews. The results of the comparative corpus study by Zeleke, Soni and Manikonda (2025) also support this aspect of LLMs, where they found out that there was a strong correlation between the length of AI-generated response and the question length, while there is no correlation between the human response length and the question length. The researchers interpreted this difference between the correlations as the more systematic and thorough processing done by the LLM, which is trained

not to miss any included element given in the prompt, while humans may choose to concentrate on one aspect of the given input and neglect providing explanations on the other components.

Another research by Culda et al. (2025) investigated and compared the characteristics of human-written and machine-generated text. They compiled the human-written part of their corpus from three sources which are Wikipedia article intros, blog articles and scientific abstracts of arXiv articles. They collected AI-generated introduction counterparts via three different LLMs, which are GPT-Neo 1.3B, Qwen2.5-1.5B-Instruct and BloomZ-560M. In their corpus, consisting of human-written and AI-generated texts, they analyzed lexical, syntactic, readability, stylistic and sentiment measures. The researchers found that LLM-generated texts show lower lexical diversity, evidenced by fewer unique words and lower TTRs, while human-written texts have larger word counts and a higher syntactic diversity (Culda et al. 2025). Besides, they argued that the most significant difference observed in terms of lexical features between human-written and AI-generated texts were in the genre of abstracts. The researchers indicate that human-written abstracts include more scientific and specialized vocabulary and higher word lengths, which is not the case for machine-generated abstracts, suggesting that LLMs are still not fully capable of imitating the technical writing style that is typical of abstracts (Culda et al., 2025).

The machine-generated texts in Culda et al. (2025) were also shown to have higher readability scores with shorter sentences and simpler clauses, which denotes simpler sentence structures and reduced syntactic depth (Leong, 2023; Liu et al., 2023, cited in Culda et al. 2025). In addition, the researchers observed that AI-generated texts have higher clarity and consistence, showing a narrower and clustered distribution with fewer extreme values, denoting a more uniform style, possibly due to their constant dependence on repetitive patterns and the reuse of common phrases. LLMs' overuse of common, formulaic syntactic patterns, and their tendency

towards a uniformity in style via reduced employment of personalized elements was also demonstrated in Wu (2025) and Zou et al. (2026), which suggest the lack of authorial voice and awareness of interdisciplinary variations in texts.

Amirjalili et al. (2024) conducted a study where they asked GPT-4 to generate seven sections of an essay by providing the model with a sample human-written text, and they found that the human-written essay included more hedges and boosters and was more personalized than its AI-generated counterpart. Furthermore, the small grammatical errors that were present in the human-written essay added to its humanness, while the AI-generated version had more factual errors, which resulted in AI texts' being more easily classified as synthetic (Toney et al., 2026). In addition, Shalevska (2024) also observed that human-written texts feature a larger range of hedging devices and boosters.

Furthermore, Fröhling and Zubiaga (2021) classified the most common characteristics of AI-generated texts into four groups, which are “lack of syntactic and lexical diversity”, “repetitiveness”, “lack of coherence” and “lack of purpose” (cited in Hashemi, Shi & Corriveau, 2025, p. 3761). Another seminal study conducted on the human-authored and LLM-generated text comparison that reproduces the results achieved by Fröhling and Zubiaga (2021) is the one by Guo et al. (2023), where they constructed a comparison corpus that composes of questions and answers provided by humans and ChatGPT. The researchers asked human annotators to summarize the features they observed in human and ChatGPT-generated answers, as a result of which certain patterns and features of ChatGPT and human-written texts were listed. Putting aside the organizational and validity-based considerations of the outputs, one core linguistic difference observed was the lack of emotion and the preference of conjunctions and adverbs over punctuation marks in ChatGPT-generated answers, unlike the human counterparts, in which

emotions were conveyed explicitly through the use of punctuation and grammar features (Guo et al. 2023). ChatGPT's higher use of conjunctions was interpreted as its effort to render the output more logical than the answers provided by humans. Furthermore, by calculating the average response length, vocabulary size and density scores, Guo et al. (2023) found that human answers are shorter than the LLM answers, but on the other hand they have a larger vocabulary size, which means that humans make use of more different words, while ChatGPT is more limited to certain vocabulary. These findings were further confirmed by Agrahari, Bisht and Sanasam (2024) and Zou et al. (2026), who also found that LLM-generated texts are more structured, and exhibit less lexical diversity than human-written texts, in spite of having a higher frequency of unique words. Another type of linguistic measure considered in Guo et al. (2023) was the distribution of part-of-speech. The authors pointed out that ChatGPT answers comprise of a high proportion of nouns, which denote a more argumentative, informative and objective nature that is typical of academic papers and official documents.

When it comes to the sphere of academic discourse, specifically, the AI-generated texts have been shown to have less variety of narrative items and explicit references, and a more persuasive tone than the human-written counterparts (Sardinha, 2024). Interestingly, in their study, Sardinha (2024) found that even though this was the case, the AI-generated texts in the academic writing register were the most human-like, when compared with the other text types such as news and essays. These findings by Sardinha (2024) call for further attention on the analysis of AI-associated features and items that are attested in and distinguishes human-written and machine-generated academic texts, which brings us to the following overview of studies specifically conducted on LLM-generated or LLM-associated academic texts.

2.3.LLM-Generated Texts and the Influence of LLMs on Human-Written Texts in the Academic Sphere

The similarity between AI-generated and human-written data becomes more alarming in the academic sphere (Yao & Liu, 2025), given the main goal of producing academic texts, which is to introduce novel contributions to the literature by building on and giving credit to previous work. Therefore, with the outstanding pace at which the current LLMs are improving, the existing matter of plagiarism in academic texts has become even more complicated, and the educators are not fully accomplished at identifying the AI-generated language yet (Singh et al., 2024; Nowacki & Wrochna, 2025). The studies performed by Ippolito et al. (2020) and Clark et al. (2021) showed that humans experienced challenges in distinguishing between texts that were written by human authors and generated by GPT-2 and GPT-3, which are among the earliest models introduced by OpenAI (cited in Geng & Poibeu, 2025).

In addition to the aspect of academic integrity, there is also the questionable quality of AI-generated texts, which have been noted to have less creativity and a lower quality (Sarkar, 2025; as cited in Geng & Poibeu, 2025). The quality also comes into question due to the way in which LLMs work. As the input provided to the LLMs consists predominantly of scholarly literature, the increasing use of GenAI in such work could lead to the LLM input becoming increasingly composed of LLM-assisted texts, which in turn could have a considerable influence on the linguistic choices of human authors in the upcoming years, with the possibility of a gradually declining textual quality (Gray, 2024). Although the current study will be focusing more on the differences between the linguistic forms attested in AI- and human-written texts instead of evaluating the content-wise quality and validity of AI-generated texts, the quality

concerns of such texts should not go unnoticed (Tudino & Qin, 2024), especially in the academic sphere.

Liang et al. (2024) aimed to detect the rising usage of AI-tools in the peer reviews by using a simple approach, which is to extract words and structures that were highly associated with AI-generated texts from their target corpus. Their study has shown that GenAI tools such as ChatGPT and others are employed in the generation of peer reviews of the conference papers written in the scope of the area of artificial intelligence, which was identified due to the ten to thirty times higher usage of certain common adjectives as of 2023, after the launch of GPT-3.5 to public in 2022. The top 100 high frequency LLM-associated adjectives and adverbs list that they gathered from their study of peer reviews have been used extensively by numerous researchers (cf. Sprung, 2025; Bao et al., 2025). Considering the corpus that Liang et al. (2024) analyzed is supposed to compose of texts that pass through detailed procedures to ensure academic integrity is not breached, the findings appear to be quite concerning for the future of academic writing.

Another influential study was conducted by Bao et al. (2025), where researchers compiled the abstracts of pre-print papers written in the fields of Computer Science, Physics and Mathematics from arXiv. They analyzed the abstracts in terms of linguistic features such as lexical and syntactic complexity, readability, and the frequency of the LLM-associated adjectives and adverbs set provided by Liang et al. (2024). In addition to these features, the researchers also observed the new words in the abstracts that occurred after the advent of ChatGPT. Bao et al. (2025) found that the number of new words and the frequency of LLM-associated adjectives and adverbs increased considerably after the public launch of ChatGPT. The starkest increase in new words was in the noun category, which was interpreted as new technical vocabulary introduced by ChatGPT. The newly emerging nouns were noted especially in the abstracts from the

Computer Science field, while Physics and Mathematics appeared to be more stable over time, which was explained as the characteristic of the Computer Science field, which is widely known to be constantly developing and evolving with higher research activity. As to the other parts-of-speech, the most significant increase in the use of LLM-associated vocabulary was observed for the adjectives “intricate”, “valuable”, “exceptional”, “comprehensive”, “innovative” and “pivotal”, and for the adverbs “primarily”, “thoroughly”, “subsequently”, “effectively” and “particularly” (Bao et al., 2025), which aligns with the results of Reinhart et al. (2025). It was noted that especially the Computer Science and Physics fields exhibited an increase in the use of words widely preferred by LLMs whereas Mathematics abstracts were more stable, which was interpreted as the nature of LLM-associated vocabulary set chosen. The set consists of adjectives and adverbs which have a more descriptive function in writing that assist with emphasizing the importance of experimental results, but the field of Mathematics, by nature, relies more on abstract concepts that do not necessitate such descriptive vocabulary use.

Furthermore, the results of Bao et al. (2025) revealed that average word length and lexical density increased after the advent of ChatGPT, whereas the syntactic complexity kept decreasing, which was explained as a more extensive word range and a tendency towards more simplified sentence structures brought about by LLMs. The findings regarding lexical complexity support the results of the exploratory study by Zindela (2023), where the researcher found that ChatGPT-generated argumentative essays outperformed student-written ones in terms of lexical diversity, lexical density and lexical sophistication.

Turning back to Bao et al. (2025), the readability score of the human-written pre-print abstracts between 2022 and 2023 also displayed a decrease, which, according to the researchers, suggests a more specific and advanced terminology use that is caused by the constant

developments in the scientific disciplines. Another interesting finding Bao et al. (2025) reported is that the gap between the lexical complexity, syntactic complexity, and cohesion scores of the pre-print abstracts written by native English-speaking scholars (NESs) and non-native English-speaking scholars (NNEs) has considerably decreased, which denotes the influence of LLMs in assisting NNEs with the overall writing quality.

Gray (2024) analyzed and tested the Dimensions database that was compiled by Herzog, Hook and Konkiel (2020), from which only the document type “article” was chosen for consistency. From the Dimensions database, the author calculated the percentages of the articles in which the selected LLM-associated adjectives, adverbs and control words were attested ranging from the publications in 2015 to 2023. He found an especially considerable rise in the use of words that are commonly preferred by LLMs in the year 2023, which is also named “the post-LLM era”, the most striking increase in the use of words being “intricate”, “commendable”, “meticulous”, “meticulously”, “methodically” and “innovatively” (Gray, 2024, p.5). The researcher also calculated the frequency of the combination of the LLM-associated words in the articles, which also yielded significant results (strong indicators being attested 87.4% more often in articles published in 2023, while the highest change was 6.6% in previous years). The author lists the lack of the analysis of papers published in 2024 as a limitation, assuming that the situation regarding the increase in the use of AI-generated words will be gradually accelerating in 2024. This limitation is one of the factors that laid the groundwork for the dataset scope of the current thesis, which will also include the master’s theses published in the year 2024.

Another seminal study on the linguistic aspects of AI-generated academic texts was conducted by Tudino and Qin (2024) through corpus linguistics methods. The researchers created their corpus of human-written and AI-generated academic articles by compiling texts

written in the social sciences area from the Elsevier OA CC-BY Corpus and generating academic articles in the same fields by using the ChatGPT interface (GPT-4o mini). They analyzed the part-of-speech distribution, n-grams, and conducted a concordance analysis on the corpus data they collected. The results of Tudino & Qin (2024) revealed that ChatGPT articles included more nominalizations than the human-written articles, aligning with the results of Reinhart et al. (2025). Confirming the results of Guo et al. (2023), Tudino & Qin (2024) also observed that the LLM-generated articles had a higher number of adjectives, which was interpreted as a natural outcome that is brought about by the higher use of nominal items. As to the other main parts-of-speech, it was discovered that ChatGPT-generated academic articles feature a slightly higher number of verbs, whereas adverbs are utilized more frequently in the human-written articles. The observation of Tudino and Qin (2024) appears to differ from the findings of Muñoz-Ortiz et al. (2024, cited in Terčon & Dobrovoljc, 2025), which reported a fewer number of adjectives in AI-generated texts. And on the other hand, Tudino & Qin's (2024) results align with Markowitz et al. (2024), who found a more frequent use of adjectives in AI-generated texts, and with Chong et al. (2023), Guo et al. (2023) and Liao et al. (2023), who reported that LLM-generated texts feature a fewer number of adverbs.

Another result that Tudino and Qin (2024) achieved from the part-of-speech distribution analysis is that human-written articles displayed more instances of subordination, which is an indicator of a higher syntactic complexity (Norris and Ortega, 2009; Wu et al., 2020, as cited in Tudino & Qin, 2024). This result, again, does not align with the findings of some studies in the literature (Culda et al. 2025; Herbold et al. 2023; Liu et al. 2023; Nkhobo and Chaka, 2023, cited in Terčon & Dobrovoljc, 2025), where researchers reported that AI-generated texts had higher syntactic complexity scores.

2.4.The Literature on the Methodology and Linguistic Features Used for Comparing Human-Written and Synthetic Texts

Toney et al. (2026) provide a survey of the literature that addressed the question of humanness in texts. The studies that were examined in their paper converge on four computational measures that could be used while comparing human-written and AI-generated texts, which are readability, rhetorical structures, sentiment and style. The current thesis selected readability and stylistic metrics among these measures instead of sentiment and rhetorical structures, both due to the lower relevance of sentiment in academic texts, and because of the lower compatibility of lengthy master's theses for rhetorical analysis.

Average sentence length, average word length and sentence count are fundamental linguistic features, which help in calculating readability ease, in addition to shedding light on the sentence complexity and lexical complexity level of a given text. Therefore, they were selected also as separate linguistic features to be measured in the current thesis. Regarding readability ease, one of the most common frameworks adapted in the majority of the previous literature is the Flesch Reading Ease Score (FRES), which calculates a score between 0 and 100, denoting how easily a text can be read based on sentence length and syllables per word (Kincaid et al., 1975, cited in Toney et al. 2026). In addition, the Gunning Fog Index makes an estimation on the number of years of education necessary to understand a text, by calculating the words per sentence and complex words, which are words with three or more syllables (Gunning, 1952, cited in Toney et al., 2026). While there are studies specifically focusing on the readability of AI-generated and human-written texts (cf. Zou et al., 2026), these two readability indices were deemed sufficient to measure the readability ease of the master's theses in the current study, which will investigate numerous other linguistic dimensions besides readability.

The next linguistic dimension that is relevant for the current study is style, which is identified as the “non-functional linguistic choices distinct to an author’s writing” (Toney et al., 2026, p. 16). Since style is usually linked to authorship attribution, it is of special value for the current study, considering that LLMs will be treated as artificial authors, whose linguistic choices will be compared with human-authors. As to this measure, Douglas Biber’s tag set (1993) was selected as a comprehensive inventory of 67 lexical, syntactic, stylistic and stance features, following previous studies that achieved illuminating results (cf. Reinhart et al., 2025; Sardinha, 2024). The use of Biber’s tag set in the current thesis derives from an effort to give place to other features such as subordination, modals, co-ordination, verb classes (e.g. public verbs used for speech acts and suasive verbs which are instruments of narration and argumentation) and lexical classes such as conjuncts and amplifiers (Biber, 1993), which provide a structured overview that is not covered by other tag sets and metrics. The current thesis assumes that these additional features may yield valuable results regarding the difference between human-written and AI-generated discourse and argumentation features, which may go further than lexical and syntactic diversity and complexity information.

Georgiou (2026) also provides an extensive literature review of the studies that compare AI-generated and human-authored texts. It was pointed out that in a considerable proportion of such studies, the linguistic comparisons were based on the analysis of surface cues (Georgiou, 2026). The surface cues refer to lexical, morphosyntactic, stylistic and readability features, whose comparison provides insights into the distinction between AI-generated and human-written texts. According to the detailed review of Georgiou (2026), most of the studies in the literature that analyzed the texts in educational and academic spheres concentrated on calculating and comparing the lexical variety, syntactic complexity, the distribution of part-of-speech and

syntactic patterns, and readability indices. Furthermore, other studies predominantly examined the use of function words, punctuation patterns, and n-grams (Georgiou, 2026). Therefore, following the seminal literature provided by Georgiou (2026), lexical variety was measured by the calculation of type-token ratio, the function words were examined through lexical density scores, punctuation patterns were analyzed through the detection of the frequency of each punctuation mark attested in the corpus, and the part-of-speech (POS) distribution was identified by calculating the POS variability scores, also adding the and noun, verb, adjective and adverb ratios separately.

As to other linguistic features, the hapax legomena ratio and rare word ratio were selected from Hewa Julige (2026), in an attempt to examine whether words with lower frequency are increasingly attested in human-written master's theses in the ChatGPT era. Furthermore, the passive voice ratio feature was also taken from the study of Hewa Julige (2026), in order to see whether the AI-generated texts or the human-written texts make more frequent use of passive constructions, and whether there is a change in the quantity of such constructions ranging from the pre-ChatGPT to the post-ChatGPT era in human-written theses.

The last lexical measure considered was the LLM-associated words, which were provided by the seminal study of Liang et al. (2024) and Sprung (2025). The aim of checking the presence of LLM-associated words in human-written texts may lead to valuable information on the possibility of a diachronically increasing integration of GenAI tools in the writing procedure of human-authored master's theses.

The syntactic complexity measures were selected from Bagdasarov & Alves (2025) because the authors' selection was based on the most common syntactic features analyzed, which are average dependency length, tree depth and average branching factor. These features are

based on dependency relations of a given text, which provide information about the structural nesting, and sentential complexity of texts.

Finally, three other stylistic features were chosen from Hewa Julige (2026), which are stop word ratio, hedge word ratio and capitalized word ratio. A possible significant increase or decrease in the scores of these features from the pre-ChatGPT to the post-ChatGPT era of human-authored master's theses could denote possible traces of GenAI language in human writing.

The description and the significance of all linguistic features measured in the current thesis that were briefly explained above are presented in detail in Section 3.3.

2.5.The Description of the Research Gap and the Importance of the Current Study

The current thesis will aim to bridge an important gap in the literature, because although the characteristics of human-written and AI-generated abstracts, academic papers and essays were previously analyzed by numerous researchers (Yildiz Durak et al., 2025; Hewa Julige, 2026; Lin & Liu, 2026; Tudino & Qin, 2024), there is currently no study that compares and analyzes the linguistic features of human-written and fully LLM-generated master's theses. One study that was conducted on theses aimed to find an answer to the question of how to estimate the extent to which AI was utilized in the writing processes of theses written in Austrian universities (Sprung, 2025), which only addresses indicators that might imply AI involvement in human-written theses. Furthermore, another research that was carried out by Nowacki and Wrochna (2025), where researchers closely followed students' whole process of writing the thesis, merely identified and analyzed the theses where the students confirmed the usage of AI in certain samples of their work. Therefore, the AI-generated part of the corpus in Nowacki and Wrochna (2025) consists merely of the samples for which students openly reported the

integration of an LLM. On the other hand, the current thesis goes further by bringing the fully AI-generated master's theses into question, while also reporting the diachronic changes in the linguistic features of human-authored master's theses.

Research Questions and Hypotheses

RQ1: Has there been an increase in the frequency of LLM-associated vocabulary in Ca' Foscari University of Venice master's theses from 2019 to 2024?

Hypothesis: The current thesis hypothesizes that the theses that were written after the introduction of ChatGPT will have more AI-associated words. Increases in the frequency of these words in human-written theses are expected to start after 2022, which is the year in which ChatGPT was launched to public use.

RQ2: Do human-written theses written after the introduction of ChatGPT have more similarity to AI-generated theses in terms of linguistic features? If so, which linguistic measures observed in the ChatGPT era human-written group show the highest convergence with AI-generated texts?

Hypothesis: It is expected that the human-authored master's theses from the ChatGPT era might display higher similarity to the linguistic characteristics of AI-generated texts, while the pre-ChatGPT theses are not likely to show results that are similar to the features of machine generated theses in the linguistic analysis.

3. Methodology

This thesis aims to identify the lexical, syntactic and stylistic changes in master's theses that were written before and after the public release of ChatGPT and make a linguistic comparison between human-written and fully AI-generated theses. In the scope of the current thesis, only the publicly available master's theses that are uploaded to the Ca' Foscari University

of Venice thesis database³ are analyzed. The number of theses was refined through a detailed selection procedure, which are explained below in the Data Collection section.

3.1.Data Collection

3.1.1. Human-Written Theses

Firstly, the seven macro areas in the Ca' Foscari University of Venice master's programs were identified. Then, each program under the macro areas were searched for in the thesis database. Since the analysis is targeted on theses written in English, the areas that do not have any English theses were excluded as the first step. The year range for the selection of theses was designated as 2019-2024, to be able to make the comparison between theses before and after the advent of ChatGPT. The final year was chosen as 2024 due to the lack of theses uploaded after 2024 to the database.

The remaining macro areas from which the dataset was collected is as follows:

- Economics and Management
- Languages and Cultures
- Science and Technology

Due to the length of the target texts (i.e. full master's theses), the dataset was initially refined to include two programs per each macro area. The programs from each macro area were chosen specifically from different departments to allow for variety in the texts and to see whether the results of this thesis could be generalized to various fields. Besides, including more than one discipline could reveal important information as to the linguistic feature values of LLM texts that may vary depending on the writing styles of each academic area. For instance, Zou et al. (2026) found that LLM-generated compositions written in the linguistics field featured lexical

³ <https://unitesi.unive.it/>

oversimplifications of the theoretical concepts that are usually more nuanced in the human-written ones, while the machine-generated computer science compositions included more technical and complex terms than their human-authored counterparts. Furthermore, as to area variability, the possibility of a bias that could originate from specific terminology choices that are peculiar to one field could be lowered by encompassing various areas, following the methodology of Zhang and Crosthwaite (2026). The target number of theses in the current dataset was fifteen theses per each year for each program. Therefore, the ideal dataset that was designed initially included 540 theses.

However, not every program had 15 theses written in English each year. For that reason, the thesis number per year and per program was further refined to nine, considering the Environmental Sciences program that has only nine theses for the year 2020. The limitation in the number of theses led to the exclusion of one program from each macro area, discarding in total three programs out of six. Thus, the final dataset included three programs, each of which belong to three different macro areas, which again allows for variety in texts.

For programs which had more than nine theses per year, a random selection method was employed through a random number generator website⁴ to prevent any kind of data selection bias and to obtain fair results⁵. Later during the analysis part, a thesis from the Environmental Sciences program was excluded due to its incompatibility with computational processing, in order to prevent any problems that could have affected the results. The final dataset for the human-written theses is as follows:

⁴ <https://www.calculator.net/random-number-generator.html>

⁵ The random selection method was not applied only for the program that had nine theses in the year 2020 (further refined to eight). In that case, all the available theses in English were automatically selected.

Table 1. Corpus of Human-Written Theses

Macro Area	Program Name	2019	2020	2021	2022	2023	2024	Total:
Economy and Management	Economics and Finance	9	9	9	9	9	9	54
Languages and Cultures	Language Sciences	9	9	9	9	9	9	54
Science and Technology	Environmental Sciences	9	8	9	9	9	9	53
SUM		27	26	27	27	27	27	161

All the theses in the dataset were collected manually from the thesis database as PDF files and converted to .txt files (UTF-8) in order to facilitate the data analysis process. The index, acknowledgements, references and appendix sections were manually removed from all texts in order to prevent any unnecessary noise that might bring about misleading results in the quantitative analysis of the linguistic measures. There was no restriction or a specific preference on the length of texts during the text selection, due to the varying nature of human-written theses.

3.1.1.1. Classification of the Human-Written Theses

Taking into consideration the year in which ChatGPT was extensively released for public use, the human-written theses were classified into two main periods, which are named “pre-ChatGPT” and “ChatGPT”. Since ChatGPT was launched to public at the end of 2022, the human-written theses from 2019, 2020, 2021 and 2022 were classified as pre-ChatGPT theses, while the ones from 2023 and 2024 were grouped as ChatGPT theses. It is important to note that although other GenAI models like GPT-2 were present before 2022, they did not have the

sufficient ability to imitate the human writing style and they did not have a widespread use until 2022 (Hashemi, Shi & Corriveau, 2025).

3.1.2. AI-Generated Theses

Considering the widespread use of ChatGPT among students for text generation and receiving assistance with writing (Hashemi, Shi & Corriveau, 2025) and the limited scope of the current thesis, the only considered LLM was ChatGPT. The AI theses were generated through gpt-5-mini-2025-08-07 between August 12-18, 2026, using OpenAI API⁶ via Google Colab⁷. The selected GPT model has the overall ideal capabilities and the reasoning level among the most recent models available during data collection. Furthermore, since the GPT-4 series and its varieties have been used in many studies, the current thesis will bring a more recent model into analysis. Considering the discrepancy in the lexical diversity results achieved by the newer version of an LLM (i.e. ChatGPT-3 vs. ChatGPT-4) in Herbold et al. (2023, cited in Zhang and Crosthwaite, 2026), the employment of one of the most recent GPT models was found to be suitable for the current thesis. The reason why the OpenAI API was preferred over the chat interface is to facilitate and automatize the data collection process for the generated texts and to maximize the usefulness of the selected GPT model. The default temperature and top_p settings (temperature default setting: 1, top-p default setting: 1) for this model were not changed, since the current thesis does not aim to compare the creativity level of the outputs.

The quantity of AI-generated theses in the corpus exactly corresponds to the number of human-written theses (see Table 1), because the aim of the current thesis is to make comparisons between LLM-generated and human-authored master's theses. For that reason, each human-written thesis had a thematically matching AI-generated counterpart.

⁶ <https://openai.com/api/>

⁷ <https://colab.research.google.com/>

3.1.2.1.Steps for the Generation of AI-Theses

Since the output length does not allow for the generation of a full master's thesis, a specific prompting strategy was used, through which a whole thesis was generated section by section (see Appendix A). This prompting technique was adapted from the study by Tudino and Qin (2024), where the researchers built their corpus of papers by using the separate prompts that they devised for each section, asking the model to follow a consistent paper outline (i.e. list of sections). The current thesis also added additional steps in the prompting procedure which ensured that each master's thesis constituted a coherent and meaningful unit, which are explained in detail below.

The first and the fundamental step for the LLM-generated thesis condition was to feed the human-written thesis to the model and ask it to summarize the main ideas of the thesis in a paragraph. Then, the output of main ideas of the human-written thesis was fed back into the model in each prompt during the AI-thesis generation to ensure that the model keeps the core ideas in its memory at every step in the data generation process. This main idea approach was used by Guo et al. (2025) and Muñoz-Ortiz et al. (2024), where the researchers built the AI-generated part of their corpus based on the actual human-written counterparts. According to Muñoz-Ortiz et al. (2024), prompting LLMs to start based on a previous content (in their case human-written news headlines and the first three words of the first paragraph) assists the models in following the domain and topic, thereby increasing their chance of finding and using the words with the highest probability that are appropriate to the theme provided.

For section-by-section thesis generation, the master's thesis outline was also provided in each prompt to make sure that the model strictly follows that structure. In addition, before each new section generation, the model was asked to summarize the previous section and use that

summary as the basis for the next section, further ensuring consistency and coherence among the sections. Following the methodology of Zhang and Crosthwaite (2026), the model was prompted not to generate any references section, and this section was manually removed from each human-written thesis (see Section 3.1.1), which enables further comparability between the two text types.

The exact procedure for thesis generation is as follows:

Step 1- The LLM was asked to provide a summary of the main ideas of the human-written thesis.

Step 2- A starter prompt was given to the model, which makes it explicit that the model would be asked to generate a master's thesis section by section in the following prompts, and that it needs to strictly follow a blueprint and the core ideas.

Step 3- Sections were generated step-by-step.

Step 4- After the generation of each thesis section, the model was asked to summarize the previous section that it created.

Step 5- For generating the next section, the previous section summary was provided along with the main ideas, research questions and hypotheses⁸, and the thesis outline in each step. This procedure was iterated until the conclusion section.

For the comparability of the computation of the linguistic features of human-written and AI-generated texts, the prompts for the AI texts included specifications of a word count, which is intended to mirror the approximate length of the corresponding human-written thesis. For example, if a human-written thesis has approximately 22,000 words in total, the AI-generated

⁸ The research questions (RQs) and hypotheses for each AI-thesis are generated after the literature review. Therefore, the RQs and the corresponding hypotheses were fed to the model only after their own generation (i.e. after the generation of the Literature Review section), while the main ideas and the thesis blueprint were provided consistently in each step.

theses that are based on that human-authored thesis were generated with a prompt that commands the model to generate a thesis around 20,000 words. Clearly, as explained above, since the model cannot produce a whole output of a 20,000-word-thesis, the word count specifications were divided into sections⁹ (e.g. If a 20,000-word-thesis is aimed for, an example prompt given to the LLM is like the following: “Create a 2750-word literature review section of the thesis.”).

3.2. Corpus

The final corpus consists of 161 human-written and 161 AI-generated theses, which amounts to 322 master’s theses in total. After the removal of one human-written thesis from the Environmental Sciences area, the corresponding thesis was also discarded from the AI-generated part of the dataset. The final distribution of all the theses in the corpus is shown in the table below.

Table 2. The Entire Corpus of Human- and AI-Written Theses

Human-Written Theses

Macro Area	Program Name	2019	2020	2021	2022	2023	2024	Total:
Economy and Management	Economics and Finance	9	9	9	9	9	9	54
Languages and Cultures	Language Sciences	9	9	9	9	9	9	54
Science and Technology	Environmental Sciences	9	8	9	9	9	9	53

⁹ The only sections for which there was no set word limit were the abstract and the RQs, due to their structural nature which does not require lengthy texts.

SUM		27	26	27	27	27	27	161
------------	--	----	----	----	----	----	----	------------

LLM-Generated Theses

Macro Area	Program Name	2019	2020	2021	2022	2023	2024	Total:
Economy and Management	Economics and Finance	9	9	9	9	9	9	54
Languages and Cultures	Language Sciences	9	9	9	9	9	9	54
Science and Technology	Environmental Sciences	9	8	9	9	9	9	53
SUM		27	26	27	27	27	27	161
TOTAL SUM								322

3.3.Linguistic Features

In quantitative data analysis, certain metrics are selected and computed for specific target texts depending on the features that are aimed to be examined. Therefore, the linguistic measure selection might change considerably based on text type, and on specific textual aspects that are the focus. For the current thesis, four categories of linguistic measures were selected, which are lexical, syntactic, readability and stylistic metrics.

3.3.1. Lexical Features

3.3.1.1.Average word length

It is the mean length of words based on characters, which gives information about the complexity level of a text (Culda et al. 2025, p. 35). When a text has a higher word length on average, this could imply that it features a more complex and sophisticated language.

3.3.1.2.Type-token ratio (TTR)

It is one of the most common features through which lexical diversity of a text is measured. Type-token ratio is calculated by dividing the number of different words (i.e. word types) by the total number of words (Lu, 2012, cited in Zindela, 2023). A higher TTR score suggests a richer vocabulary range and a higher writing proficiency whereas a lower score indicates repetitiveness in vocabulary usage (Zou et al., 2026).

3.3.1.3.Hapax legomena ratio

This feature indicates the proportion of words that appear only once in a text, which denotes the originality of an author (Culda et al., 2025, p. 35). It is calculated by dividing the number of hapax legomena (i.e. words occurring not more than once in a text) by the total number of words/tokens in a text.

3.3.1.4.Rare word ratio

The definition and the calculation of this feature was adapted from Hewa Julige (2026), who describe rare words as the tokens that have a Zipf frequency that is lower than 3. It was argued that LLMs tend to use a higher number of rare words, while humans prefer using more common words in scientific texts (Varga & Baksa, 2025).

3.3.1.5.Lexical density

This feature is calculated by dividing the number of lexical words (i.e. content words) by the total number of words in a text (Zindela, 2023). It could provide insights into vocabulary sophistication and writing proficiency (Pan, Oliveira & Mello, 2025, p. 905).

3.3.1.6.Noun ratio, Verb ratio, Adjective ratio, Adverb ratio

These ratios in the content word category are calculated by dividing the total number of the specific part-of-speech by the total number of tokens (e.g. noun ratio = the number of nouns / the total number of tokens).

3.3.1.7.Part-of-speech diversity

This feature is calculated through dividing the distinct POS tags in a text by the total number of tokens. Just like TTR, the part-of-speech diversity also gives information about variety. A higher part-of-speech diversity value would indicate the employment of more various grammatical categories, which is expected to be higher in human-written texts in the pre-ChatGPT era, and gradually decreasing in the ChatGPT era, with the influence of GenAI.

3.3.1.8.LLM-associated word frequency

Previous researchers conducted word frequency analyses on human-written texts that were discovered to be assisted by AI. The frequency lists provided by Liang et al. (2024) and Sprung (2025) are taken as source lists in the diachronic comparison (pre-ChatGPT vs. ChatGPT era, and year-by-year) of the occurrence of the LLM-associated words in human-written part of the corpus (the entire word list can be viewed in Appendix C).

3.3.2. Syntactic Features

3.3.2.1.Sentence count

The sentence count is simply the total number of sentences in a text. This feature is the basis for other syntactic measures, in that, combined with other measures, it gives information on whether a high number of words are stacked into a sentence, or dispersed into more sentences.

3.3.2.2.Average sentence length

The average sentence length is calculated by dividing the total number of words by the total number of sentences. Therefore, a higher average number of words per sentence might mean a higher sentence complexity.

3.3.2.3.Average dependency length

This feature is measured by calculating the distance between the heads and their dependents. A higher dependency length is considered as an indicator of a higher syntactic complexity (Bagdasarov & Alves, 2025).

3.3.2.4.Average tree depth

Tree depth is the maximum distance from the root node to any leaf node in a dependency tree, and it gives information on structural nesting (Bagdasarov & Alves, 2025).

3.3.2.5.Average branching factor

This feature is “the mean number of immediate dependents (children) per internal node”, and it is calculated by dividing the total number of children nodes minus one, by the number of internal nodes (Bagdasarov & Alves, 2025, p.41). Tian (2025) argues that a higher value of average branching factor would indicate denser modifiers, which would necessitate higher dependency and recognition capabilities, and lower values would signify a more linear structuring (Bagdasarov & Alves, 2025).

3.3.3. Readability Metrics

3.3.3.1.Flesch reading ease

It is a readability metric that provides information about the accessibility of a reading text, with a higher reading ease score between the 0-100 scoring range usually interpreted as the text being less challenging and therefore easier to read. The exact formula is given below (Sudharshan et al., 2024, p.2):

$$\text{FRE} = 206.835 - (1.015 * \text{mean sentence length}) - (84.6 * \text{mean syllable count per word})$$

In line with this formula, complex vocabulary and long and rare words generally lead to a lower reading ease score (Pan, Oliveira & Mello, 2025).

3.3.3.2. Gunning fog index

It is another readability metric that makes an estimation on the required number of years of formal education to comprehend a text, which takes the sentence length and the percentage of complex words into account. Words with three or more syllables are considered as complex words in the calculation of this feature (Zou et al., 2026). The exact formula is as follows (Habib et al., 2026, p. 369):

$$\text{GFI} = 0.4 * ((\text{number of words} / \text{number of sentences}) + 100 * (\text{complex words} / \text{total number of words}))$$

According to this formula, a higher frequency of complex words increases the gunning fog index score, which would signify an increased level of formal education needed to comprehend the text, thereby a lower readability ease.

3.3.4. Stylistic Features

3.3.4.1. Stop word ratio

This feature is simply the proportion of stop words to the total number of tokens. A higher stop word ratio is usually attributed to human-authored texts, since humans tend to use more function words in writing (Banat, 2024).

3.3.4.2.Capitalized word ratio

Capitalized word ratio is the proportion of words starting with a capital letter to the total number of tokens. To get proper nouns specifically, the sentence-initial capitalized words were excluded through a pattern filtering on Python (RegEx). It is expected that since humans make use of more personal experiences and referential knowledge, they will employ proper nouns more frequently.

3.3.4.3.Punctuation count

This feature is used to analyze the types and frequencies of each punctuation mark in a text. It can give information regarding the writing rhythm and structural complexity (Pan, Oliveira & Mello, 2025). Furthermore, according to Sprung (2025), there are certain punctuation marks that are highly associated with ChatGPT-generated texts, which are colons, straight quotation marks (as opposed to the typographic quotation marks), exclamation marks, ellipses and em-Dashes (p. 4). Therefore, for the current thesis, analyzing the frequency of each punctuation mark, focusing on the ones listed by Sprung (2025) can provide useful information with regards to discovering the LLM integration in human-written master's theses.

3.3.4.4.Hedge word ratio

The hedge word ratio is calculated by dividing the number of hedge words (e.g. *somewhat, appears, almost*) by the total number of words. It is argued that use of hedging devices is one of the features that characterize human writing, which is a means of communicating authorial attitude and degrees of certainty, and ensuring reader engagement, appearing to be employed much less in LLM-generated content (Mo & Crosthwaite, 2025, cited in Wu, 2025).

3.3.4.5.Passive voice ratio

Passive voice ratio is measured through dividing the number of sentences that include a passive structure by the total number of sentences in a text. This feature was analyzed by Culda et al. (2025), where the researchers found that LLM-generated texts make a less frequent use of passive voice. According to the researchers, a higher percentage of passive voice reflects a fundamental quality of academic writing, which LLMs cannot entirely grasp yet.

3.3.4.6. Biber features (Biber Tag Set)

The Biber features were proposed by Douglas Biber in 1988, and it includes the scores for 67 linguistic features of a given text. The Biber tag set is composed of multiple categories for linguistic analysis, some of which are subordination features, co-ordination, and modals (Biber, 1988, the full tag set can be viewed in Appendix B).

3.4. The Implementation of the Linguistic Features

For the calculation of the selected metrics, Python scripts were constructed via Google Colab. The main Python NLP library chosen for the linguistic analysis was *spaCy* (en_core_web_sm, 3.8.0) (Honnibal et al., 2020), which formed the basis for the processing of the texts and the calculation of most of the linguistic measures thanks to its multidimensional and comprehensive features. Different Python packages and libraries were also utilized besides *spaCy*. A list of all the Python packages and for which linguistic features they were used are listed below:

- *elfen* (1.3.5) (Efficient Linguistic Feature Extraction for Natural Language Datasets; Maurer, 2026): The Python package *elfen* provides a great number of modules that concentrate on the calculation of various linguistic features. In the current thesis, extractor module was used as the initiator for the calculation of all the features in the

elfen feature inventory, after which other specific modules were used. The lexical richness module was used to calculate lexical density, type-token ratio and the hapax legomena ratio; POS module was employed to measure POS diversity, noun ratio, verb ratio, adjective ratio and adverb ratio; and finally, the semantic module was used to get the hedge word ratio.

- *pybiber* (0.3.1): This is the Python package through which 67 Biber features in the Biber tag set were automatically extracted (browndw, 2026).
- *PassivePy* (0.2.23): This package was used to measure the passive voice ratio in texts. The final output provides all passive structures, passive counts and the passive percentages, among which only the passive percentages were considered in the quantitative analysis (Sepehri, et al., 2023).

3.5.The Quantitative Analysis of Variables

The current study has four variables for analysis, which are year, source, course and the feature results. The presence of three categorical variables and one numeric variable do not allow for plots and analyses that can contain all variables at the same time. For that reason, separate plots were created with two different variable sets. Firstly, violin plots (grouped violin plots with split violins) and boxplots were created with year, source, and feature variables. Then, the same types of plots were created by replacing the year variable with the course variable. For LLM-associated words, punctuation marks and Biber features, only the most significant results were chosen for visualization due to the abundance of items analyzed in these measures.

3.6.Statistical Analysis

Following the methodology in Culda et al. (2025), the Kruskal-Wallis test was applied for the statistical analysis. This test allows for exploring whether there are statistically significant

differences among the groups, but it doesn't specify which groups differ from another. For that reason, in order to explore the specific connections among the groups, a further test was applied only for conditions which yielded statistically significant results in the Kruskal-Wallis test. This further analysis was conducted through Dunn's post-hoc test with Bonferroni correction, which makes a pairwise comparison, and states which specific pairs of groups show differences that are statistically significant. This test is useful for clarifying the differences when there are three or more groups to compare in a variable. For the current thesis, it was used for the year-based (six years) and course-based (three courses) comparisons but not applied for source-based comparisons because the source condition contains only two groups to compare, which are human and AI. This source-based comparison among the two groups (human vs. AI) was made applying the paired Wilcoxon signed rank test in order to discover the overall differences between the human and AI theses, and to see which features differ the most between these two groups. The paired Wilcoxon signed rank test was deemed necessary for this comparison, because the human- and AI-written theses are related in terms of topic, theme and area. Therefore, pairing each human-authored thesis with its AI-counterpart during the analysis enables a clearer comparison of the different feature values that could be achieved across these two sources.

In the case where no specific pairs show significant differences in the Dunn's post-hoc test, it is interpreted as the existence of only an overall difference. The Kruskal-Wallis test and the paired Wilcoxon signed rank test were performed through the Python library *scipy* (1.18.1) (Pauli Virtanen et al., 2020) and the Dunn's post-hoc test was carried out using the *scikits-posthocs* library (0.17.0) (Terpilowski, 2019). In addition, the *pingouin* package (0.7.0) (Vallat, 2018) was used for calculating the effect sizes (rank-biserial correlation) of the results achieved

from the paired Wilcoxon signed rank test. All these statistical analysis procedures were implemented on Google Colab.

In addition to these statistical tests, to show if the values of the linguistic features of human-written and AI-generated theses become closer after the introduction of ChatGPT (i.e. 2022), signed gap was measured for each linguistic feature. For a better comparison, the values for each feature were converted to z scores through the *scipy* library (1.18.1) to standardize the varying scales of the linguistic measures (e.g. While TTR has a range between 0-1, the Flesch Readability ease ranges from 0 to 100). The signed gap calculation focused on the changes observed over time for each feature, and the direction of the changes. The values from each year in human-written theses were subtracted from those of AI-written theses. Therefore, a positive signed gap would signify the cases where AI-generated theses values are higher than the human-authored ones, whereas a negative signed gap would denote the cases where the values of human-written theses surpass the AI-generated ones. Only the features which show similar values across sources in the ChatGPT era will be reported in the Results section, while the non-significant results can be viewed in the Appendix E.

4. Results

The results of the current study will be reported with the previously mentioned plots, along with the results of the statistical analysis tests that were explained in the previous section. For clarity, the reporting of the results will take place under four linguistic feature subcategories, which are lexical, syntactic, readability and stylistic features. Each feature listed under these subcategories will have three dimensions for plots and results, which are year-based, course-based and source-based. In the course-based comparison plots, the label “EcoFin” refers to the

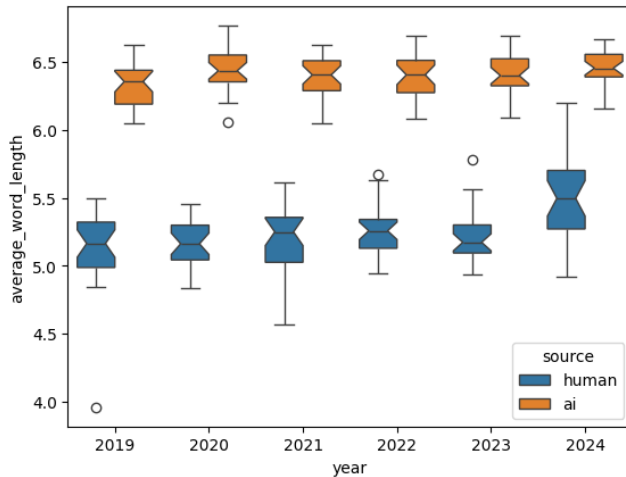
Economics and Finance area, “EnvSci” refers to the Environmental Sciences course, and finally “LangSci” represents the theses in the Language Sciences program.

4.1.Lexical Features

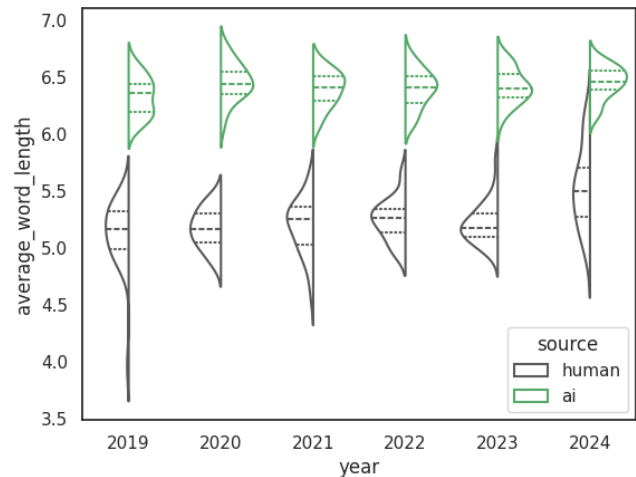
4.1.1. Average Word Length

4.1.1.1.Year-Based Comparison

The results reveal that GPT-generated master’s theses have a higher average word length, which is consistent in both the pre-ChatGPT and the ChatGPT era. As to the human-written theses, it could be noted that the average word length values peaked in the year 2024.



Year-based average word length values of human- and AI-written theses (box plot)

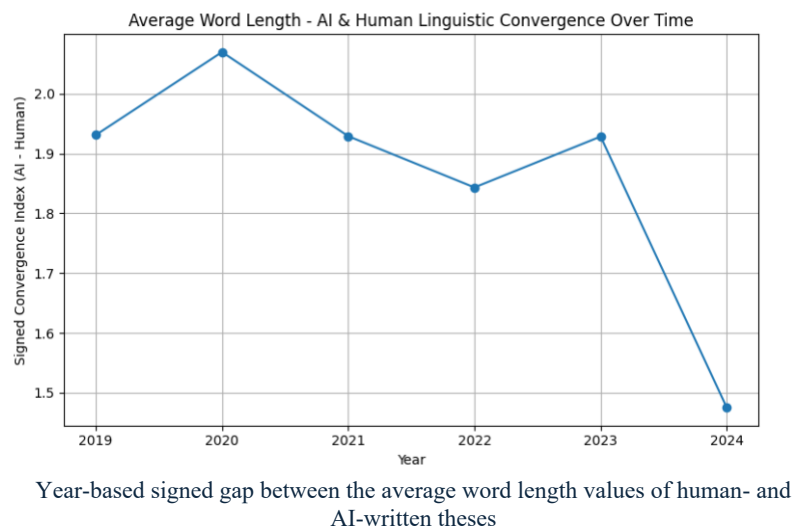


Year-based average word length values of human- and AI-written theses (violin plot)

The Kruskal-Wallis test results for the average word length values exhibit that there is a statistically significant difference in the year-based comparison of the human-written theses condition ($H = 25.642$, $df = 5$, $p < .001$). Dunn’s pairwise post-hoc test further revealed significant differences between the following pairs in the human-only group: 2019-2024 ($p < .001$), 2020-2024 ($p < .001$), 2021-2024 ($p = .005921$) and 2023-2024 ($p = .004378$). This pairwise comparison further confirms the significant increase in the average word length values of the human-authored theses in the last year of the ChatGPT era (2024).

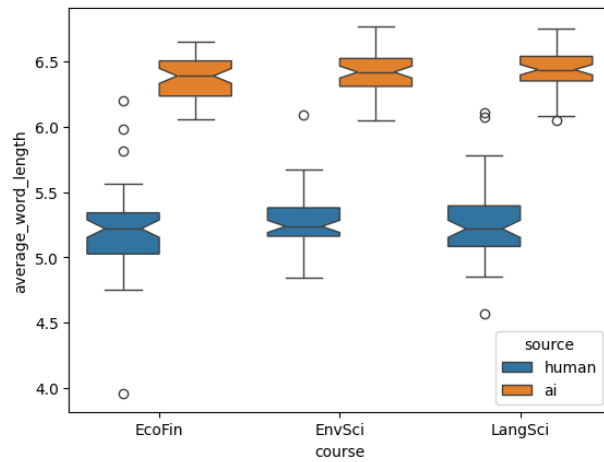
As to the year-based comparison between the average word length results of the two sources, the pairwise comparison showed that the values of human-theses from each year differs significantly from the GPT-generated theses from each year ($p < .001$), which is further evidence to the overall discrepancy between the two sources for this feature.

Furthermore, the signed gap calculation for this feature yielded valuable results. It was exhibited that the difference between human- and AI-generated thesis average word length values are lower in the ChatGPT era, specifically in the year 2024.

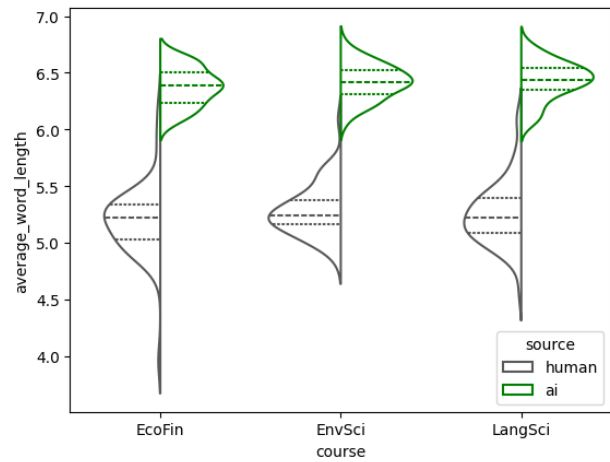


4.1.1.2.Course-Based Comparison

The average word length values of human-written and AI-generated theses, again, show an evident difference in the course-based comparison plots, with GPT-generated theses consistently having higher average word lengths in each micro area.



Course-based average word length values of human- and AI-written theses (box plot)



Course-based average word length values of human- and AI-written theses (violin plot)

The Kruskal-Wallis test did not yield any statistically significant results in the course condition when the source variable was isolated (course-based comparison among human-theses: $p = .254$; course-based comparison among AI-theses: $p = .231$). When the values of both sources were examined in the course-based comparison (i.e. human and AI combined), the Kruskal-Wallis and Dunn's post-hoc test showed significant p-values for each course between the human and AI-theses ($p < .001$ for both Kruskal-Wallis and Dunn's test, consistent among all pairs). The statistical results, therefore, confirm the data visualized on the plots, signifying that the average word length for each micro area differs significantly between the two sources, with no specific pairs standing out.

4.1.1.3.Source-Based Comparison

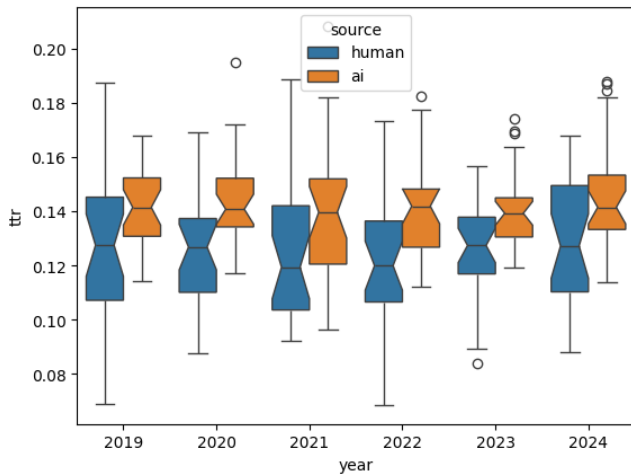
The paired Wilcoxon signed-rank test results on the overall average word length values of human- and AI-written theses revealed significant differences between the two sources ($W =$

0.0, $p < .001$, $r = 1.0$, Direction: AI > Human). In other words, the average word length values of GPT-generated theses are considerably higher than human-authored ones overall.

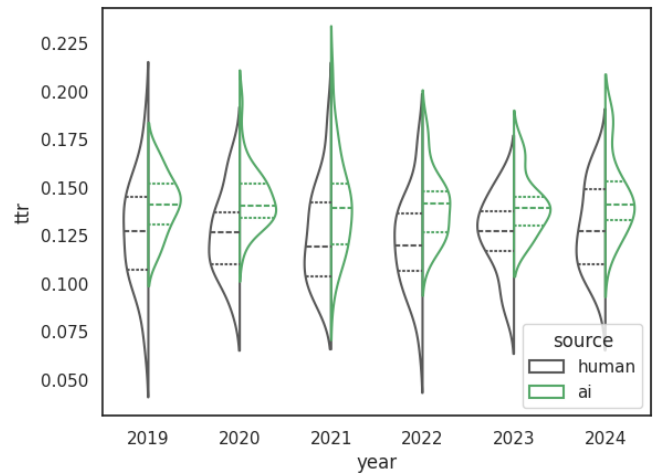
4.1.2. TTR

4.1.2.1. Year-Based Comparison

The calculation of the type-token ratio values of human-written and AI-generated master's theses showed that LLM-generated master's theses have higher TTR scores than human-authored ones in each year from 2019 to 2024. The figures below display the considerable discrepancy between the confidence intervals of the TTR values of the human-written and GPT-generated theses for each year, the latter surpassing the former consistently.



Year-based TTR values of human and AI-written theses (box plot)



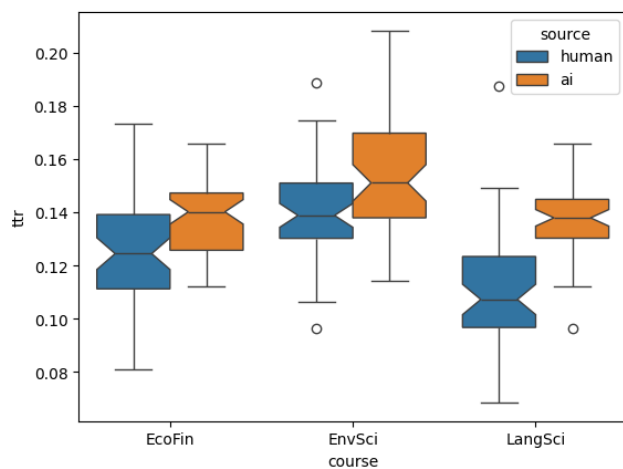
Year-based TTR values of human and AI-written theses (violin plot)

The Kruskal-Wallis test does not demonstrate a statistically significant difference in the year-based comparison among the human-written theses ($H = 1.337$, $df = 5$, $p = .931$), and the year-by-year comparison of AI-generated theses ($H = 1.657$, $df = 5$, $p = .894$). However, when the source variable is added to the comparison (i.e. human and AI combined group), the tests revealed that the human-written theses in 2022 differed significantly from the theses of three years in the AI condition: 2019 ($p = .0492$), 2020 ($p = .0124$) and 2024 ($p = .0218$). In other

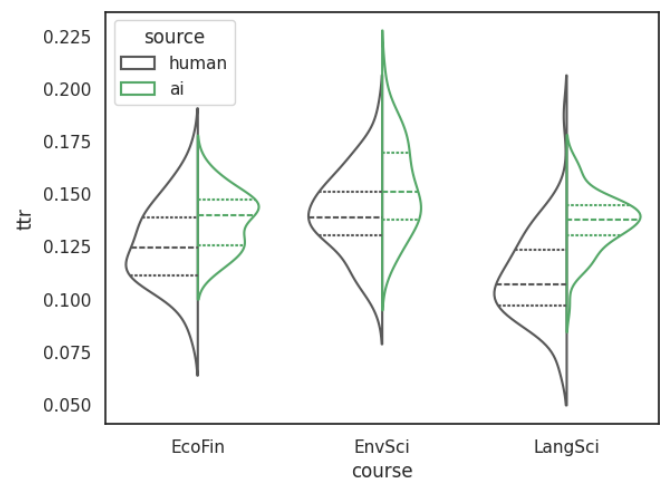
words, the highest discrepancies between the TTR values of human-written and AI-generated theses in the year-based comparison were observed between 2022 human-authored theses and the LLM-generated theses in 2019, 2020 and 2024, where the human-written theses showed considerably lower values.

4.1.2.2.Course-Based Comparison

As to the TTR values of human and AI-theses, the results reveal that the AI-generated theses have a higher TTR than human-written theses in each micro area.



Course-based TTR values of human and AI-written theses (box plot)



Course-based TTR values of human and AI-written theses (violin plot)

Furthermore, statistically significant differences were found in the course-based comparisons of both human-only ($H = 45.086$, $df = 2$, $p < .001$) and AI-only text conditions ($H = 21.269$, $df = 2$, $p < .001$), in addition to the comparison across sources (human vs. AI, $H = 101.323$, $df = 5$, $p < .001$) achieved from the Kruskal-Wallis test.

Dunn's post-hoc test further revealed that significant differences exist among all courses in the human-written theses ($p < .05$ between each pair), while AI-generated texts only show discrepancy in the following micro area pairs: Economics and Finance and Environmental

Sciences ($p < .001$); and Environmental Sciences and Language Sciences ($p < .001$). As to the course-based comparison for two sources together, the Dunn's post-hoc test shows that the human-authored theses in the Language Sciences area differ significantly from all AI-theses of each course ($p < .05$), with considerably lower values.

4.1.2.3.Source-Based Comparison

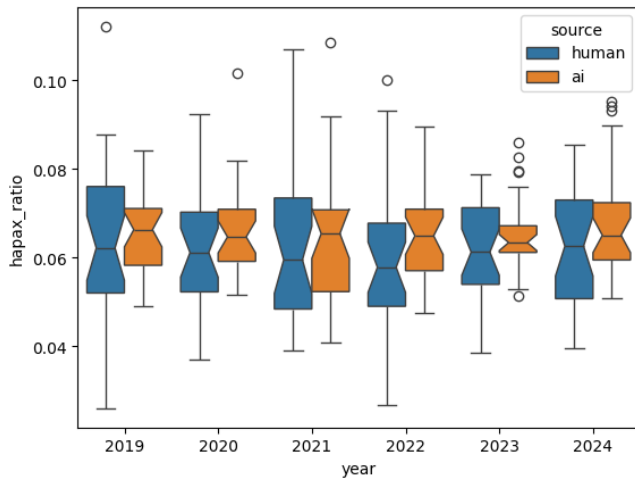
The TTR results that were further analyzed with the paired Wilcoxon signed rank test validate the consistently higher values observed in GPT-generated theses ($W = 1495.0$, $p < .001$, $r = 0.770$, Direction: AI > Human). So, it was once again confirmed that LLM theses have a higher TTR than human-authored theses.

4.1.3. Hapax Legomena Ratio

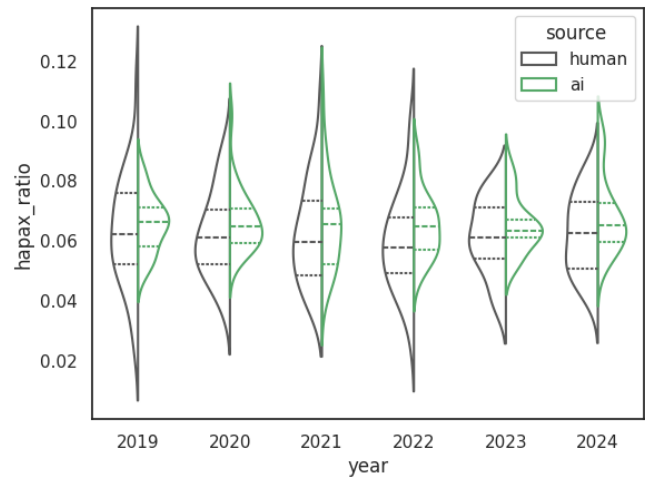
4.1.3.1.Year-Based Comparison

The results of hapax legomena ratio demonstrate that except for the year 2019, AI-generated master's theses feature higher median values of hapax legomena than human-written theses. In addition, AI-generated theses show almost no change in the ratio results in different years compared, except for 2023, while human-authored theses show slightly more change throughout the years, though again, not following a clear or regular pattern.

The Kruskal-Wallis test did not reveal any statistically significant differences in the year-based comparison in any of the groups (human-only: $H = 1.022$, $df = 5$, $p = .96$; AI-only: $H =$



Year-based hapax legomena ratio values of human- and AI-written theses (box plot)

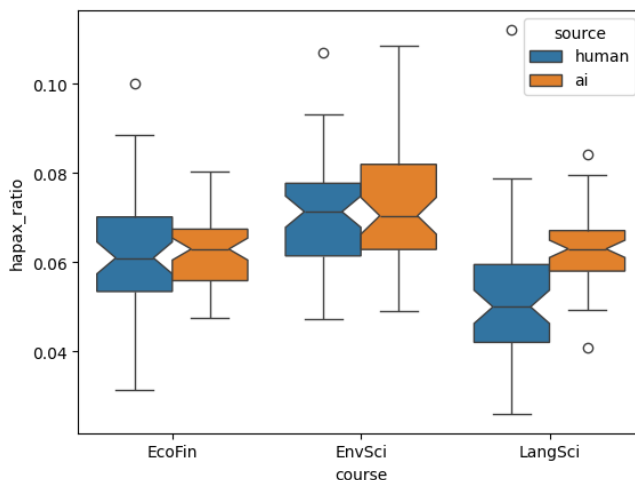


Year-based hapax legomena ratio values of human- and AI-written theses (violin plot)

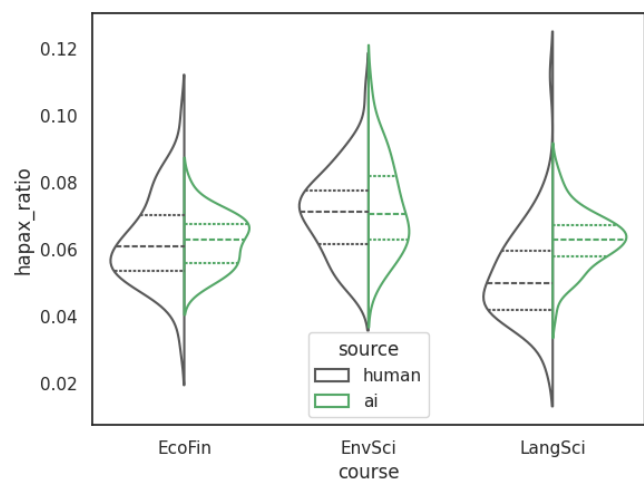
1.124, $df = 5$, $p = .951$; human and AI: $H = 10.401$, $df = 11$, $p = .49$). Therefore, it appears that the year variable does not play a determinant role in the changes observed for the hapax legomena ratio.

4.1.3.2.Course-Based Comparison

Both from the boxplot and the violin plot, it could be seen that the largest difference in the confidence intervals of the displayed data is between the human-written and AI-generated



Course-based hapax legomena ratio values of human- and AI-written theses (box plot)



Course-based hapax legomena ratio values of human- and AI-written theses (violin plot)

theses written in the Language Sciences area. Specifically, the human-authored theses in Language Sciences have most of the values situated in the lower scales of the plot, which shows that most of the theses in that area have a lower ratio of hapax legomena. And the GPT-generated theses in Environmental Sciences show more values situated on the higher scales of the plot.

The Kruskal-Wallis test yielded statistically significant p-values among the human-written theses ($H = 43.710$, $df = 2$, $p < .001$), AI-generated theses ($H = 19.960$, $df = 2$, $p < .001$), and across human-written and AI-generated theses ($H = 74.457$, $df = 5$, $p < .001$). Therefore, Dunn's post-hoc test was subsequently performed.

According to the pairwise group comparisons, the groups that differed from the others the most are the human-authored theses from Language Sciences ($p < .05$ across all pairs in the human-only and human vs. AI condition) with considerably lower values, and GPT-generated theses from Environmental Sciences with substantially higher values ($p < .05$ except for the pairwise comparison with human-written Environmental Sciences, where $p = 1.000$).

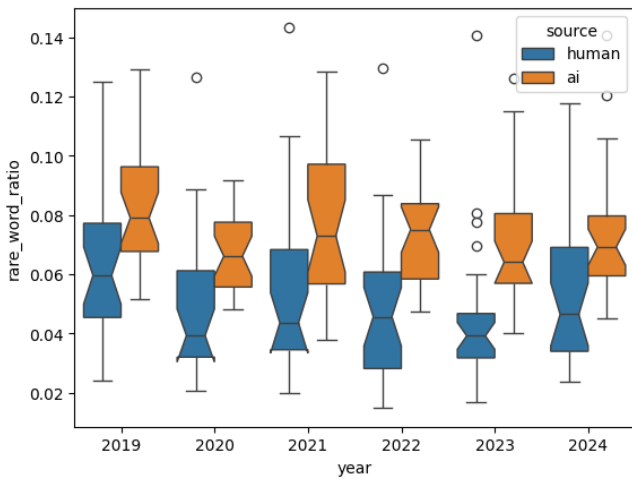
4.1.3.3.Source-Based Comparison

The values of hapax legomena ratio across sources exhibit that the AI-generated theses feature a higher proportion of hapax legomena ($W = 4191.0$, $p < .001$, $r = 0.36$, Direction: AI > Human), thereby confirming the median values previously visualized.

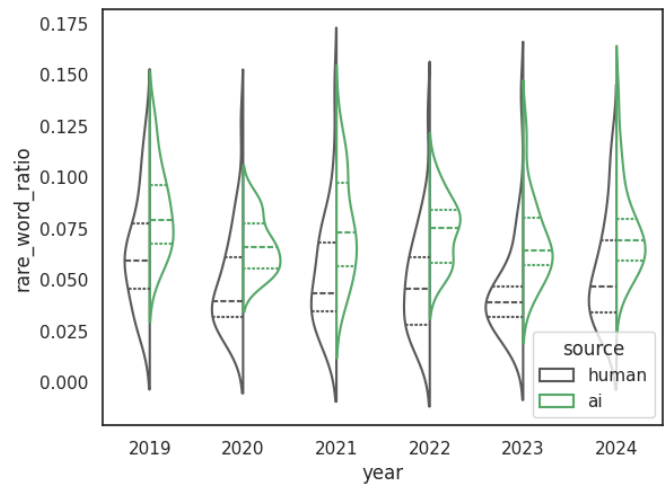
4.1.4. Rare Word Ratio

4.1.4.1.Year-Based Comparison

The results of the rare word ratio display that the AI-generated theses have a higher proportion of rare words than human-written theses in each year.



Year-based rare word ratio values of human- and AI-written theses (box plot)

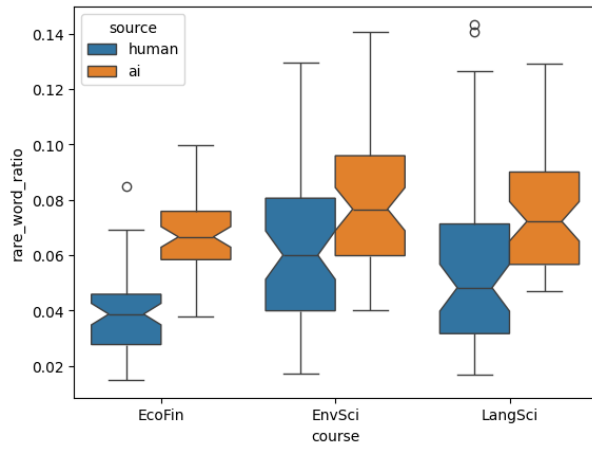


Year-based rare word ratio values of human- and AI-written theses (violin plot)

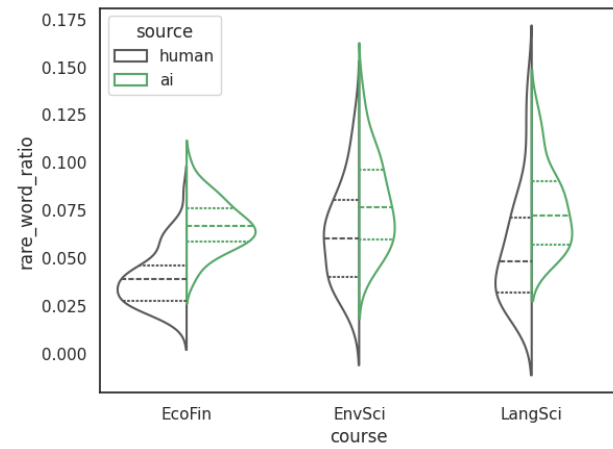
The Kruskal-Wallis test confirms this with the statistically significant difference between human-authored and GPT-generated theses ($H = 84.683$, $df = 11$, $p < .001$). As to the year-by-year discrepancy among human-written theses, the test did not display a significant p-value ($H = 9.859$, $df = 5$, $p = .079$). Regarding the year-based differences of all texts (i.e. human and AI-combined set), the Dunn's post-hoc test reveals that human-written theses from the year 2023 diverges more significantly from all AI-texts ($p < .01$) when compared to the human-theses of other years, showing considerably lower values.

4.1.4.2.Course-Based Comparison

According to the analysis of the course variable, the results demonstrate that the AI-generated texts have a higher rare word ratio than human-theses in each area, the most significant



Course-based rare word ratio values of human- and AI-written theses (box plot)



Course-based rare word ratio values of human- and AI-written theses (violin plot)

discrepancy being observed in the Economics and Finance area, where human-written theses have considerably lower values than the AI-generated counterparts. As to the distribution of the data, it could be seen that the data density of human and AI theses are considerably similar in all three courses.

Kruskal-Wallis test and Dunn's post-hoc test further confirmed that the human-authored theses from the Economics and Finance area differed significantly from the other areas of both sources ($p < .01$ in all pairwise comparisons except for the comparison with the human-written Language Sciences theses, where $p = .0262$).

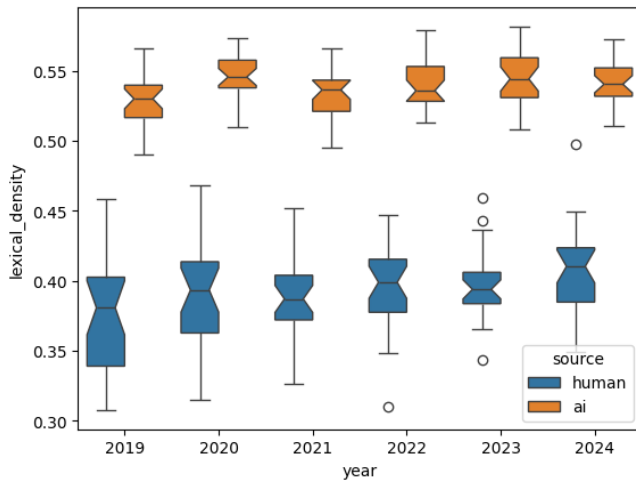
4.1.4.3. Source-Based Comparison

The paired Wilcoxon signed rank test demonstrated that AI-generated theses have significantly higher rare word ratios than their human-authored counterparts ($W = 873.0$, $p < .001$, $r = 0.866$, Direction: AI > Human).

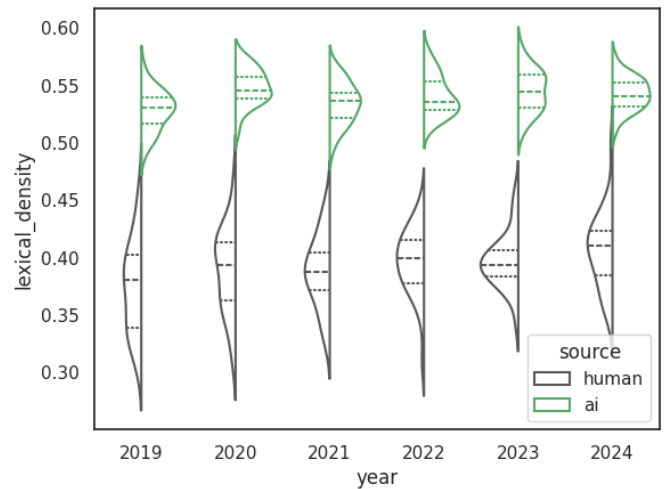
4.1.5. Lexical Density

4.1.5.1. Year-Based Comparison

According to the analysis results, there is an evident discrepancy in the lexical density values between the human-authored and GPT-generated master's theses, with the latter showing higher results than the former for each year.



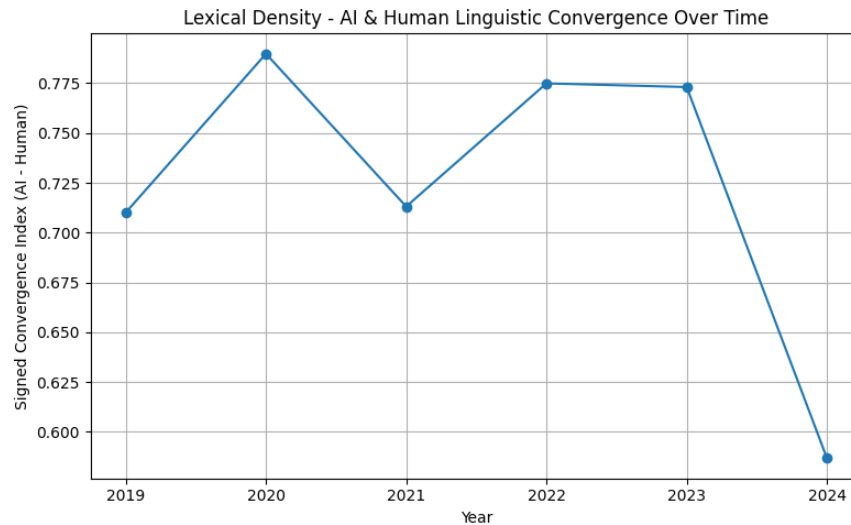
Year-based lexical density values of human- and AI-written theses (box plot)



Year-based lexical density values of human- and AI-written theses (violin plot)

Further statistical tests demonstrated that in the human-written group, theses from two specific years differed significantly from one another ($p = .012$), which are 2019 (pre-ChatGPT) and 2024 (ChatGPT era). The year-based comparison across two sources (human and AI) demonstrated that theses from each year and from both sources differed significantly ($p < .01$), which validates the overall considerable discrepancy in the lexical density between AI- and human-theses.

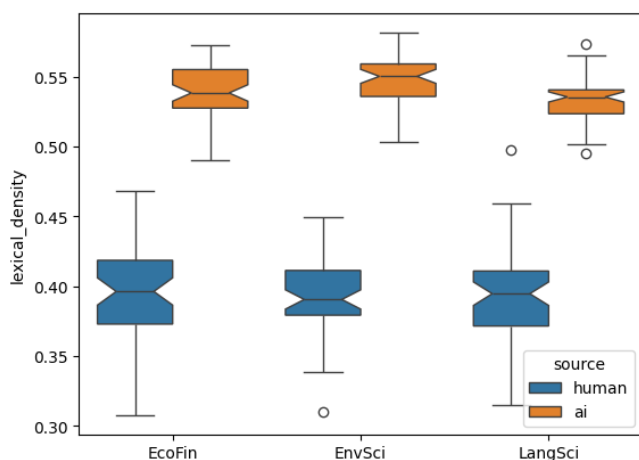
The signed gap calculation for lexical density further reveals considerable results. The change in the difference between human and AI values over time demonstrate that human-written theses have the most similar lexical density values to GPT-generated thesis values in 2024.



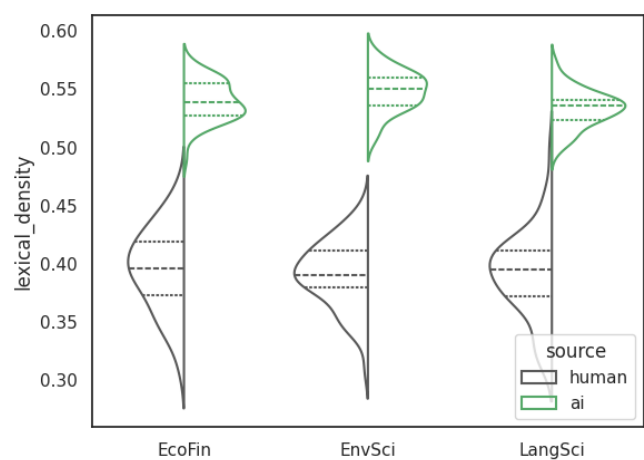
Year-based signed gap between the lexical density values of human- and AI-written theses

4.1.5.2.Course-Based Comparison

The figures exhibit that human-authored and AI-generated theses differ also in terms of micro areas, with AI-texts surpassing the lexical density values of human-theses as expected.



Course-based lexical density values of human- and AI-written theses (box plot)



Course-based lexical density values of human- and AI-written theses (violin plot)

The statistical tests reveal that the Environmental Sciences group in the AI-only condition diverges most significantly from other courses (Economics and Finance, $p = .0368$; Language Sciences, $p < .001$). As to the course-based comparison of the lexical density values across two text sources, each course in the human-written group shows a significant difference from each course of LLM-generated theses ($p < .01$). So, there was not a specific group that diverged from other groups in the overall course-based comparison.

4.1.5.3.Source-Based Comparison

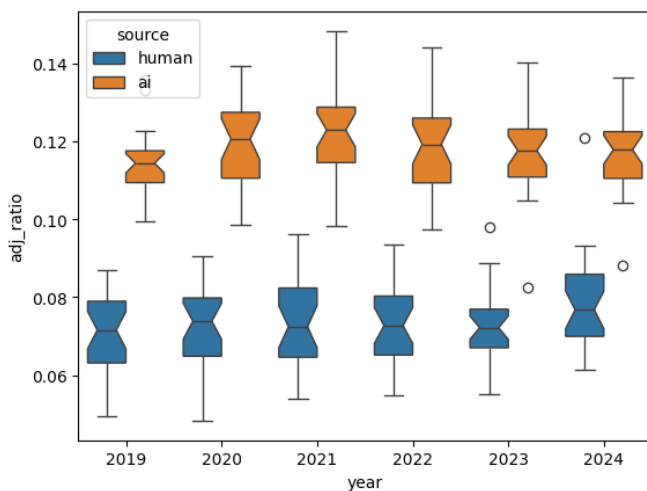
The proportion of content words present in GPT-generated master's theses were further revealed to be substantially higher than human-authored theses overall ($W = 0.0$, $p < .001$, $r = 1.0$, Direction: AI > Human).

4.1.6. Adjective, Adverb, Noun and Verb Ratio

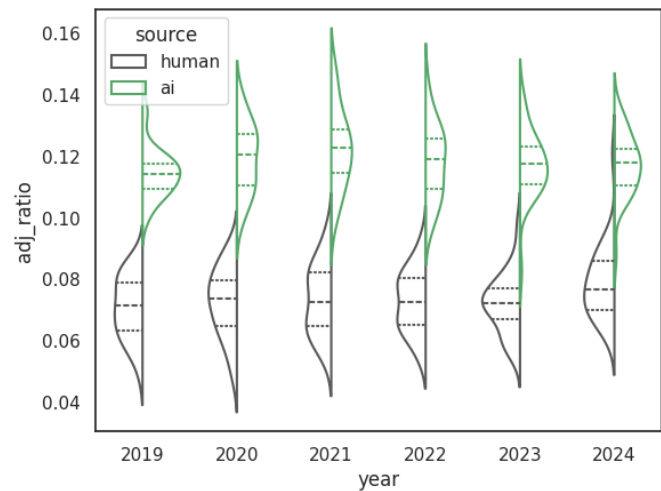
4.1.6.1.Adjective Ratio

4.1.6.1.1. Year-Based Comparison

The results of the linguistic analysis revealed that the adjective ratio of AI-generated theses is higher than human-written theses consistently in each year. The statistical tests confirm



Year-based adjective ratio values of human- and AI-written theses (box plot)

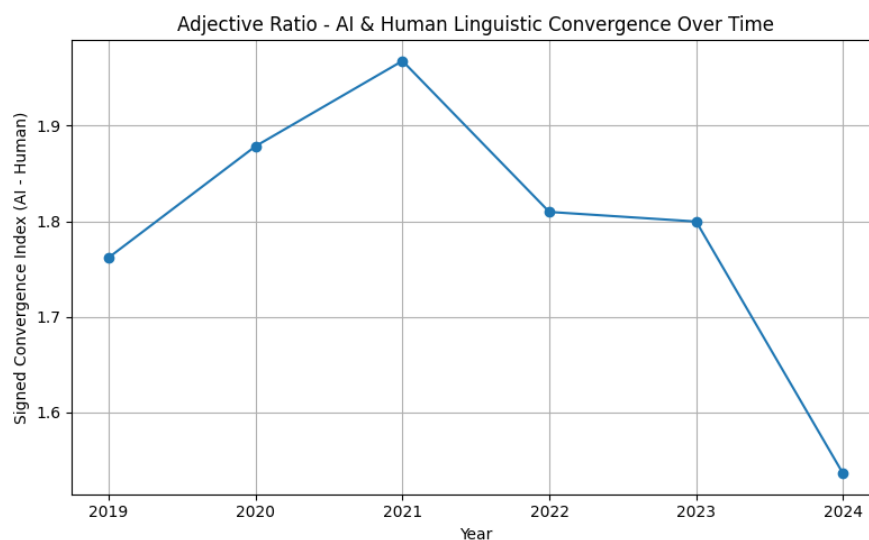


Year-based adjective ratio values of human- and AI-written theses (violin plot)

the data that can be seen on the figures; in that, human-written theses from each year differ significantly from GPT-theses in each year ($p < .001$).

Regarding year-based comparison in the human-written group, the Kruskal-Wallis test did not find any statistically significant difference among the six years ($H = 6.283$, $df = 5$, $p = .279$).

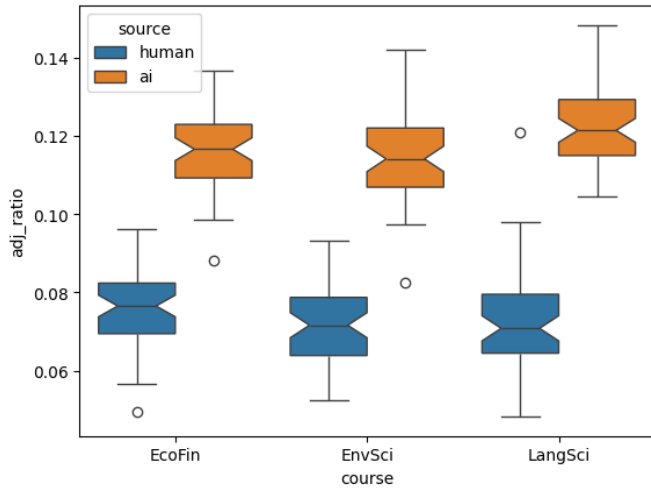
Although there were no significant year-based differences in the human-only group, the signed gap result demonstrates that the discrepancy between the proportion of adjectives in human- and AI-written theses reached the lowest value in 2024. Namely, the adjective ratios of human and LLM theses have the most similar values in 2024 (final year of the ChatGPT era).



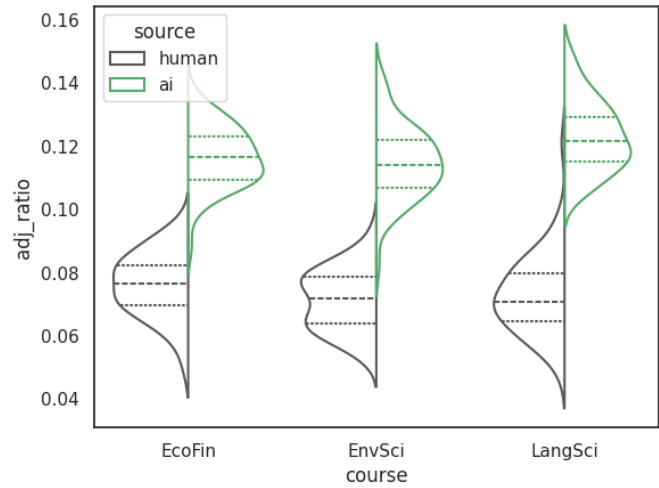
Year-based signed gap between the adjective ratio values of human- and AI-written theses

4.1.6.1.2. Course-Based Comparison

The figures below demonstrate the discrepancy between the adjective ratios of human-authored and LLM-generated theses. According to the results, the AI-generated texts have a higher proportion of adjectives than human-written theses in each micro area.



Course-based adjective ratio values of human- and AI-written theses (box plot)



Course-based adjective ratio values of human- and AI-written theses (violin plot)

The only statistical difference in the human-only group was between the theses from Economics and Finance and Environmental Sciences area ($p = .0348$), whereas for the AI-only group, the most significantly differing program was Language Sciences ($p < .01$ for both pair combinations). However, among all theses (i.e. human & AI theses combined), there was no specific discrepancy observed in the pairwise comparison. Rather, there is a significant course-based difference between all human-written and LLM-generated theses ($p < .001$ for each pairwise comparison).

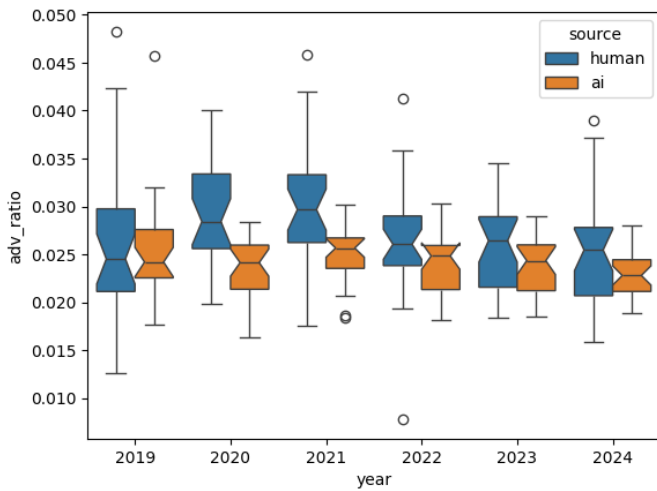
4.1.6.1.3. Source-Based Comparison

The differences between the adjective ratio across the two sources were validated with the paired Wilcoxon signed rank test ($W = 0.0$, $p < .001$, $r = 1.0$, Direction: AI > Human). To clarify, the GPT-generated theses had a higher proportion of adjectives than their human-authored counterparts.

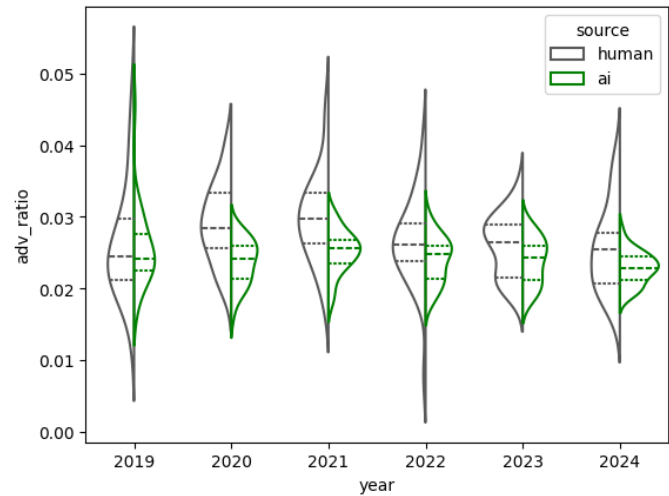
4.1.6.2. Adverb Ratio

4.1.6.2.1. Year-Based Comparison

According to the results of the linguistic analysis, the human-written theses feature a higher ratio of adverbs than AI-written theses in each year compared. The human-authored



Year-based adverb ratio values of human- and AI-written theses (box plot)



Year-based adverb ratio values of human- and AI-written theses (violin plot)

theses show an increase from 2019 to 2021 (pre-ChatGPT era), and then a decrease from 2021 to 2024 (ChatGPT era). Since the AI-theses have lower adverb ratios overall, it can be argued that the values of human-written theses from the pre-ChatGPT era have more different values from AI theses.

The Kruskal-Wallis test conducted on the year variable display significant variations in the human-written set ($H = 12.969$, $df = 5$, $p = .0236$), while the year-by-year variations did not result to be statistically significant for the AI-generated set ($p = .071$). This is also evident on the plots, where the adverb ratio in AI theses appear to be more stable when compared to the human-written ones. Dunn's post-hoc test on the pairwise comparisons for the human-only condition, however, did not yield statistically significant p-values for specific year pairs ($p > .05$), which indicates that there is only an overall difference among the human-written year groups.

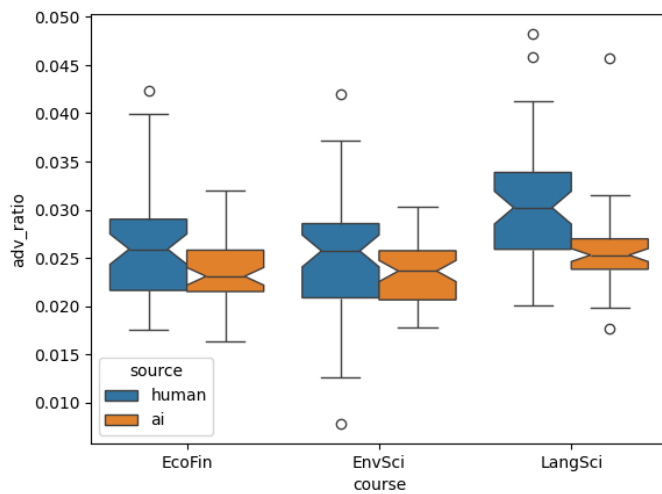
Post-hoc test on the pairwise comparisons for the human and AI-combined condition, on the other hand, revealed that human-written theses in 2020 and 2021 differ significantly from AI-

generated theses in 2020 (human-2020, $p = .013$; human-2021, $p = .0049$), 2022 (human-2020, $p = .0134$; human-2021, $p = .011$), 2023 (human-2020, $p = .015$; human-2021, $p = .005$) and 2024 (human-2020, $p < .001$; human-2021, $p < .001$). Therefore, it can be inferred that the majority of the pre-ChatGPT human-theses differ significantly from the adverb ratio values of LLM texts, denoting less LLM-like lexical values.

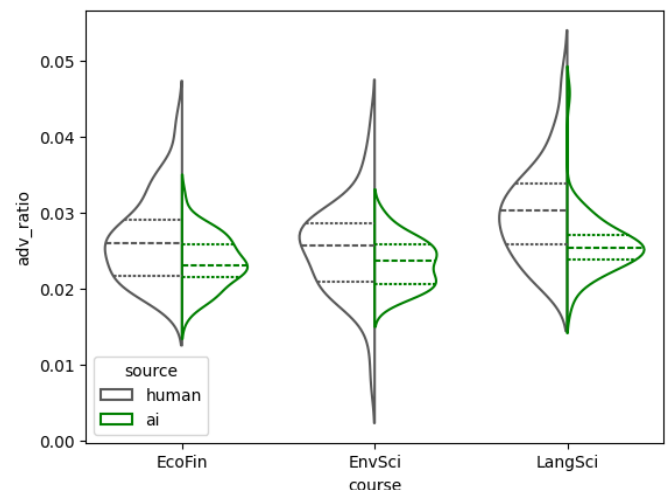
4.1.6.2.2. Course-Based Comparison

In the course-based comparison between the AI and human condition, it can, again, be seen that human-written theses have a higher ratio of adverbs across all course conditions.

Regarding the distribution of the data, Language Sciences theses display a more differing distribution when compared to the other micro areas analyzed.



Course-based adverb ratio values of human- and AI-written theses (box plot)



Course-based adverb ratio values of human- and AI-written theses (violin plot)

Dunn's post-hoc test confirmed these findings; in that, both for human-only and AI-only conditions, the theses from the Language Sciences group differ more significantly from the texts from other courses (human-only, $p < .01$ with both courses; AI-only, $p < .01$ with both courses). Finally, with respect to the human and AI-combined condition, the human-written theses from

the Language Sciences area showed the highest divergence by being the group that showed statistically significant differences in all pairwise comparisons ($p < .05$).

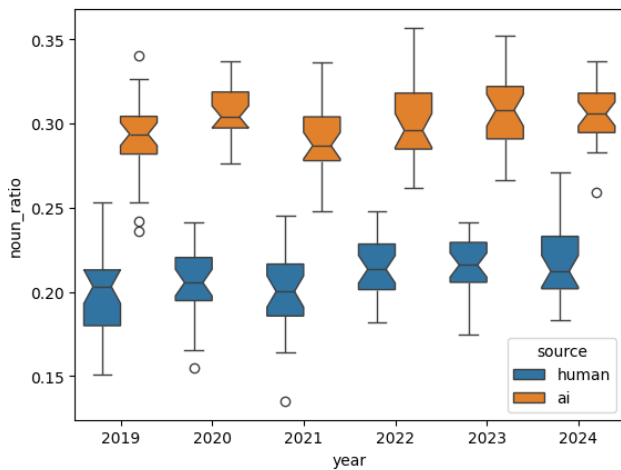
4.1.6.2.3. Source-Based Comparison

The paired Wilcoxon signed rank test demonstrated that the proportion of adverbs in human-authored master's theses considerably surpass their LLM-generated counterparts ($W = 3457.0, p < .001, r = -0.47$, Direction: Human $>$ AI).

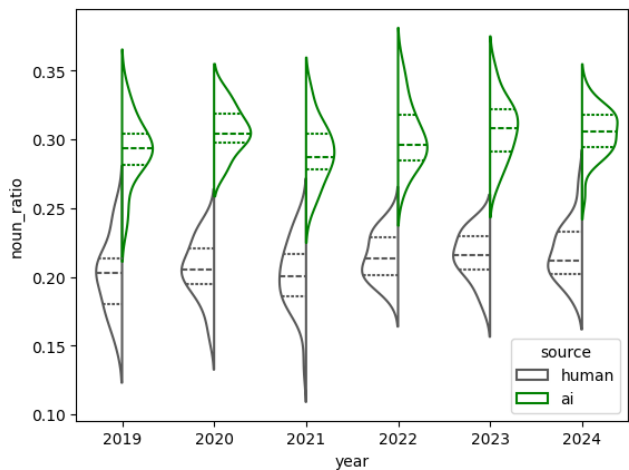
4.1.6.3. Noun Ratio

4.1.6.3.1. Year-Based Comparison

The analysis results show that AI-generated theses feature a higher proportion of nouns than human-written theses in each year. As to the human-only and AI-only conditions, the Kruskal-Wallis test revealed significant p-values for the year-based comparisons (human-only: $H = 11.397, df = 5, p < .001$; AI-only: $H = 14.657, df = 5, p < .001$). Nonetheless, Dunn's post-hoc test for pairwise comparisons among the years did not yield statistically significant differences between any specific pairs ($p > .05$). Therefore, it can be said that there is an overall year-based difference between the theses in the human-only and AI-only conditions.



Year-based noun ratio values of human- and AI-written theses (box plot)

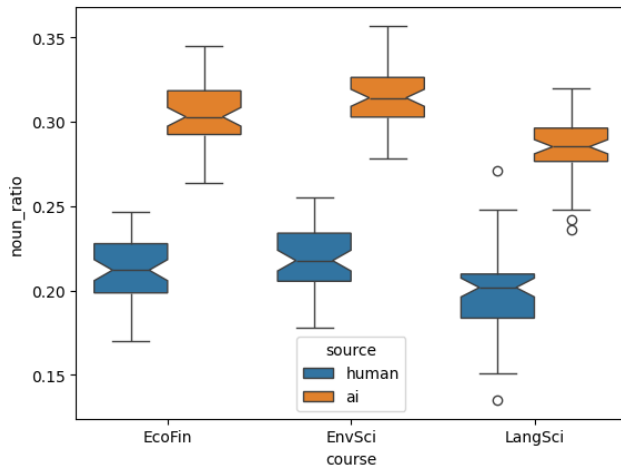


Year-based noun ratio values of human- and AI-written theses (violin plot)

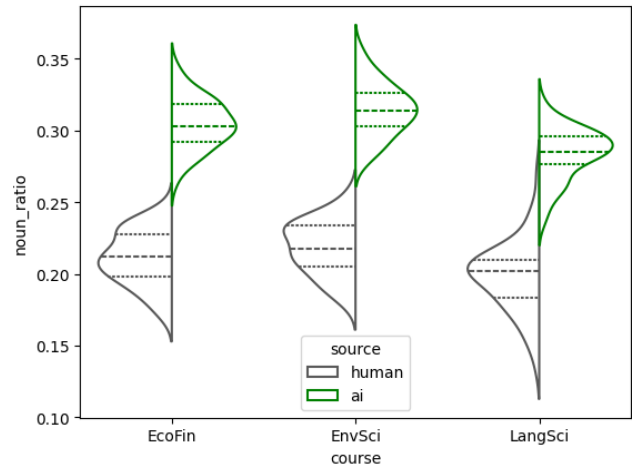
Regarding the year-based comparison across human and AI theses, Dunn's post-hoc test revealed that theses from each year in the human-written text condition differs significantly from the theses from each year in the AI-generated theses group ($p < .001$). The absence of specific divergences among human-written and AI-written theses denotes an overall statistically significant difference in the noun ratios between these groups, which can be visualized in the figures.

4.1.6.3.2. Course-Based Comparison

The analysis of the course variable demonstrates that, again, AI-generated theses have a higher noun ratio than human-written ones for each micro area. The statistical tests conducted on both human-only and AI-only groups reveal that the Language Sciences area differs from the other courses the most in terms of noun ratio (human-only: $p = .0164$ with Economics and



Course-based noun ratio values of human- and AI-written theses (box plot)



Course-based noun ratio values of human- and AI-written theses (violin plot)

Finance, and $p < .001$ with Environmental Sciences; AI-only: $p < .001$ with both courses), having considerably lower values than other courses.

Regarding the comparison across human and AI theses altogether, Dunn's post-hoc test showed that all theses from all courses in the human condition differs significantly from all

theses from all courses in the AI condition ($p < .001$). Besides, two specific pairs in the AI-generated theses group were shown to differ from one another, which are Environmental Sciences and Language Sciences areas ($p < .01$), the former having a higher proportion of nouns than the latter, as can be viewed on the figure above.

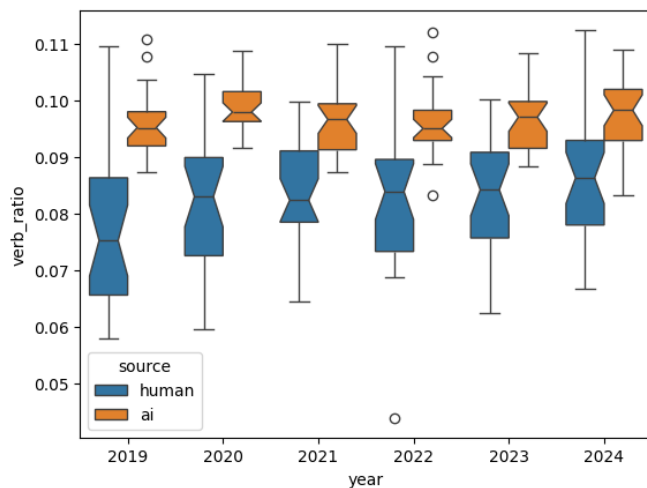
4.1.6.3.3. Source-Based Comparison

The comparison of the noun ratios across the two sources demonstrate that the AI-generated texts consistently have a higher proportion of nouns than human-authored theses, which is statistically significant ($W = 0.0$, $p < .001$, $r = 1.0$, Direction: AI > Human).

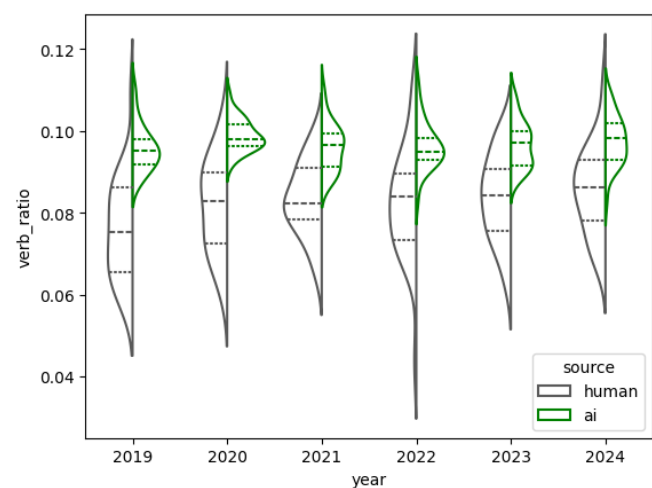
4.1.6.4. Verb Ratio

4.1.6.4.1. Year-Based Comparison

The analysis of the verb ratio in human-written and AI-generated texts show that the latter surpass the former in terms of the overall proportion of verbs, consistently in each year.



Year-based verb ratio values of human- and AI-written theses (box plot)



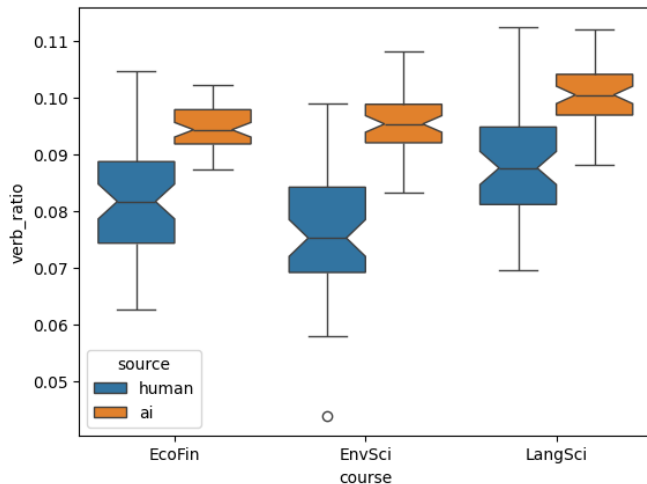
Year-based verb ratio values of human- and AI-written theses (violin plot)

The Kruskal-Wallis and Dunn's post-hoc test reveal an overall statistically significant difference between the verb ratios of human- and AI-written theses among all pairs ($p < .01$). Another result that can be viewed from the figures is that the verb ratios of human-written theses

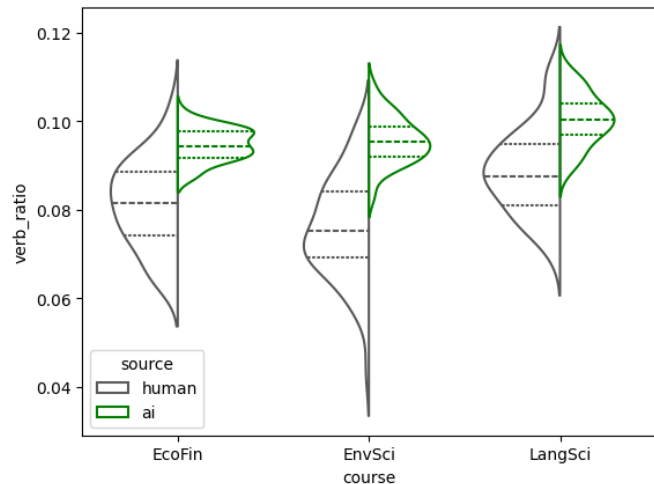
tend to have a larger range of variance in each year than the AI-theses, where the values concentrate in a more compact range.

4.1.6.4.2. Course-Based Comparison

The results of the comparison of different courses reveal that both in human-only and AI-only conditions, the Language Sciences show the highest divergence from other courses with a



Course-based verb ratio values of human- and AI-written theses (box plot)



Course-based verb ratio values of human- and AI-written theses (violin plot)

higher verb ratio, confirmed by the statistically significant p-values (human-only: $p < .01$ for both courses; AI-only: $p < .001$ for both courses). The results are considerably similar when the course variable is compared in the human and AI-combined group, where, again, in addition to the overall differences between the human and AI theses in all courses, the theses in the Language Sciences area (both human and AI) especially differ from four comparison pairs out of five ($p < .001$, except for the non-significant pairwise comparison with the AI-generated theses from Environmental Sciences, where $p = .128$).

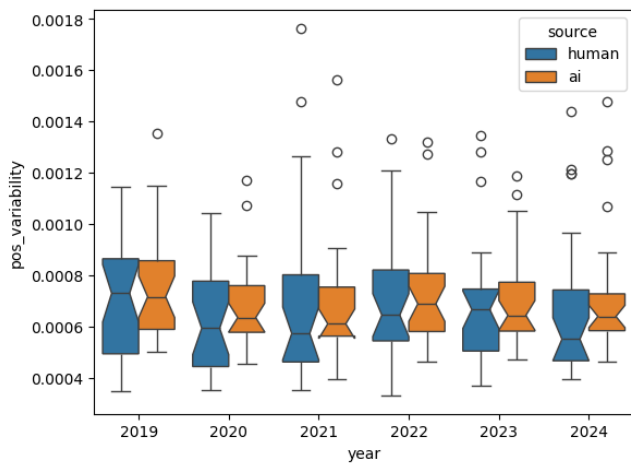
4.1.6.4.3. Source-Based Comparison

The proportion of verbs were shown to be substantially higher in LLM-generated theses as a result of the comparison across the two sources ($W = 257.0$, $p < .001$, $r = 0.960$, Direction: AI > Human).

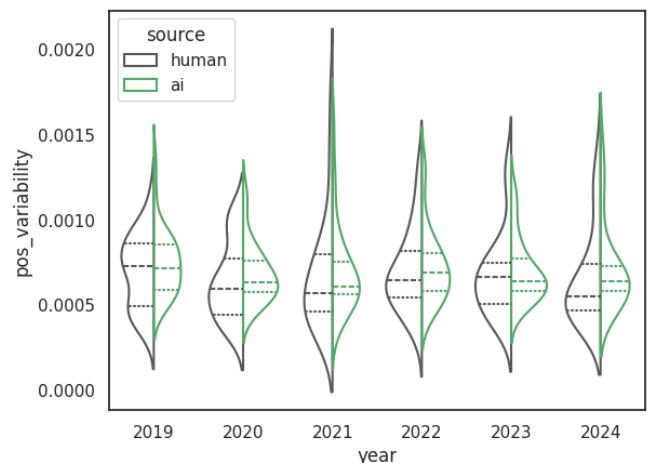
4.1.7. Part-of-Speech Diversity

4.1.7.1. Year-Based Comparison

The analysis results for the part-of-speech diversity feature did not show any considerable year-based differences, as can be seen from the figures. The Kruskal-Wallis test further confirmed that there is not any statistically significant difference between the year-by-year results of any of the source-based comparisons (human-only: $H = 2.757$, $df = 5$, $p = .737$; AI-only: $H = 3.337$, $df = 5$, $p = .648$; human vs. AI: $H = 11.476$, $df = 11$, $p = .404$).



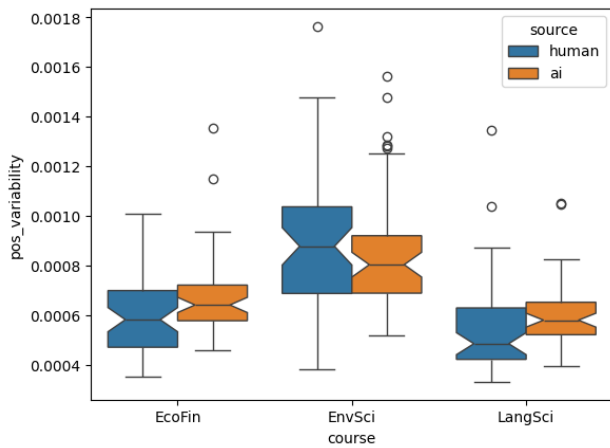
Year-based part-of-speech diversity values of human- and AI-written theses (box plot)



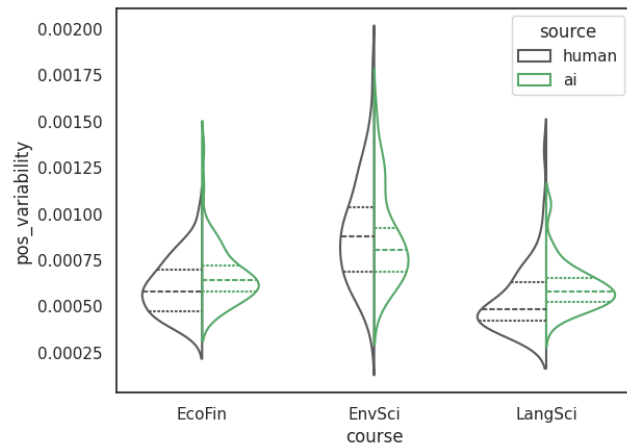
Year-based part-of-speech diversity values of human- and AI-written theses (violin plot)

4.1.7.2. Course-Based Comparison

As can be seen from the figures below and according to the results of the statistical analysis, the Environmental Sciences theses in both human-written and AI-generated condition differed significantly from the theses from other courses ($p < .001$), except for a non-significant pairwise difference between LLM-generated and human-written Environmental Sciences theses ($p = 1.000$). The aligning distribution of the theses from both sources for this specific area is displayed on the plots. In addition, it can be observed that the human-written and AI-generated comparisons for each course show a similar density distribution.



Course-based part-of-speech diversity values of human- and AI-written theses (box plot)



Course-based part-of-speech diversity values of human- and AI-written theses (violin plot)

4.1.7.3. Source-Based Comparison

The overall comparison across human- and AI-written theses pairs revealed that AI-generated master's theses have a higher part-of-speech diversity than their human-written counterparts ($W = 3638.0$, $p < .001$, $r = 0.44$, Direction: AI > Human).

4.1.8. LLM-Associated Word Frequency

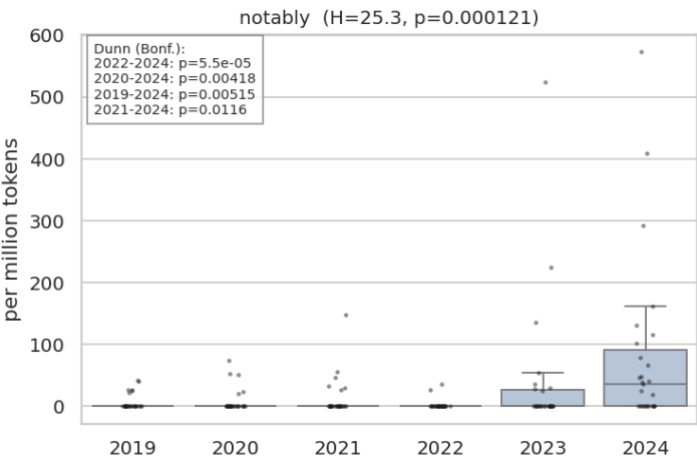
For the analysis of the word frequencies of LLM-associated, only human-written master's theses were taken into consideration. Since LLM-generated theses were all generated in 2026 and because they are only thematically connected to the human-written theses, solely the human-authored theses could provide insights into diachronic changes in word usage.

For a more robust comparison, the relative frequency for each word was calculated by dividing the occurrence of each LLM-associated word by the total number of tokens in the relevant thesis. For further consolidation, the frequencies were normalized per million tokens. Moreover, as in the previous linguistic features, the statistical significance was checked through the Kruskal-Wallis test, and Dunn’s post-hoc test with Bonferroni correction. Dunn’s test was conducted only on the significant groups detected in the Kruskal-Wallis test. All the Bonferroni-adjusted p-values that were achieved in the Dunn’s pairwise comparison test are reported on the plots. It is also important to mention that the columns (i.e. LLM-associated words) that solely consisted of zero values were removed, which were in total 22 columns¹⁰ (the list of words that did not occur in the present corpus are provided in Appendix D).

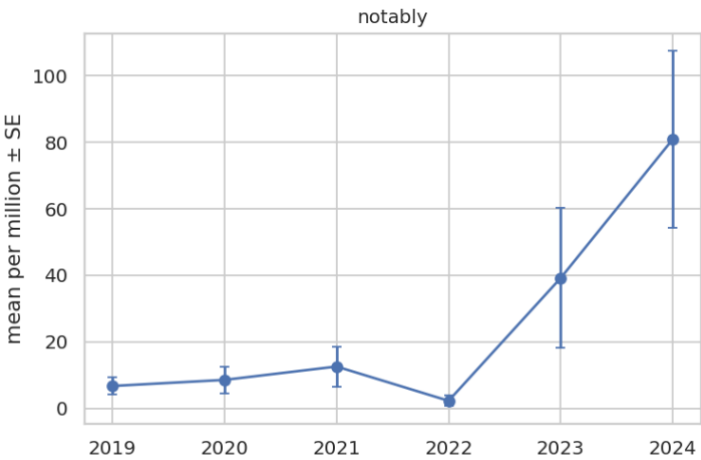
As a result of the analysis, the words whose occurrence showed a statistically significant difference between the pre-ChatGPT and ChatGPT era were “across”, “additionally”, “among”, “beyond”, “challenges”, “comprehensive”, “comprehensively”, “cultural”, “distinctive”, “effective”, “effectively”, “essential”, “foundational”, “greater”, “including”, “inherent”, “innovative”, “insights”, “intricate”, “markedly”, “notable”, “notably”, “ongoing”, “operational”, “particularly”, “pivotal”, “potent”, “practices”, “predominantly”, “primarily”, “seamless”, “seamlessly”, “strategic”, “subsequently”, “thereby”, “thoughtful”, “universally”, and “valuable”. The relative frequency results for each of these words were higher in 2024.

¹⁰ The zero results were expected, considering that the comparison list was taken from Liang et al. (2024) and Sprung (2025). Since the researchers reported their own LLM-associated word lists, it is natural that not each token is attested in the current study.

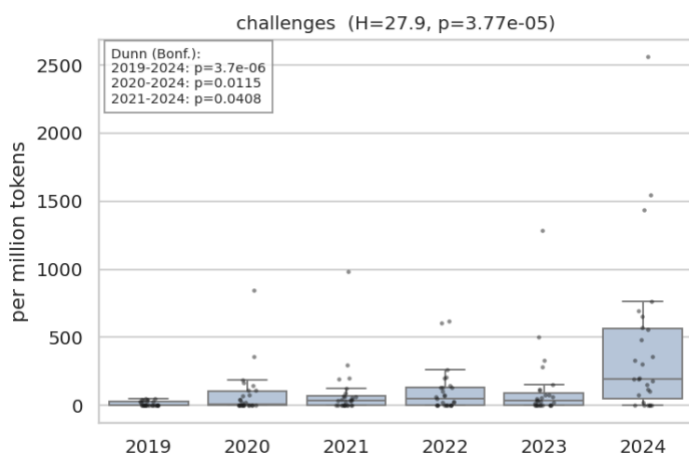
The plots displaying the top 10 words with the highest statistical difference in the pre- and post-ChatGPT diachronic comparison are provided below. It should be noted that each of these words showed their highest frequency in 2024, aligning with most of the significant year-based results of the previous linguistic features, which also showed a likely convergence with AI-generated theses values in the year 2024.



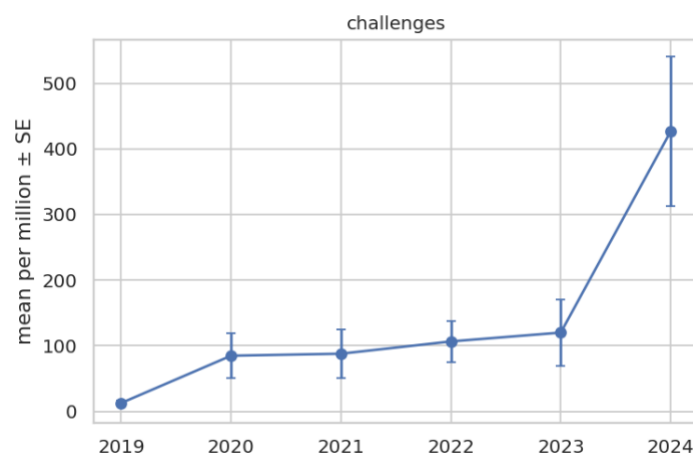
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "notably"



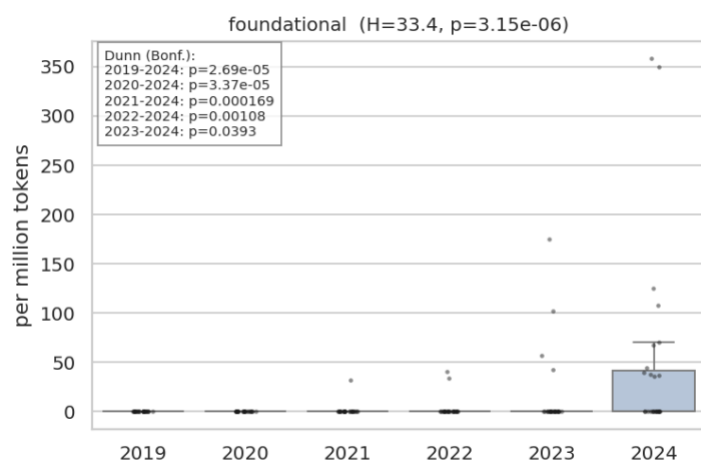
Year-based relative frequency change for the occurrence of the word "notably"



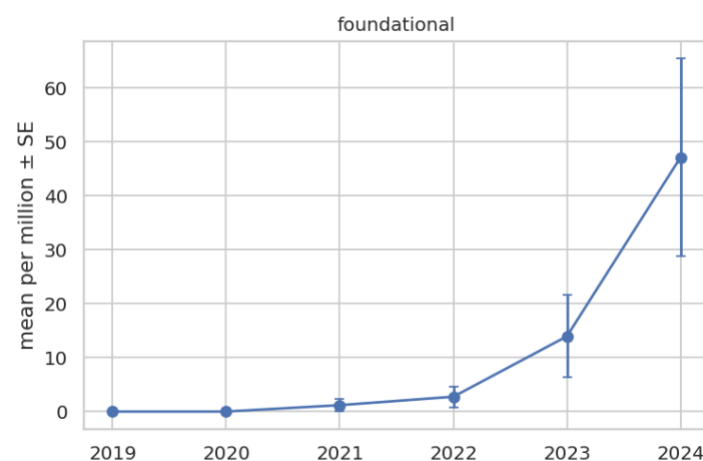
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "challenges"



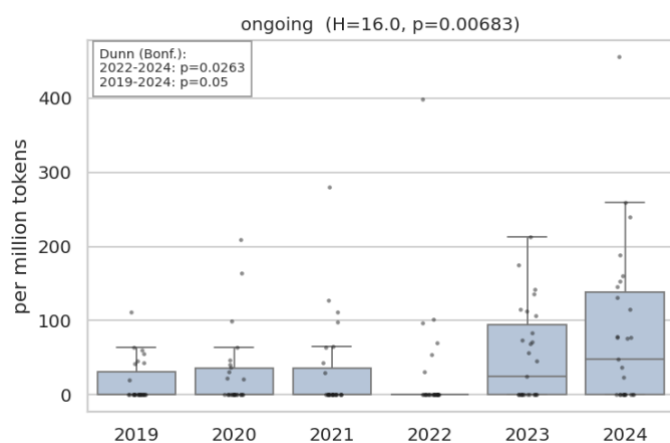
Year-based relative frequency change for the occurrence of the word "challenges"



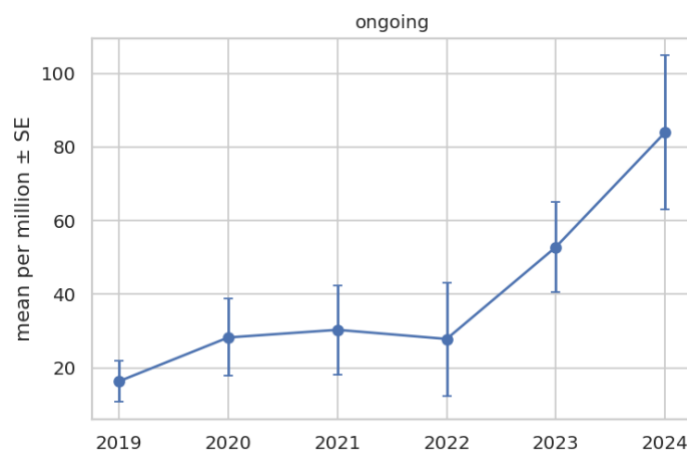
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "foundational"



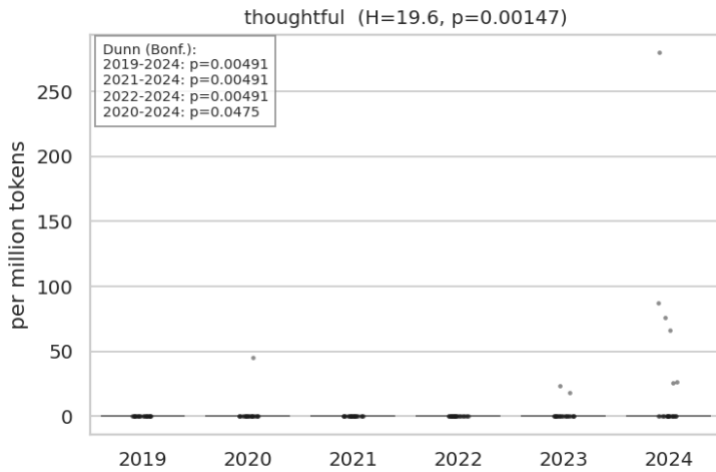
Year-based relative frequency change for the occurrence of the word "foundational"



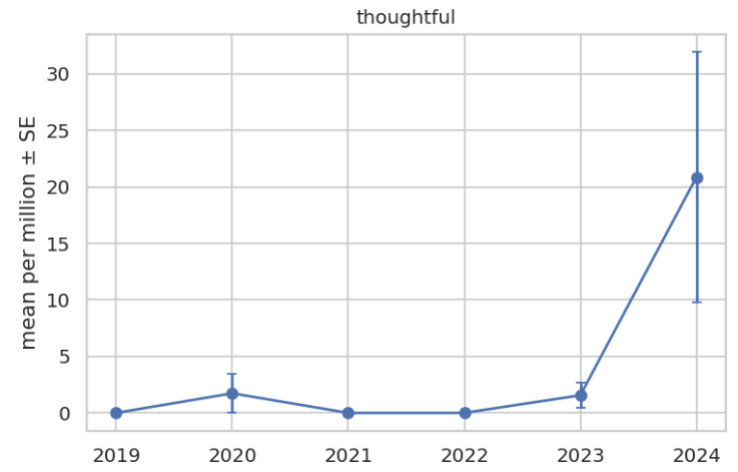
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "ongoing"



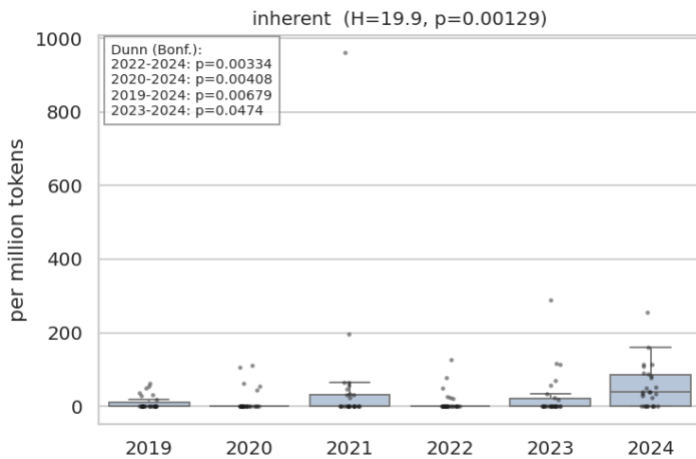
Year-based relative frequency change for the occurrence of the word "ongoing"



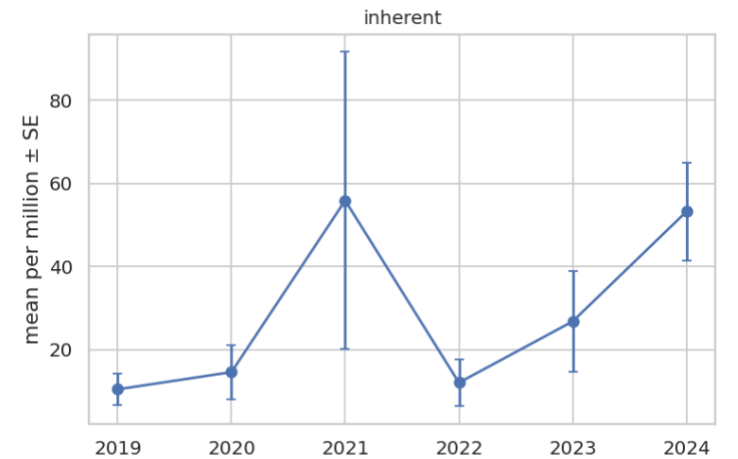
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "thoughtful"



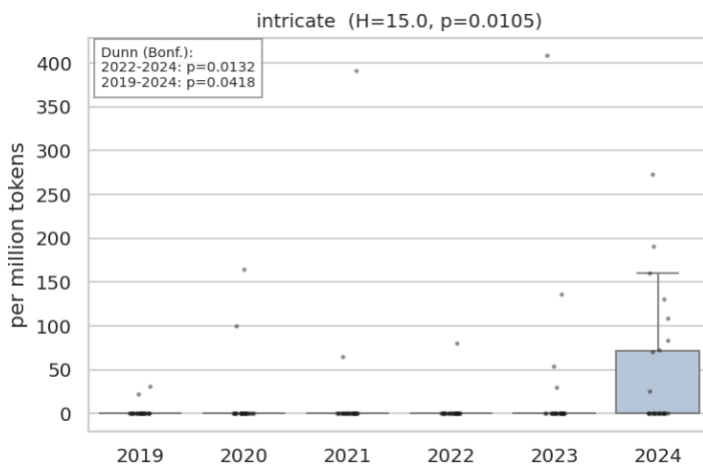
Year-based relative frequency change for the occurrence of the word "thoughtful"



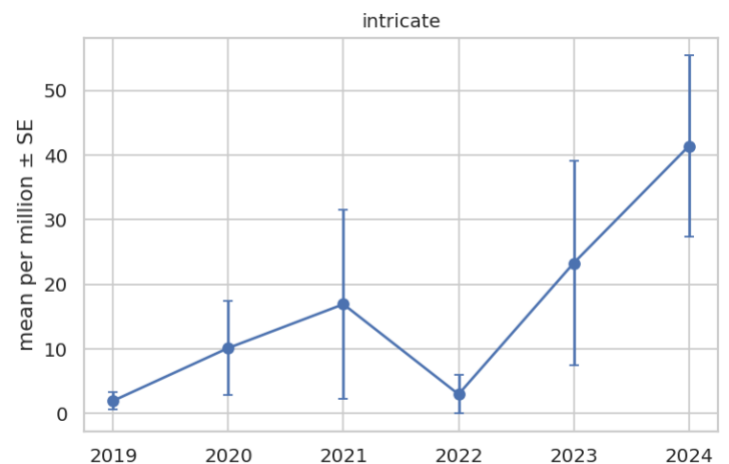
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "inherent"



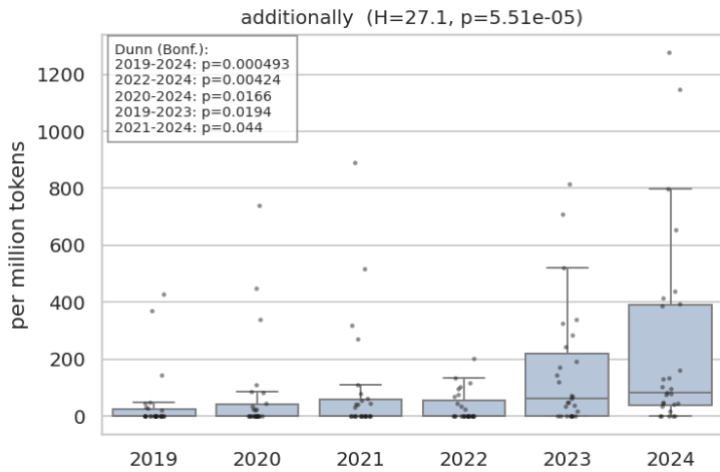
Year-based relative frequency change for the occurrence of the word "inherent"



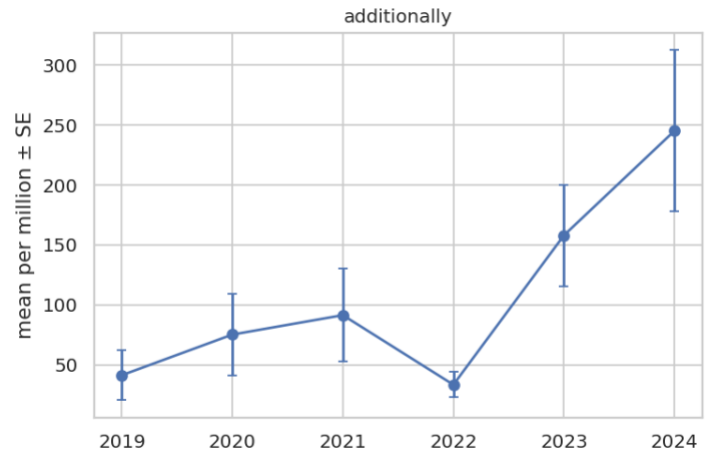
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "intricate"



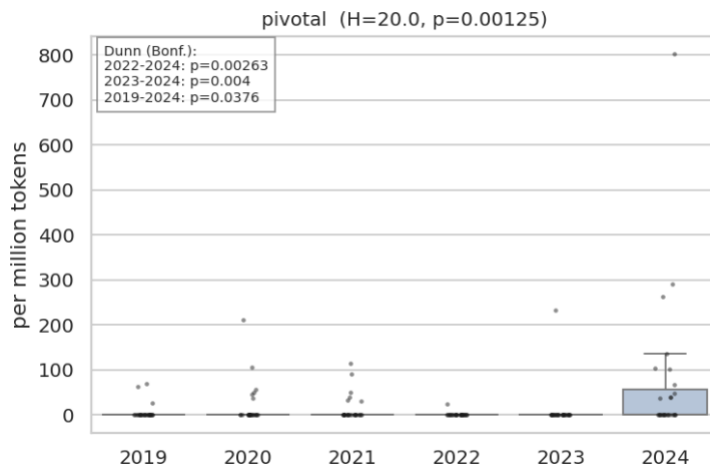
Year-based relative frequency change for the occurrence of the word "intricate"



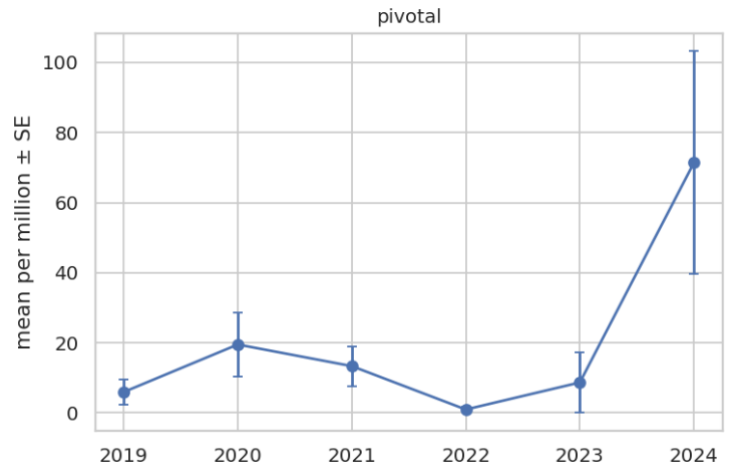
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "additionally"



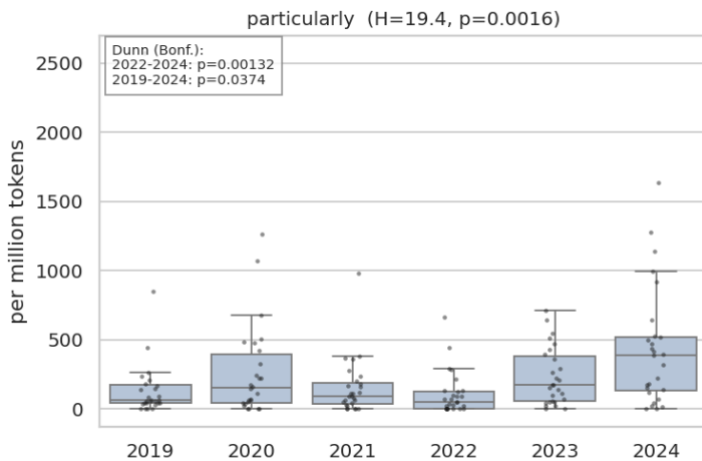
Year-based relative frequency change for the occurrence of the word "additionally"



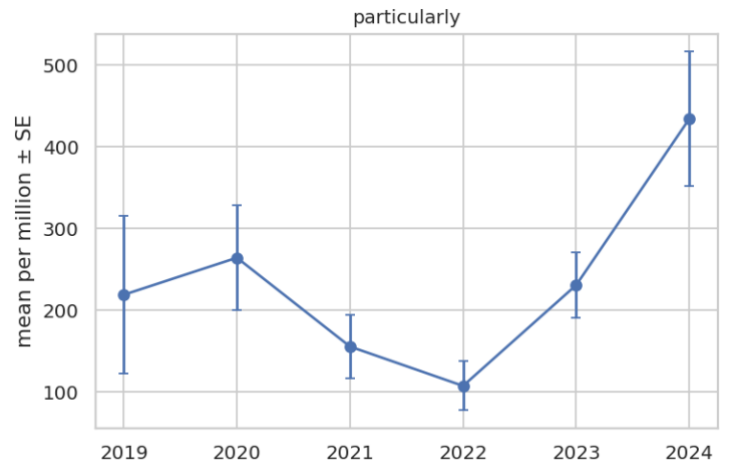
Dunn's pairwise post-hoc test results for the year-based occurrence of the word "pivotal"



Year-based relative frequency change for the occurrence of the word "pivotal"



Dunn's pairwise post-hoc test results for the year-based occurrence of the word "particularly"



Year-based relative frequency change for the occurrence of the word "particularly"

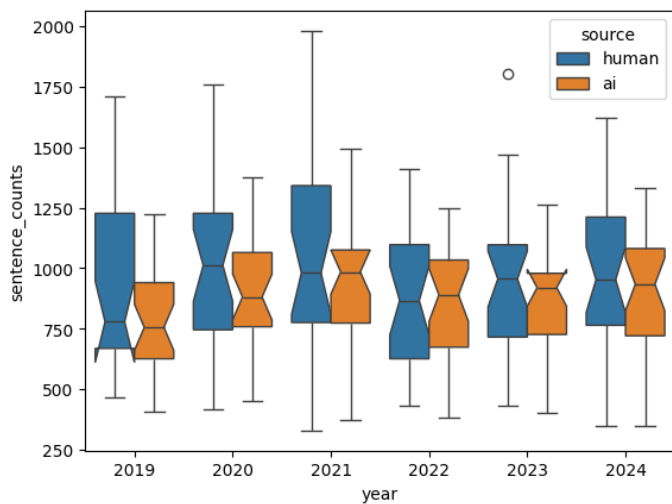
4.2.Syntactic Features

4.2.1. Sentence Counts

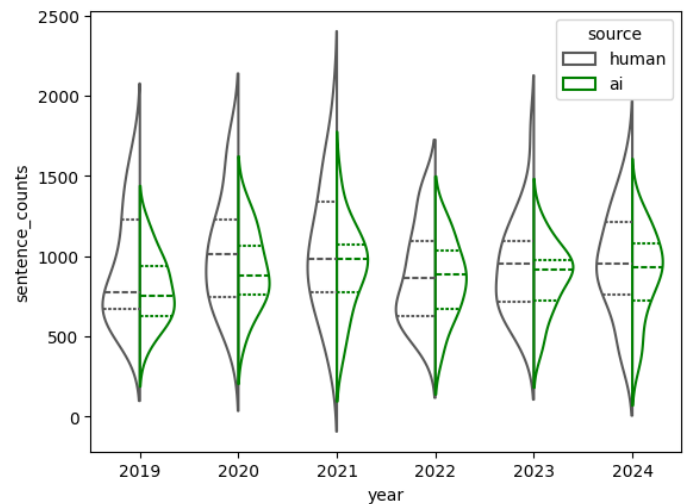
4.2.1.1.Year-Based Comparison

The results of the year-based comparison of sentence counts of human-written and AI-generated master's theses show that the former have a higher number of sentences overall than the latter.

The distribution of the values that can be seen on the figures signify that human-written theses have a larger range of variety in sentence count values while the AI-theses are slightly more homogenous in terms of value distribution.



Year-based sentence count values of human- and AI-written theses (box plot)

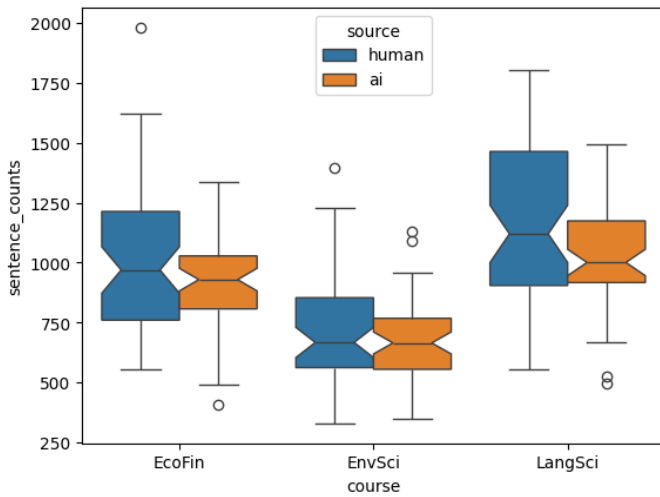


Year-based sentence count values of human- and AI-written theses (violin plot)

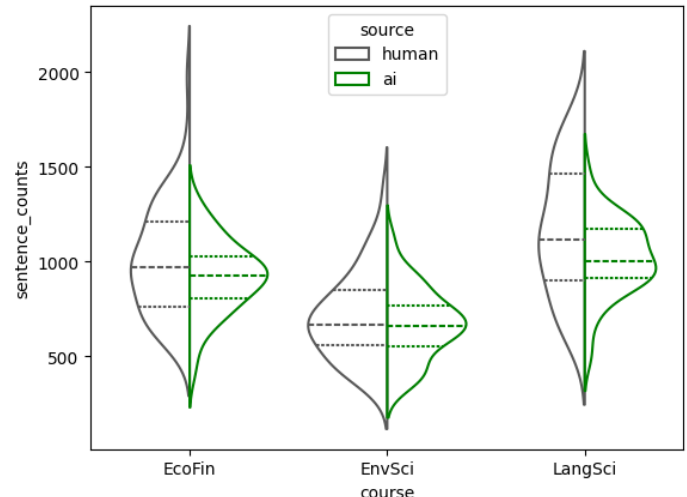
The statistical analysis tests that were carried out on human-only, AI-only, and human and AI-combined groups did not demonstrate statistically significant results in terms of the year-based variation in the values (human-only: $H = 4.630$, $df = 5$, $p = .462$; AI-only: $H = 5.549$, $df = 5$, $p = .352$; human and AI: $H = 12.751$, $df = 11$, $p = .309$).

4.2.1.2.Course-Based Comparison

The comparison of the values of the human-written and AI-generated theses based on the course variable exhibits that human-written and AI-written theses show similar median values in terms of sentence counts. In addition, the values of human-authored texts show a more varied distribution as opposed to the AI-generated master's theses which show a more compact range.



Course-based sentence count values of human- and AI-written theses (box plot)



Course-based sentence count values of human- and AI-written theses (violin plot)

The Kruskal-Wallis test found an overall statistically significant difference between the human-authored and GPT-generated theses for the course variable ($H = 110.181$, $df = 5$, $p < .001$). Therefore, Dunn's post-hoc test was applied in order to explore the specific differences among the groups. This pairwise test yielded valuable results; in that, the theses written in the area of Environmental Sciences in the human-only, AI-only and human and AI-combined groups all showed a clear divergence in the pairwise comparisons ($p < .001$). The most interesting result is that the only pair that did not show a statistically significant difference is the human-written Environmental Sciences theses and AI-generated Environmental Sciences theses ($p = 1.000$).

4.2.1.3.Source-Based Comparison

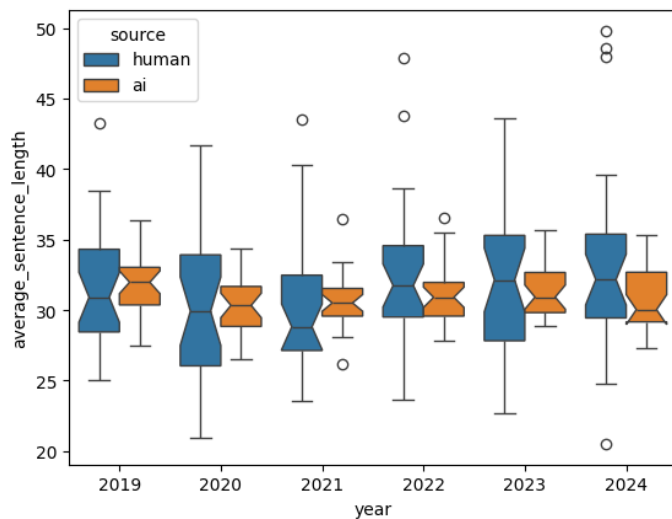
The paired Wilcoxon signed rank test found the differences between the sentence counts of LLM-generated and human-written theses to be significant, with the latter showing

substantially higher values than the former ($W = 3695.5$, $p < .001$, $r = -0.43$, Direction: Human > AI).

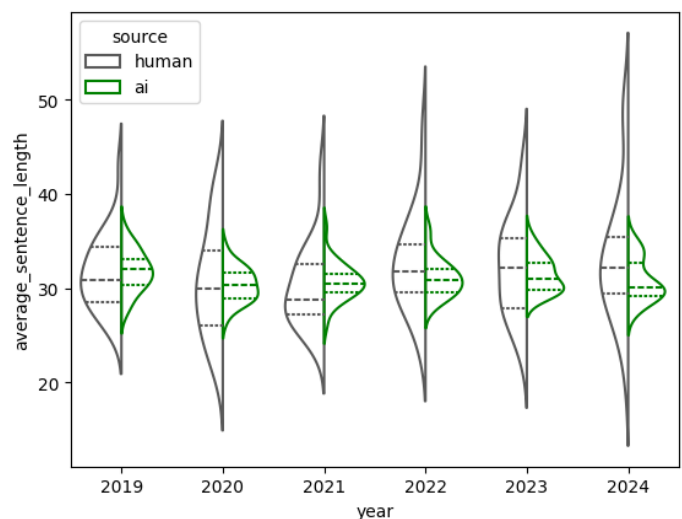
4.2.2. Average Sentence Length

4.2.2.1. Year-Based Comparison

The data reveals that the distribution of the average sentence lengths of the human-written theses are situated more towards higher values than AI-generated theses in each year overall. Nevertheless, the data neither displays considerable differences between the median values of human- and AI-texts, nor discrepancies in the values observed in each year for human-written theses only. Regarding the range of variation in the values, it can be observed that the values of AI-generated theses in each year display a more concentrated distribution while the human-written theses have a larger range of variety in the sentence length values.



Year-based average sentence length values of human- and AI-written theses (box plot)

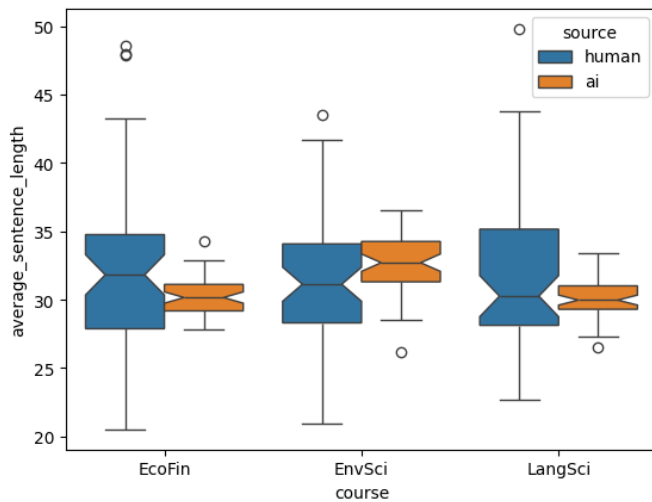


Year-based average sentence length values of human- and AI-written theses (violin plot)

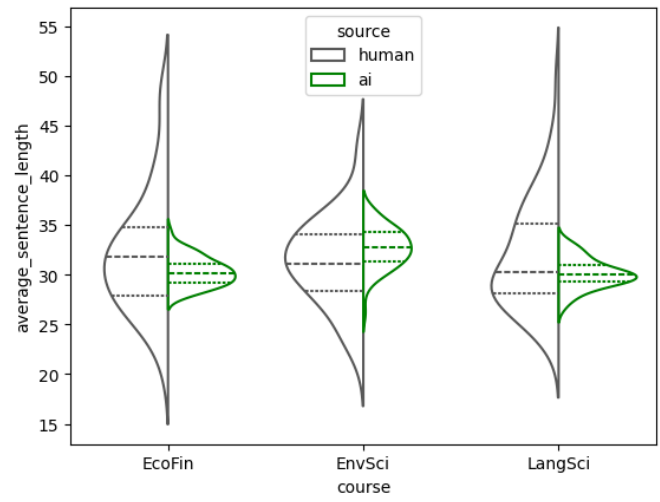
The Kruskal-Wallis demonstrated that there is not any statistically significant difference inside the human-only ($H = 6.372$, $df = 5$, $p = .271$), AI-only ($H = 9.191$, $df = 5$, $p = .101$) and human and AI-combined groups ($H = 14.708$, $df = 11$, $p = .196$) for the year variable.

4.2.2.2. Course-Based Comparison

Regarding the course-based differences in the human and AI-written theses, again, no overall remarkable differences in the median values appear, except for the Economics and Finance area, which exhibits a higher median in the human-written master's theses. The distribution of data for each course in the human-only group; on the other hand, seems to be similar to one another, while the values of the Environmental Sciences course in the AI-only group show a divergence from the other courses.



Course-based average sentence length values of human- and AI-written theses (box plot)



Course-based average sentence length values of human- and AI-written theses (violin plot)

This was further confirmed by the statistical analyses; in that, while there was no significant difference among the courses in the human-only condition ($H = 0.017$, $df = 2$, $p = .991$), the tests found a statistically significant divergence in the AI-only ($H = 50.119$, $df = 2$, $p < .001$) and human and AI-combined groups ($H = 30.198$, $df = 5$, $p < .001$), with the AI-written theses in the Environmental Sciences area showing higher median values in the pairwise comparisons (human-Language Sciences: $p = .0381$; AI-Economics and Finance: $p < .001$; AI-Language Sciences: $p < .001$).

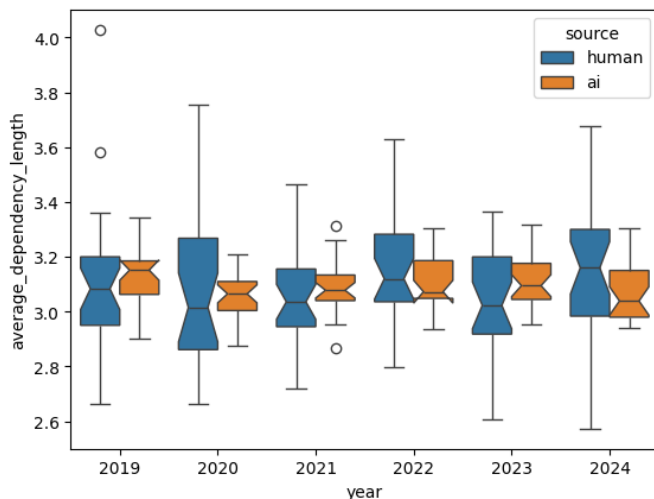
4.2.2.3.Source-Based Comparison

The further tests conducted to discover the overall value differences across the two sources demonstrated that human-written theses have higher average sentence lengths than their LLM-generated counterparts, which, however, was not statistically significant ($W = 5999.0, p = .3787, r = -0.08$, Direction: Human > AI).

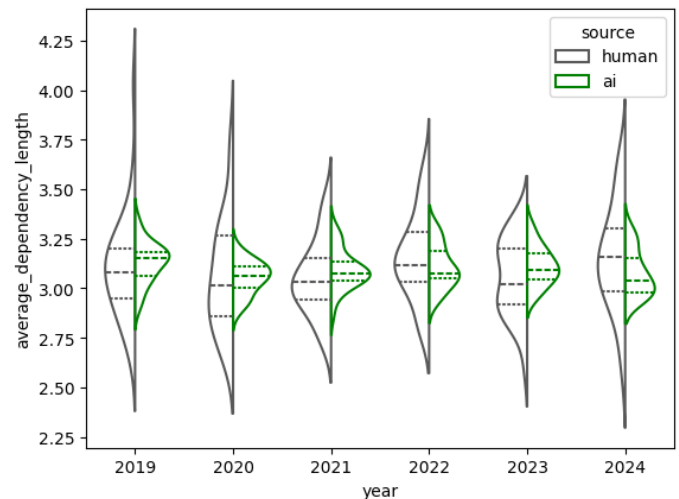
4.2.3. Average Dependency Length

4.2.3.1. Year-Based Comparison

The linguistic analysis results show that the median values of human-written and AI-generated theses do not follow a regular pattern. While the median values of human-authored theses are higher than LLM-generated ones in 2022 and 2024, the other years show the opposite. In addition, as in the previous syntactic features, the range between the lowest and the highest value for each year in human-written theses is larger than AI-generated theses. Regarding the year-by-year changes in human-authored texts, it could be observed from the figures that there does not appear to be a considerable difference between the average dependency length values of texts from each year.



Year-based average dependency length values of human- and AI-written theses (box plot)

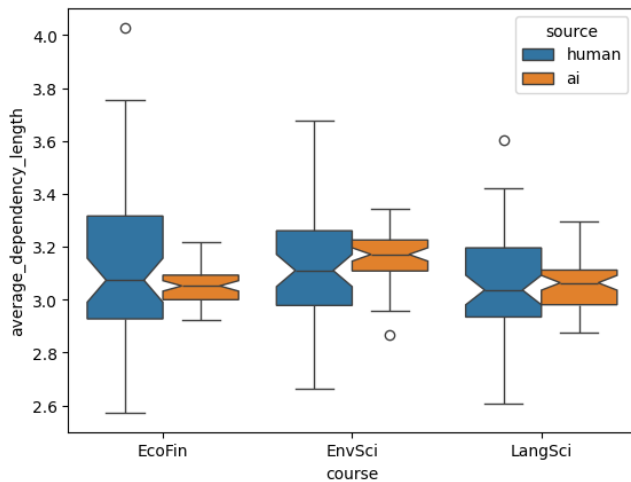


Year-based average dependency length values of human- and AI-written theses (violin plot)

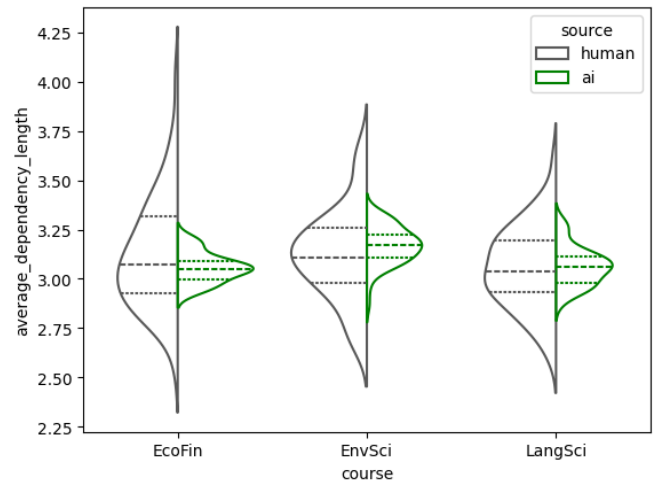
The Kruskal-Wallis test further confirms the data observed in the figures, revealing no statistically significant difference in the human-only ($H = 6.595$, $df = 5$, $p = .252$), AI-only ($H = 9.949$, $df = 5$, $p = .076$), and the human and AI-combined group ($H = 15.539$, $df = 11$, $p = .159$).

4.2.3.2.Course-Based Comparison

As to the comparison of the theses from different university programs, it can be observed that except for the slightly larger difference in the Economics and Finance area, the human-written theses and AI-generated theses did not show a considerable discrepancy in the medians of average length values, although human-theses have higher maximum values for each course. The data distribution, again, has a larger variety range in the case of human theses in each course than AI theses.



Course-based average dependency length values of human- and AI-written theses (box plot)



Course-based average dependency length values of human- and AI-written theses (violin plot)

The analyses show that the most divergent group which demonstrated statistically significant differences in almost all its pairwise comparisons (AI-only: $p < .001$; human vs. AI: $p < .05$) is the AI-generated Environmental Sciences theses group. It is important to note that it does not show the significant differences only with the human-written Environmental Sciences

theses ($p = .266$). For the human-only condition, no statistically significant differences were found in the course-based comparison ($H = 2.146$, $df = 2$, $p = .341$).

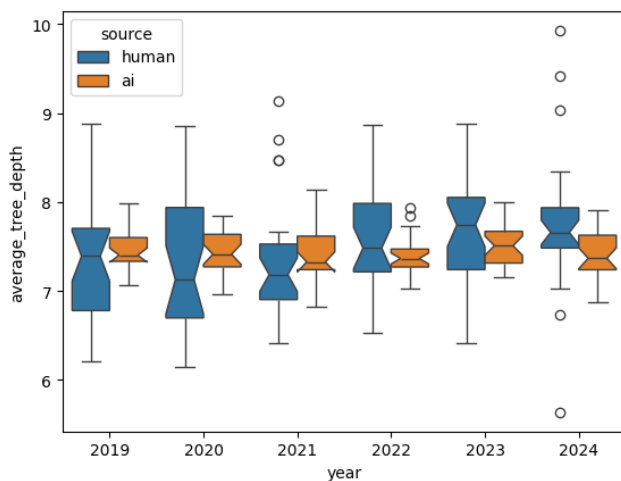
4.2.3.3.Source-Based Comparison

Like the overall comparison of the values across the two sources reported for the average sentence length, the human-authored theses were found to have higher dependency lengths than GPT theses, but the difference was not statistically significant ($W = 6227.0$, $p = .6203$, $r = 0.05$, Direction: Human > AI).

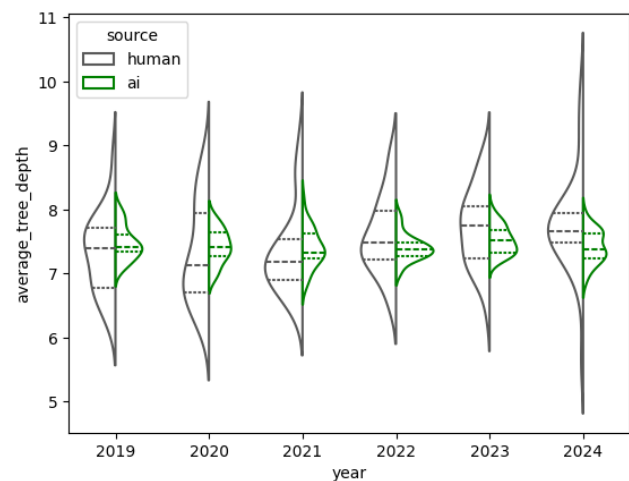
4.2.4. Average Tree Depth

4.2.4.1.Year-Based Comparison

As to the results of the average tree depth values, it can be observed that the year-based comparison does not reveal a regular, interpretable pattern, and no considerable difference



Year-based average tree depth values of human- and AI-written theses (box plot)



Year-based average tree depth values of human- and AI-written theses (violin plot)

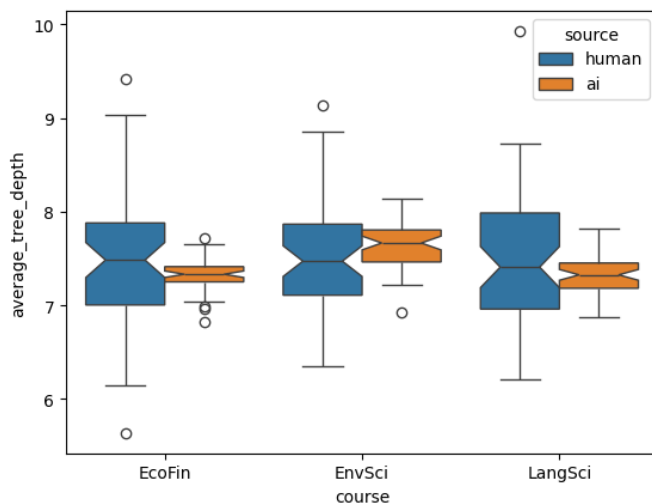
between the median values of human- and AI-written theses. The year-by-year comparison of the average tree depth values of human-written theses, however, appear to demonstrate variance between different years. Besides, the data distribution range of human-written theses is larger than the AI counterparts, as in the previously observed syntactic complexity features.

The results of the Kruskal-Wallis test for the human-only group reveal an overall statistically significant difference for the year-based values in the human-written group ($H = 15.625$, $df = 5$, $p = .0079$), while the Dunn's post-hoc test does not demonstrate any specific groups that differ from one another ($p > .05$ for all pairwise comparisons), which confirms the data viewable on the plots. This finding indicates that there is just an overall difference among the groups, and no specific pairwise discrepancies. There also were not any statistically significant differences in the theses from the AI-only group ($H = 4.294$, $df = 5$, $p = .507$).

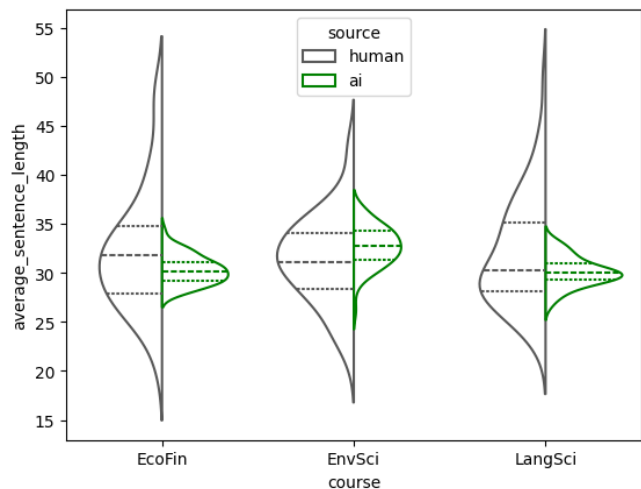
Regarding the year-based comparison of human and AI-combined group, interestingly, the only statistically differing pair was the human-written theses in 2021 (pre-ChatGPT) and 2024 (ChatGPT era) ($p = .0071$), which was not significant in the human-only group test.

4.2.4.2.Course-Based Comparison

The comparison of the course variable demonstrated that while the human-written theses



Course-based average tree depth values of human- and AI-written theses (box plot)



Course-based average tree depth values of human- and AI-written theses (violin plot)

from three courses do not considerably differ from one another ($H = 0.155$, $df = 2$, $p = .925$), the Environmental Sciences area in the GPT-generated theses show the highest divergence from the

other areas in the only-AI condition ($p < .001$ with both other courses) and almost all areas in the human and AI-combined condition ($p < .05$), exhibiting a higher median value.

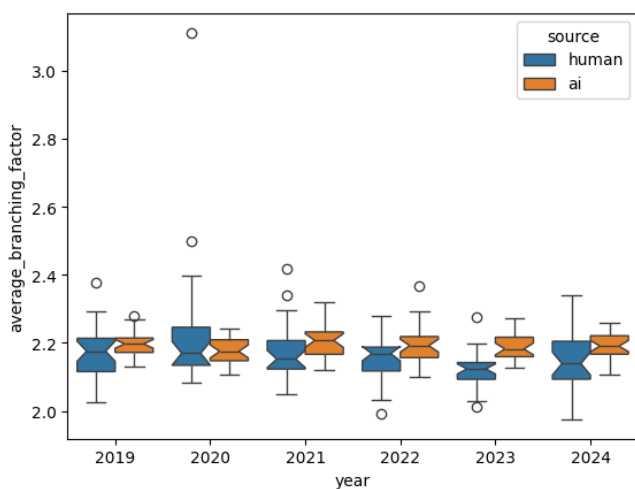
4.2.4.3. Source-Based Comparison

Similar to the average sentence lengths and average dependency lengths, the average tree depths of human-authored master's theses were revealed to be higher than the AI-generated counterparts, but this discrepancy was not found to be substantial ($W = 6130.0$, $p = .5098$, $r = -0.06$, Direction: Human $>$ AI).

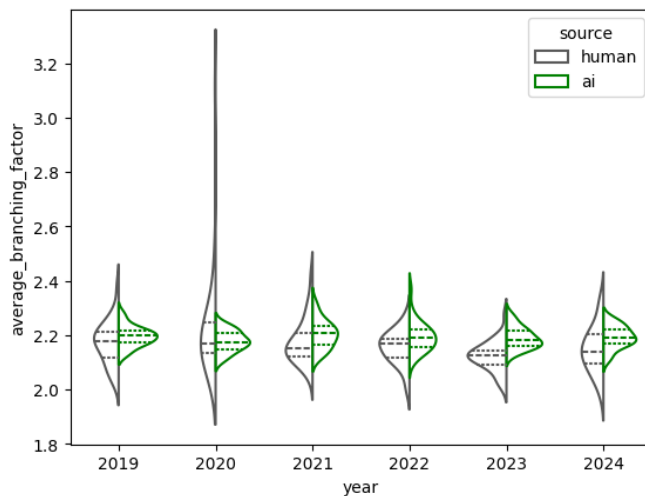
4.2.5. Average Branching Factor

4.2.5.1. Year-Based Comparison

The linguistic analysis for the average branching factor reveals different results from other syntactic results, exhibiting that AI-generated median values notably surpass the human-



Year-based average branching factor values of human- and AI-written theses (box plot)



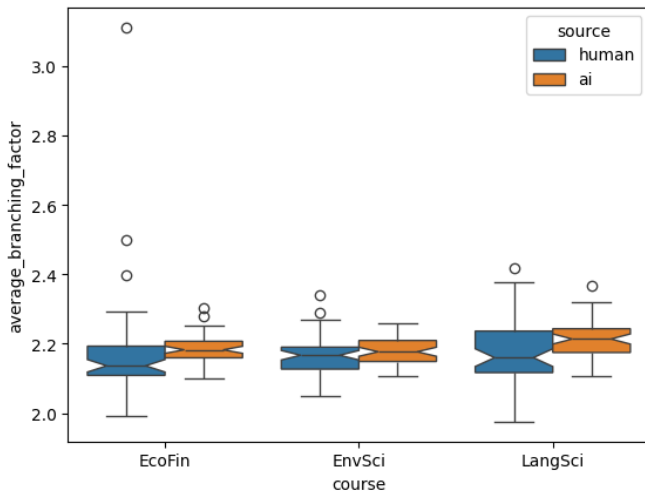
Year-based average branching factor values of human- and AI-written theses (violin plot)

written ones in most of the years, except for 2019 and 2020. The range of values for the human-authored theses show a larger distribution, although less remarkably than in the previous syntactic features.

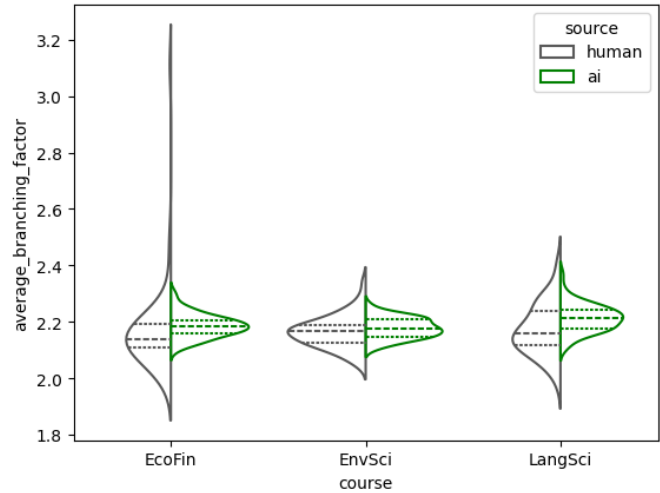
The statistical analyses reveal a significant difference between the values of the human-written theses in 2020 (pre-ChatGPT) and 2023 (ChatGPT era) ($p = .0201$), whereas the AI-only condition does not have any overall statistically significant differences among the year-based values ($H = 3.869$, $df = 5$, $p = .568$). The human and AI-combined group reveals significant discrepancies between the values ($H = 43.066$, $df = 11$, $p < .001$), the human-written theses in 2023 being the most divergent with considerably lower values, differing with almost all AI-generated theses ($p < .01$). The only pairwise comparison in which 2023 human-theses do not exhibit a difference with the LLM-generated theses is the year 2020 ($p = .082$).

4.2.5.2.Course-Based Comparison

When it comes to the course variable, based on the figures derived from the analysis results, it appears that course-based values in both human-written and AI-generated theses seem to be similar, except for the AI-written Language Sciences theses, which are slightly more different from the other courses in the AI-generated group.



Course-based average branching factor values of human- and AI-written theses (box plot)



Course-based average branching factor values of human- and AI-written theses (violin plot)

The statistical analysis confirms the data visualized on the figures. While the human-only group does not show course-based variance among the theses ($H = 2.619$, $df = 2$, $p = .269$), the

values of the Language Sciences theses in the AI-only condition significantly differ from other courses ($p < .01$). Regarding the course-based comparison across human and AI-theses, it was found that AI-written Language Sciences theses substantially differ from four comparison groups out of five ($p < .01$). Another significantly differing value was also found between the human-written Economics and Finance theses and GPT-generated Economics and Finance texts ($p = .0147$).

4.2.5.3.Source-Based Comparison

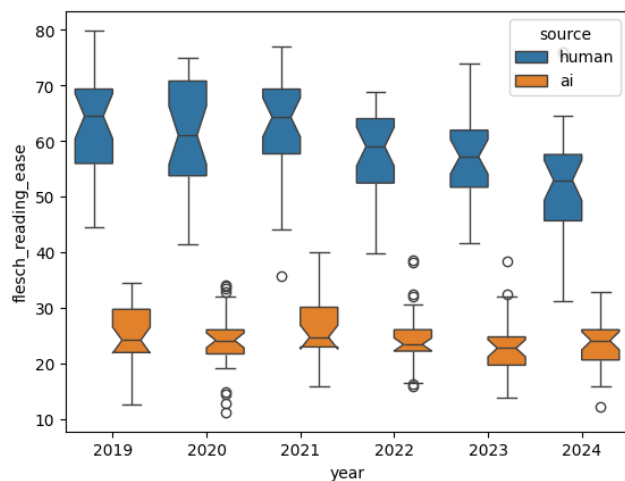
The discrepancies between the human-written and AI-generated theses values were found to be statistically significant in the paired Wilcoxon signed rank test, with the latter showing higher values overall ($W = 4154.0$, $p < .001$, $r = 0.36$, Direction: AI > Human).

4.3.Readability Features

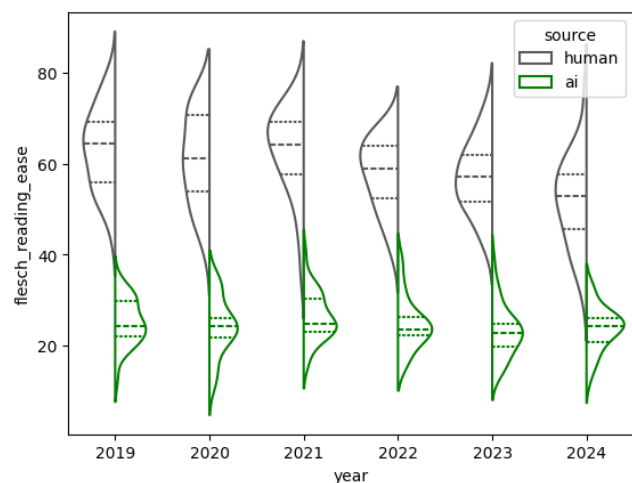
4.3.1. Flesch Reading Ease

4.3.1.1.Year-Based Comparison

It can be observed in the figures that human-written theses in each year have a higher readability ease score than the AI-generated theses. As to the year-by-year comparison among human-written theses, it could be seen that the theses in years 2022, 2023 and 2024 show a lower



Year-based Flesch readability ease values of human- and AI-written theses (box plot)

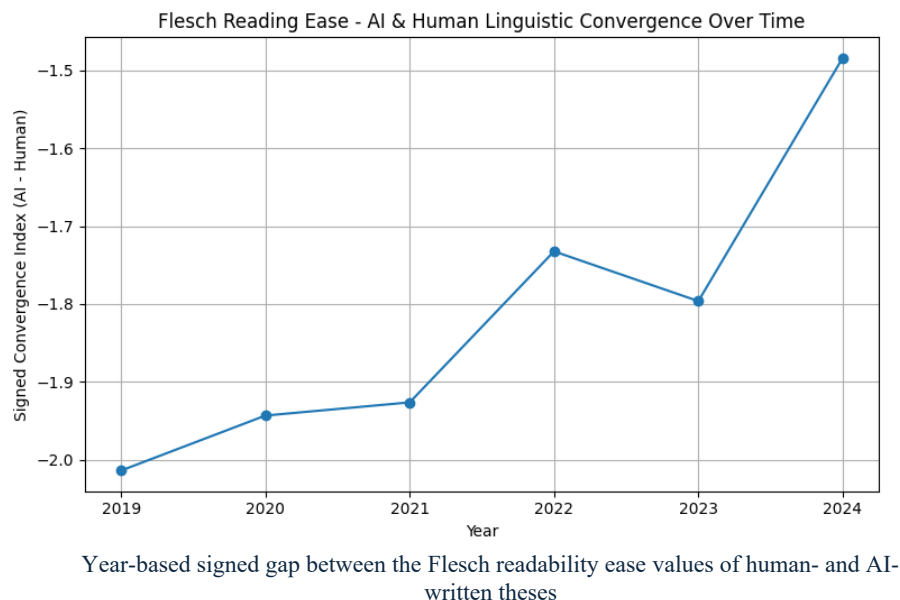


Year-based Flesch readability ease values of human- and AI-written theses (violin plot)

readability score, being closer to LLM-generated theses values, in comparison with the pre-ChatGPT years which differ more from the values of GPT-generated theses.

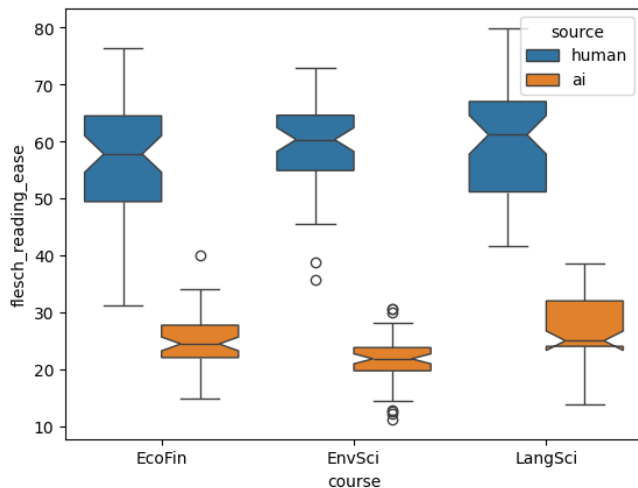
The statistical analyses confirm the data visible on the figures, revealing that the year pairs 2019-2024 ($p < .001$), 2020-2024 ($p = .0202$), and 2021-2024 ($p < .001$) in human-written theses are significantly different. These results denote a discrepancy between theses from pre-ChatGPT years and theses from 2024 (ChatGPT era). As to the AI-written theses, no overall considerable difference was found in the year-based comparison ($H = 5.737$, $df = 5$, $p = .332$). Regarding the year-based comparison across sources, human-written theses in each year were shown to significantly differ from GPT-generated theses from each year ($p < .001$ for all pairs).

The signed gap between the year-based results for the two sources demonstrated that the gap between the values of human- and AI-written theses becomes narrower in 2024. In other words, the human and GPT theses have the most similar Flesch reading ease scores in 2024 (ChatGPT era).

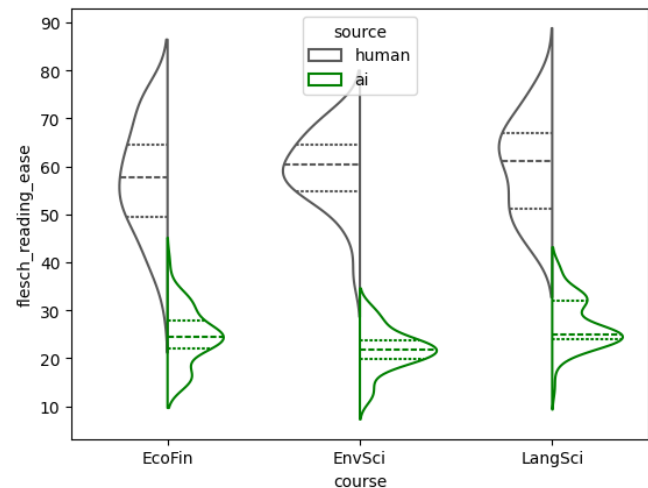


4.3.1.2.Course-Based Comparison

In the plots displaying the comparison of human and AI-theses across sources, it can be seen that human-written theses consistently have higher reading ease scores than AI-written theses in all courses. In addition, the distribution of the values shows a larger range in human-theses, while the distribution of the AI-theses concentrates on closer values.



Course-based Flesch readability ease values of human- and AI-written theses (box plot)



Course-based Flesch readability ease values of human- and AI-written theses (violin plot)

Regarding the statistical analyses, the human-only group did not yield significant discrepancies among courses ($H = 1.399$, $df = 2$, $p = .496$), whereas the readability ease values of Environmental Sciences theses in the AI-only group considerably differed from other courses ($p < .01$ in both pairs), having the lowest readability ease. As to the course-based comparison across sources, the results exhibit significant differences between all groups of human-written theses and all groups of AI-generated theses ($p < .001$), once again showing the overall discrepancy between the readability ease of human and AI writing style.

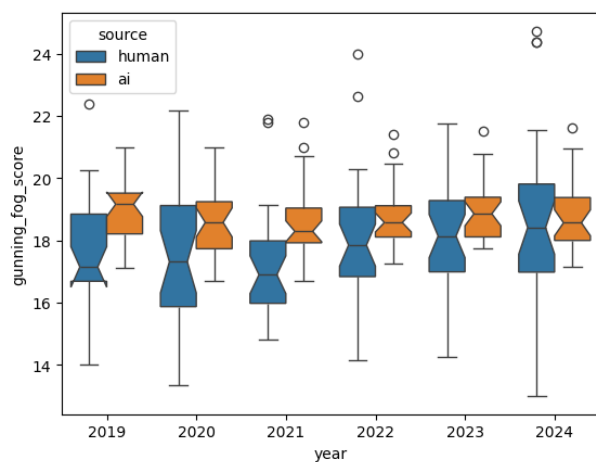
4.3.1.3.Source-Based Comparison

The readability ease scores that were analyzed in each human and AI thesis pair exhibited that the human-authored master's theses have a higher ease of readability, showing considerably higher Flesch reading ease values than their AI-counterparts ($W = 0.0, p < .001, r = -1.0$, Direction: Human > AI).

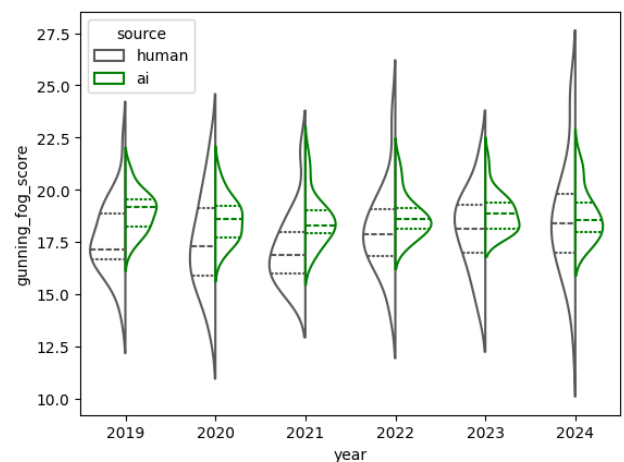
4.3.2. Gunning Fog Index

4.3.2.1. Year-Based Comparison

According to the figures which were achieved from the linguistic analysis results, the gunning fog scores seem to be higher in the AI-generated group, although it is not consistently so



Year-based Gunning fog index values of human- and AI-written theses (box plot)

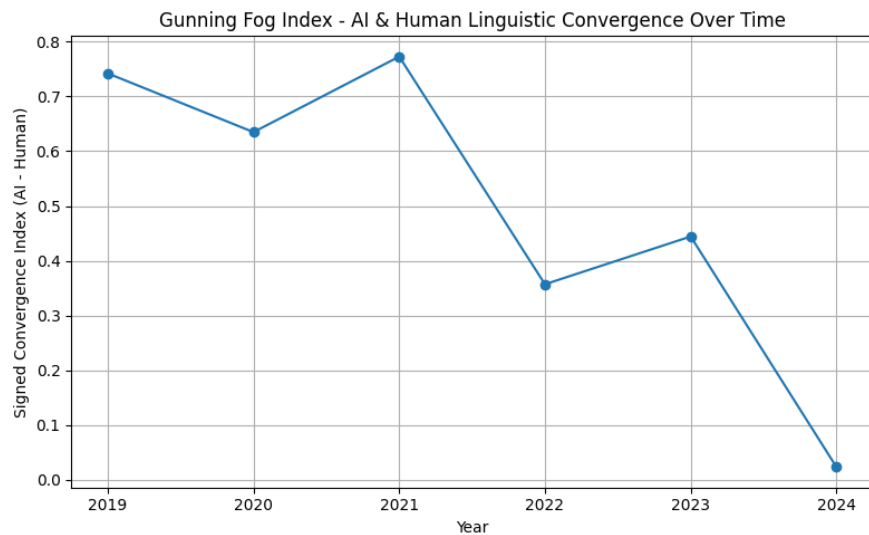


Year-based Gunning fog index values of human- and AI-written theses (violin plot)

for each year. At this stage, the statistical analyses exhibit clearer results. According to the Kruskal-Wallis test, the human-only ($H = 8.790, df = 5, p = .117$) and AI-only group ($H = 4.465, df = 5, p = .484$) do not display significant year-based differences. With respect to the year-based comparison across sources, it was found that the human-written theses in 2019, 2020 and 2021 differ significantly from LLM-generated theses in 2019 (human-2019: $p = .0435$; human-2020: $p = .0402$; human-2021, $p < .001$). Moreover, the human-authored theses in 2021 are considerably

diverse from the AI theses in 2022 ($p = .0245$), 2023 ($p = .0026$) and 2024 ($p = .0197$), most of which are the thematic counterparts of the ChatGPT era human-written theses.

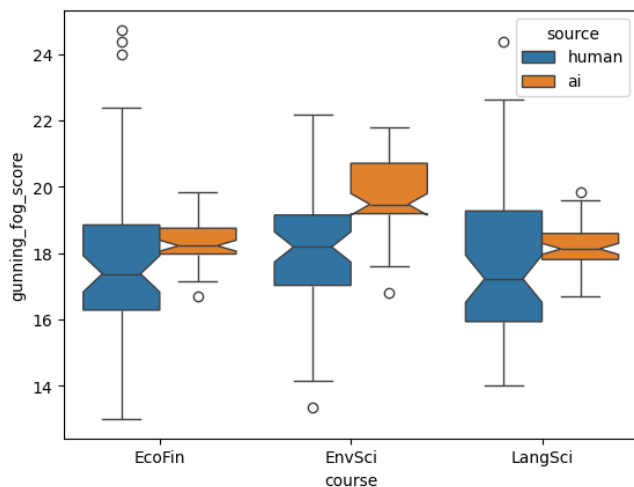
The calculation of the signed distance between human- and AI-written thesis values over time also provided significant results. It appears that the difference between the gunning fog scores of the two sources reaches the lowest value in 2024.



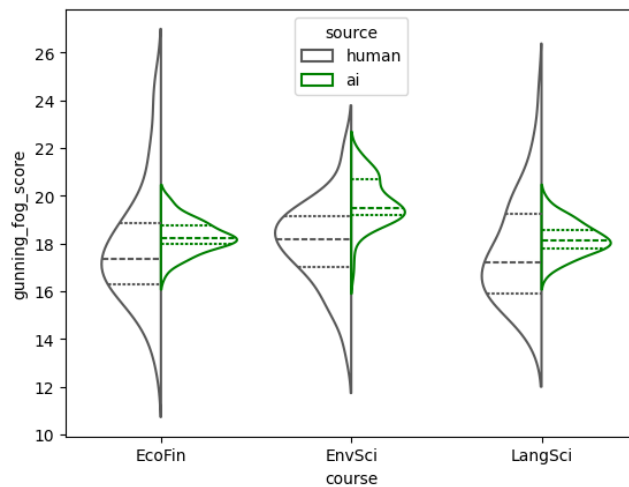
Year-based signed gap between the Gunning fog index values of human- and AI-written theses

4.3.2.2.Course-Based Comparison

The figures below demonstrate that while the human-written theses do not have substantial course-based differences, the LLM-generated Environmental Sciences theses exhibit considerably higher scores than all groups (i.e. more challenging to read).



Course-based Gunning fog index values of human- and AI-written theses (box plot)



Course-based Gunning fog index values of human- and AI-written theses (violin plot)

The statistical analyses confirm the linguistic analysis data visible on the figures. While the human-only group does not exhibit statistically significant difference in gunning fog score values ($H = 2.052$, $df = 2$, $p = .358$), the LLM-generated Environmental Sciences theses considerably diverge from all courses in the AI-only ($p < .001$) and human and AI-combined group ($p < .001$).

4.3.2.3. Source-Based Comparison

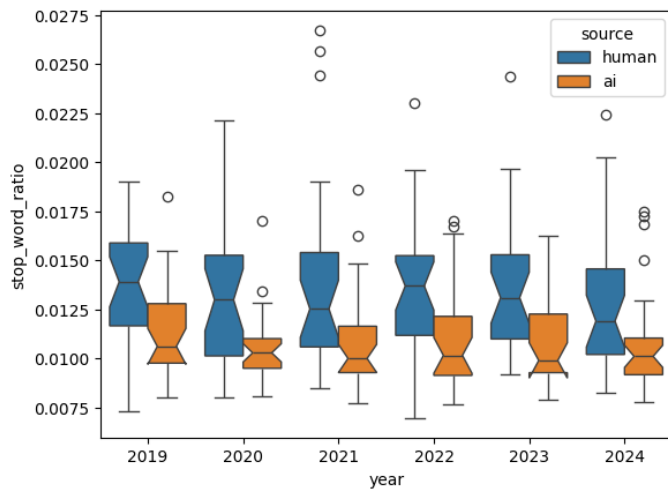
The paired Wilcoxon signed rank test revealed that GPT-generated master's theses have higher gunning fog scores than human-written theses ($W = 3503.0$, $p < .001$, $r = 0.46$, Direction: AI > Human). These results signify that the LLM-generated theses require a higher educational level to comprehend, thereby having a lower readability ease than human-authored theses.

4.4. Stylistic Features

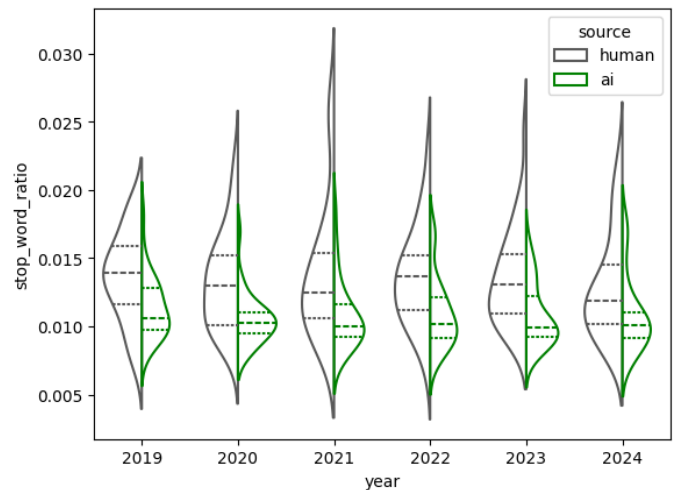
4.4.1. Stop Word Ratio

4.4.1.1. Year-Based Comparison

When the year-based stop word ratio values are analyzed based on the figures, it can be seen that the values are higher in human-written theses, consistently for every year, and there is a considerable difference between the values of human and AI-theses. Regarding the year-by-year comparison of the proportion of stop words in only human-authored texts, there does not appear to be a stark discrepancy between the values achieved.



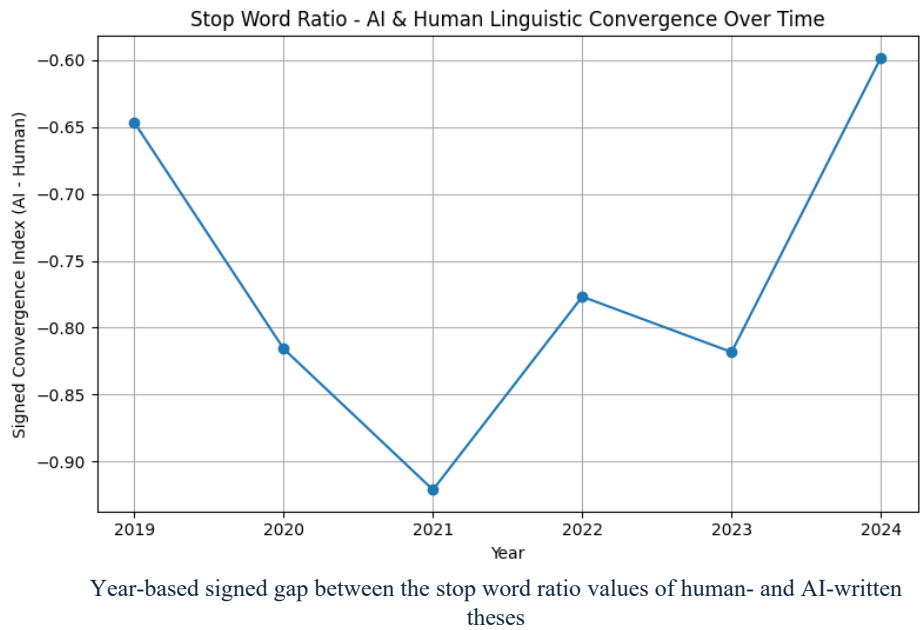
Year-based stop word ratio values of human- and AI-written theses (box plot)



Year-based stop word ratio values of human- and AI-written theses (violin plot)

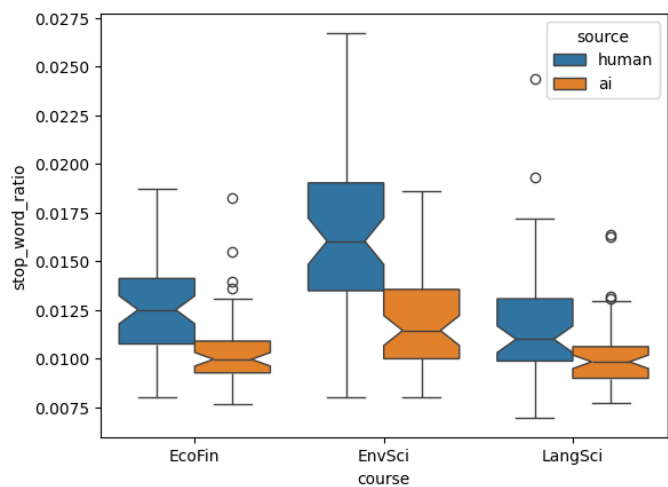
The statistical analyses that were run confirm the absence of a statistically significant variance among the values calculated for the theses from different years in human-authored theses ($H = 1.758$, $df = 5$, $p = .881$). With respect to the year-based comparison among human and AI-written theses, the statistical tests demonstrated that the human-authored theses in 2019 (pre-ChatGPT) specifically differ from the AI theses in 2020 ($p = .0237$), 2021 ($p = .0277$), 2023 ($p = .0307$) and 2024 ($p = .0378$), half of these AI theses being the thematic counterparts of the post-ChatGPT human-written theses.

The signed gap calculation yielded valuable results; in that, the distance between the stop word values of human- and AI-written theses were shown to be the narrowest in 2024, in the ChatGPT era.

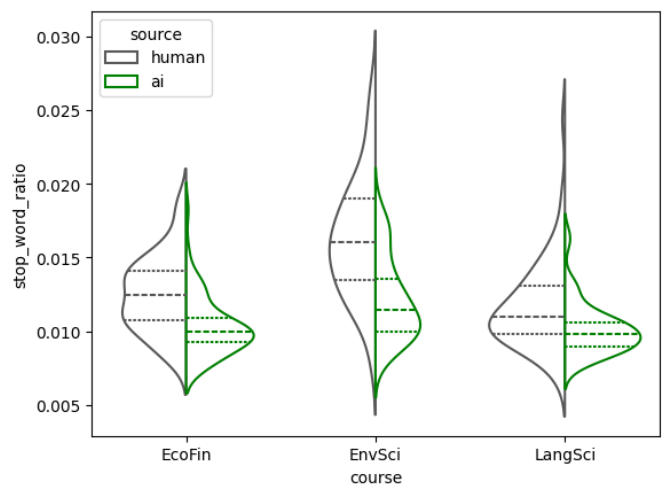


4.4.1.2.Course-Based Comparison

The comparison of the course variable demonstrated that while human- and AI-written theses in the Economics and Finance and Language Sciences areas appear to have a similar



Course-based stop word ratio values of human- and AI-written theses (box plot)



Course-based stop word ratio values of human- and AI-written theses (violin plot)

distribution and similar values with one another, the theses in the Environmental Sciences area seem to differ from the others substantially in both human and AI group.

The statistical tests confirm this finding, reporting that the values of Environmental Sciences theses in both human-only ($p < .001$) and AI-only groups ($p < .001$) show a statistically significant discrepancy from other courses. With respect to the course-based comparison across the sources, it was demonstrated that the Environmental Sciences theses in the human group differed significantly from the values of the theses from each course across both sources ($p < .001$ for each pairwise comparison). The Environmental Sciences theses in the AI-group shows almost exactly the same differences, except for the result that it does not show discrepancies with the Language Sciences and Economics and Finance theses in the human group ($p > .05$), which can be clearly confirmed from the similar values reported on the box plot.

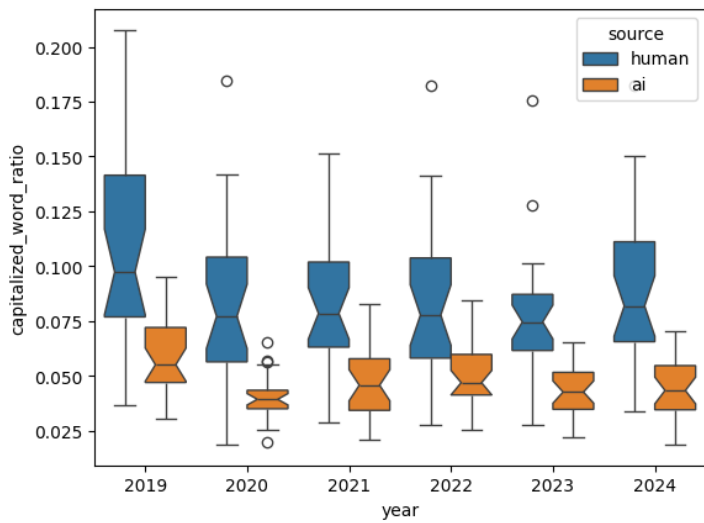
4.4.1.3.Source-Based Comparison

Confirming the previous findings on stop word ratio, the comparison across the sources also demonstrate that human-authored theses feature a higher proportion of stop words ($W = 385.0$, $p < .001$, $r = -0.94$, Direction: Human > AI).

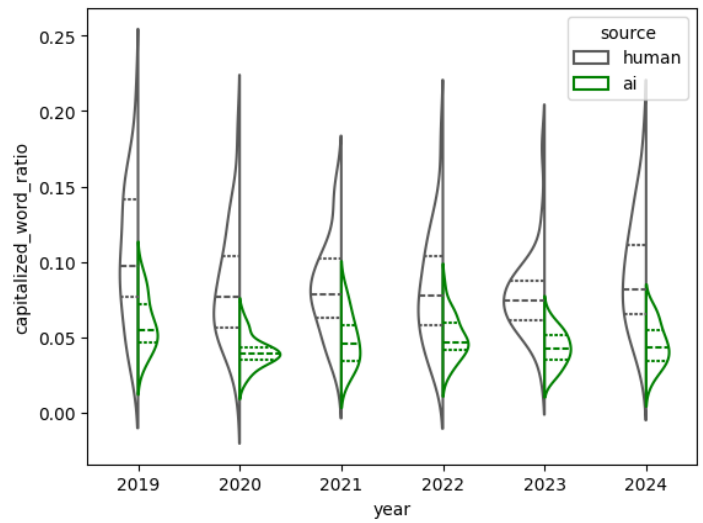
4.4.2. Capitalized Word Ratio

4.4.2.1.Year-Based Comparison

The data demonstrate that the capitalized word ratio in human-written theses surpass the values in the AI-generated texts in every year. Although the proportion of capitalized words in human-written theses exhibits a decrease in 2022, and a respective increase in 2023, reaching the highest mean value in 2024, the Kruskal-Wallis and Dunn's post-hoc test did not find this to be a statistically significant difference ($H = 7.617$, $df = 5$, $p = .178$). On the other hand, there is a substantial discrepancy in the year-based comparison of the GPT-generated theses ($H = 21.045$, $df = 5$, $p < .001$).



Year-based capitalized word ratio values of human- and AI-written theses (box plot)



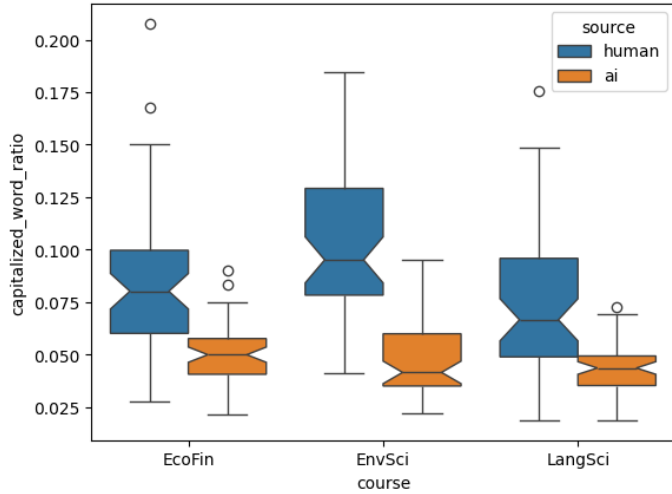
Year-based capitalized word ratio values of human- and AI-written theses (violin plot)

Regarding the comparison in the human and AI-combined theses group, the statistical tests once again reveal important results. The AI-written theses in 2019 is the only group that does not significantly differ from the human-written theses in each year ($p > .05$), showing higher ratios than other years, and therefore being closer to the human-authored text values.

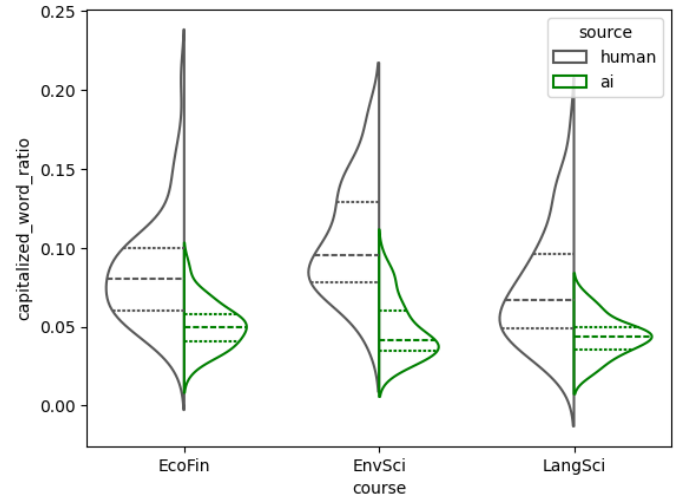
4.4.2.2.Course-Based Comparison

Although no considerable course-based variance was found in the AI-only group ($H = 5.047$, $df = 2$, $p = .080$), the Environmental Sciences theses in the human-only group exhibited

significant differences with the other two areas ($p < .05$). As to the human and AI-combined group, each pairwise comparison (the values of human and AI, and each course) displayed a



Course-based capitalized word ratio values of human- and AI-written theses (box plot)



Course-based capitalized word ratio values of human- and AI-written theses (violin plot)

statistical significance ($p < .01$), a further difference being between the human-written theses in Environmental Sciences and Language Sciences ($p < .001$). The course-based comparison, thus, further proves the considerable discrepancy between the capitalized word ratios of human and AI-theses.

4.4.2.3. Source-Based Comparison

The paired Wilcoxon signed rank test demonstrates that there is a statistically significant overall discrepancy between the human- and AI-written theses, where the former showed considerably higher capitalized word ratios than the latter ($W = 420.0$, $p < .001$, $r = -0.94$, Direction: Human > AI).

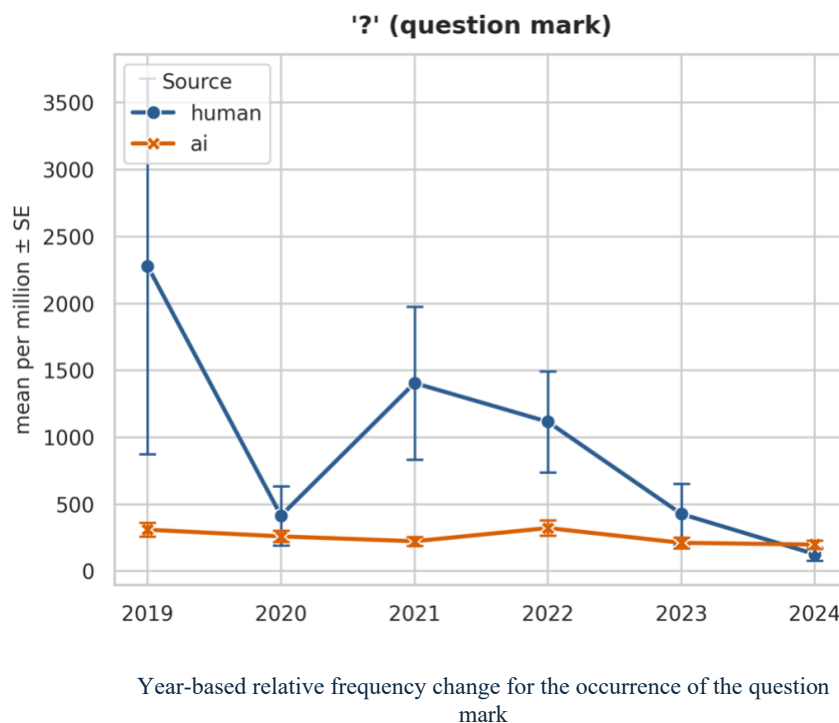
4.4.3. Punctuation Count

The categorization and the reporting of the punctuation mark counts follow a similar procedure with the LLM-associated words. However, in this case, both human-written and AI-

generated theses were considered and compared with one another, because no predetermined list of LLM-associated punctuation marks was set for analysis.

Relative frequency of each punctuation mark was calculated, which was normalized per million tokens to prevent any length-related effects that may have accounted for significant results. Then, Kruskal-Wallis test was carried out, which was followed by the Dunn's post-hoc test with Bonferroni correction.

Among the results, the most possible diachronic AI trace can be viewed for the question mark, the frequency of which gradually decreased up until its lowest value in 2024. It can be seen on the plots that this mark was preferred considerably less in LLM-generated texts, denoting that humans' including fewer punctuation marks in their texts could be a stylistic similarity with AI-texts.

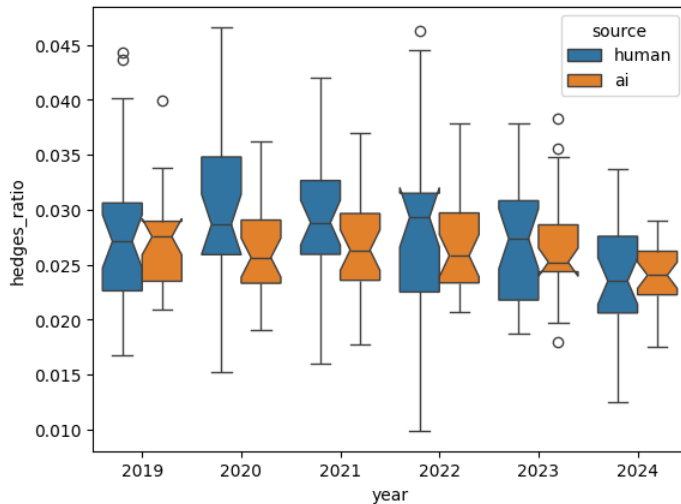


Dunn's post-hoc test further confirmed the visible difference in the year-based values of question mark usage in human-written texts (human-2021 and human 2023: $p = .0430$; human-2021 and human-2024: $p = .0044$; human-2022 and human-2024: $p = .0110$). The pairwise comparison further revealed that there was no statistically significant difference between the specific punctuation counts of human-authored and AI-generated theses in the year 2024, which indicates that the values for this year in two sources are highly similar.

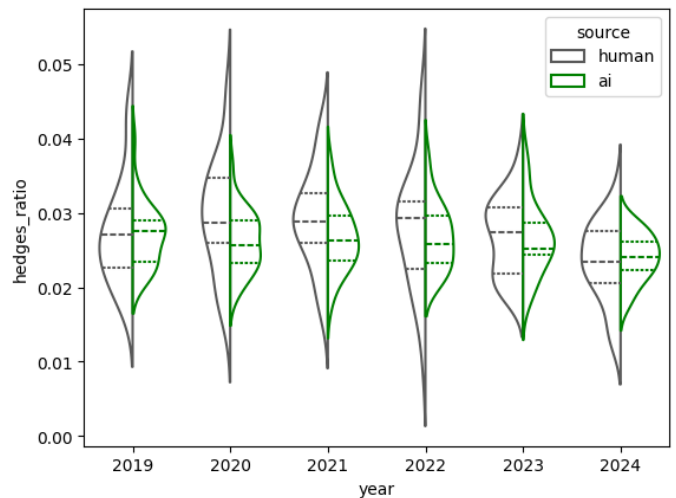
4.4.4. Hedge Word Ratio

4.4.4.1. Year-Based Comparison

The linguistic analysis revealed that human-written theses have a substantially higher hedge word ratio than GPT-generated theses in 2020 and 2021. So, as can be seen from the figures, these two years in pre-ChatGPT period show the highest median of values for the human-written theses, which are in turn the years that exhibit the largest discrepancy from the LLM-theses.



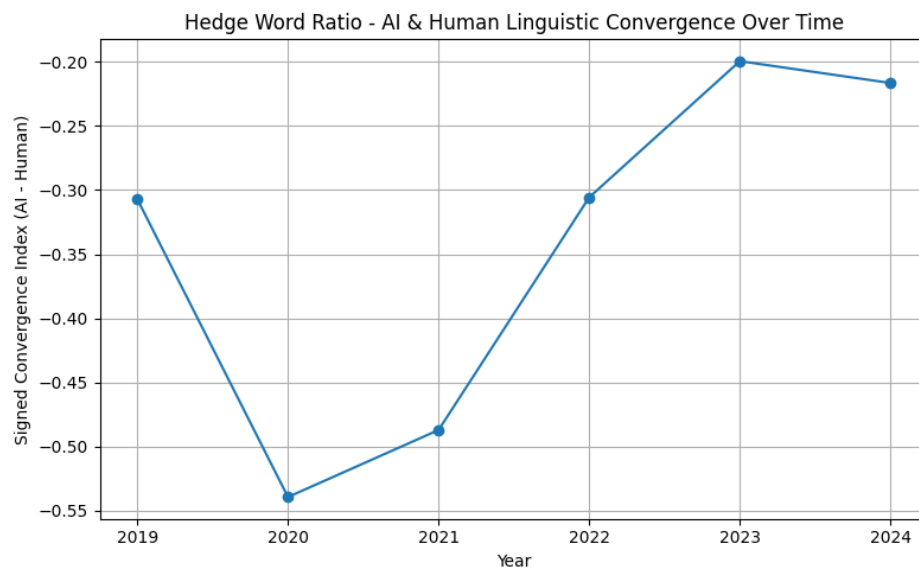
Year-based hedge word ratio values of human- and AI-written theses (box plot)



Year-based hedge word ratio values of human- and AI-written theses (violin plot)

The statistical analysis validates these results, revealing a significant difference between the following human-written theses pairs: 2020-2024 ($p = .0297$), and 2021-2024 ($p = .0200$). The Kruskal-Wallis test also found an overall significant difference in the year-based values across human and AI groups. Dunn's post-hoc test demonstrated that the specifically differing pairs were human-written theses in 2021 and 2024 ($p = .0237$), and human-written theses in 2021 and AI-generated theses in 2024 ($p = .0259$). The findings show the discrepancy between a pre-ChatGPT human-written thesis and a ChatGPT era thesis, which implies a similarity in the human-written and AI-generated hedge word values in 2024.

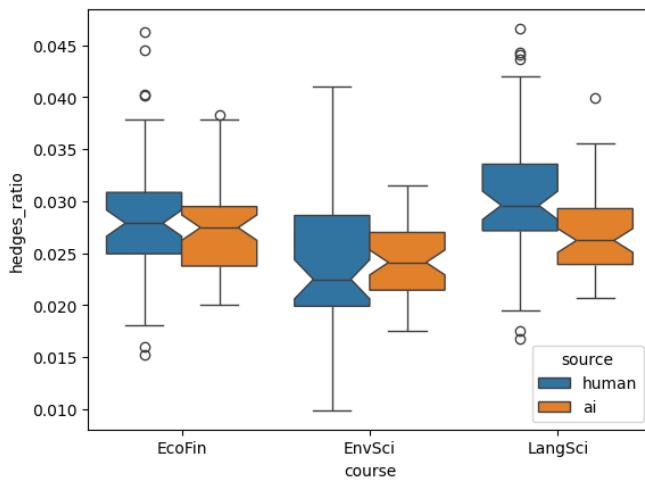
The signed gap result for the hedge word ratio exhibits that the differences between human- and AI-written theses values reached the lowest point in 2023, while the second lowest difference was observed in 2024, which could be due to the high outlier values of GPT-generated theses in 2023.



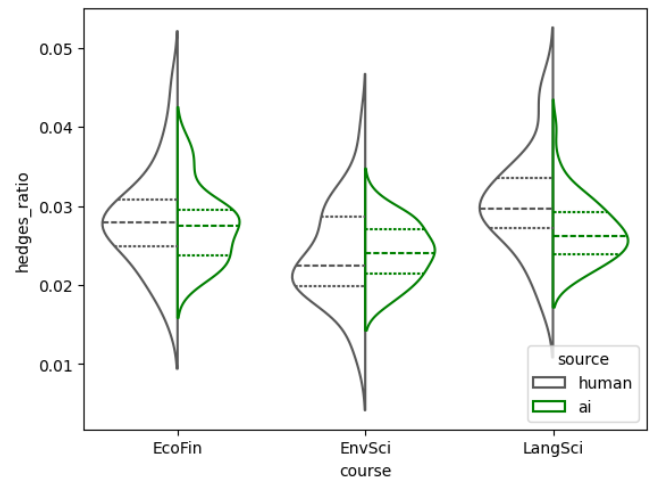
Year-based signed gap between the hedge word ratio values of human- and AI-written theses

4.4.4.2.Course-Based Comparison

The findings of the linguistic and statistical analysis of the course-based values in human-only, AI-only, and human and AI-combined comparison groups revealed that the theses in the Environmental Sciences program in both human and AI groups show the highest difference in the pairwise comparisons ($p < .05$ in all pairwise comparisons where they show differing results). It is crucial to note that the Environmental Sciences theses in human-written and AI-generated condition showed statistically similar results ($p = 1.000$), having a lower value distribution when compared to the other courses in both human and AI-theses.



Course-based hedge word ratio values of human- and AI-written theses (box plot)



Course-based hedge word ratio values of human- and AI-written theses (violin plot)

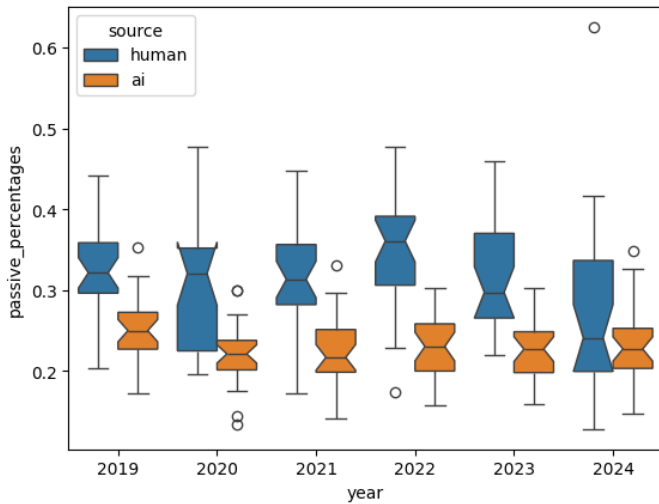
4.4.4.3.Source-Based Comparison

Although there were significant discrepancies in the year-based hedge word ratios of human-written theses, the overall source-based comparison did not yield statistically significant differences between human and AI theses, while the former was found to have slightly higher values ($W = 5439.0$, $p = .0679$, $r = -0.17$, Direction: Human > AI).

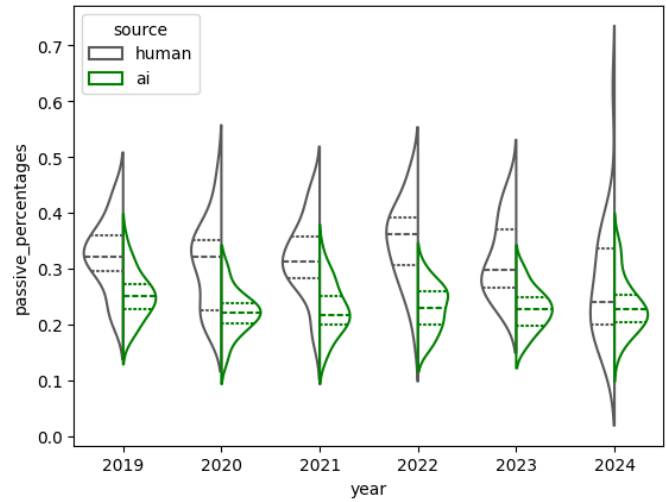
4.4.5. Passive Voice Ratio

4.4.5.1.Year-Based Comparison

The results reveal that human-written theses employ a higher proportion of passive voice structures than GPT-theses, which is consistent in every year. Another specific observation is the decrease in passive percentages from 2022 to 2023, and successively to 2024, with the year 2024 showing the lowest distribution of values among other years in the human-written theses set.



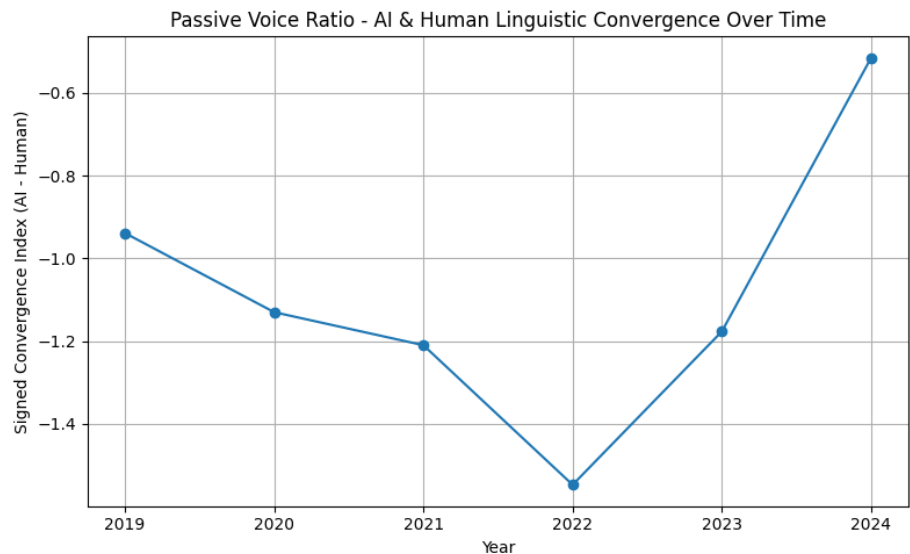
Year-based passive voice ratio values of human- and AI-written theses (box plot)



Year-based passive voice ratio values of human- and AI-written theses (violin plot)

The statistical tests found a significant difference between the human-written theses in 2022 and 2024 ($p = .0039$), the final year of the post-ChatGPT theses in the dataset. While the AI-only group does not show substantial year-based discrepancies in the values ($H = 7.204$, $df = 5$, $p = .205$), the statistical comparison across sources (i.e. human and AI-combined set) demonstrates that the only year in which human-written theses do not show a discrepancy with any of the LLM-theses is the year 2024 ($p > .05$). This specific finding implies that the human-written theses in 2024 show the highest similarity in passive voice ratio values to the GPT-generated theses.

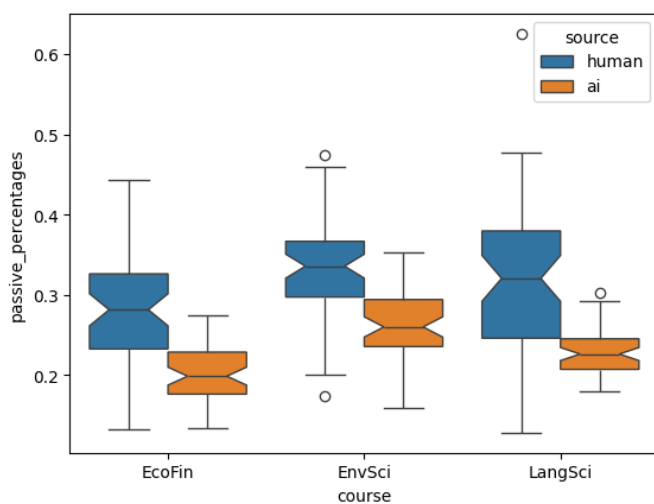
The increasing similarity between the passive voice ratios of human-authored and AI-generated theses in 2024 (ChatGPT era) was further shown to be significant with the signed distance calculation.



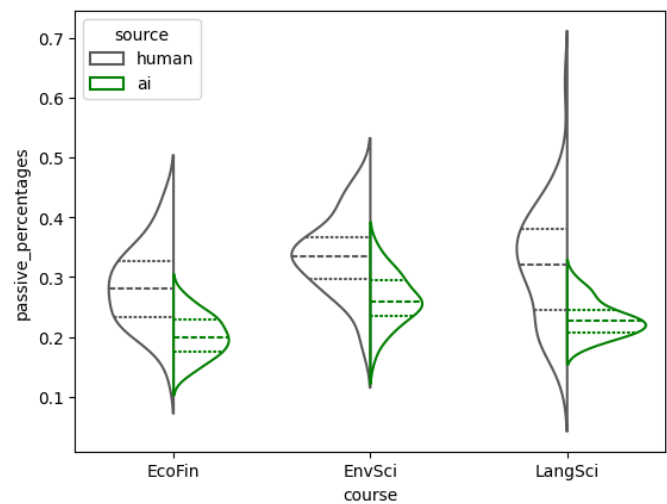
Year-based signed gap between the passive voice ratio values of human- and AI-written theses

4.4.5.2.Course-Based Comparison

The linguistic and statistical analyses report a high course-based variance inside the human-only, AI-only and human and AI-combined sets.



Course-based passive voice ratio values of human- and AI-written theses (box plot)



Course-based passive voice ratio values of human- and AI-written theses (violin plot)

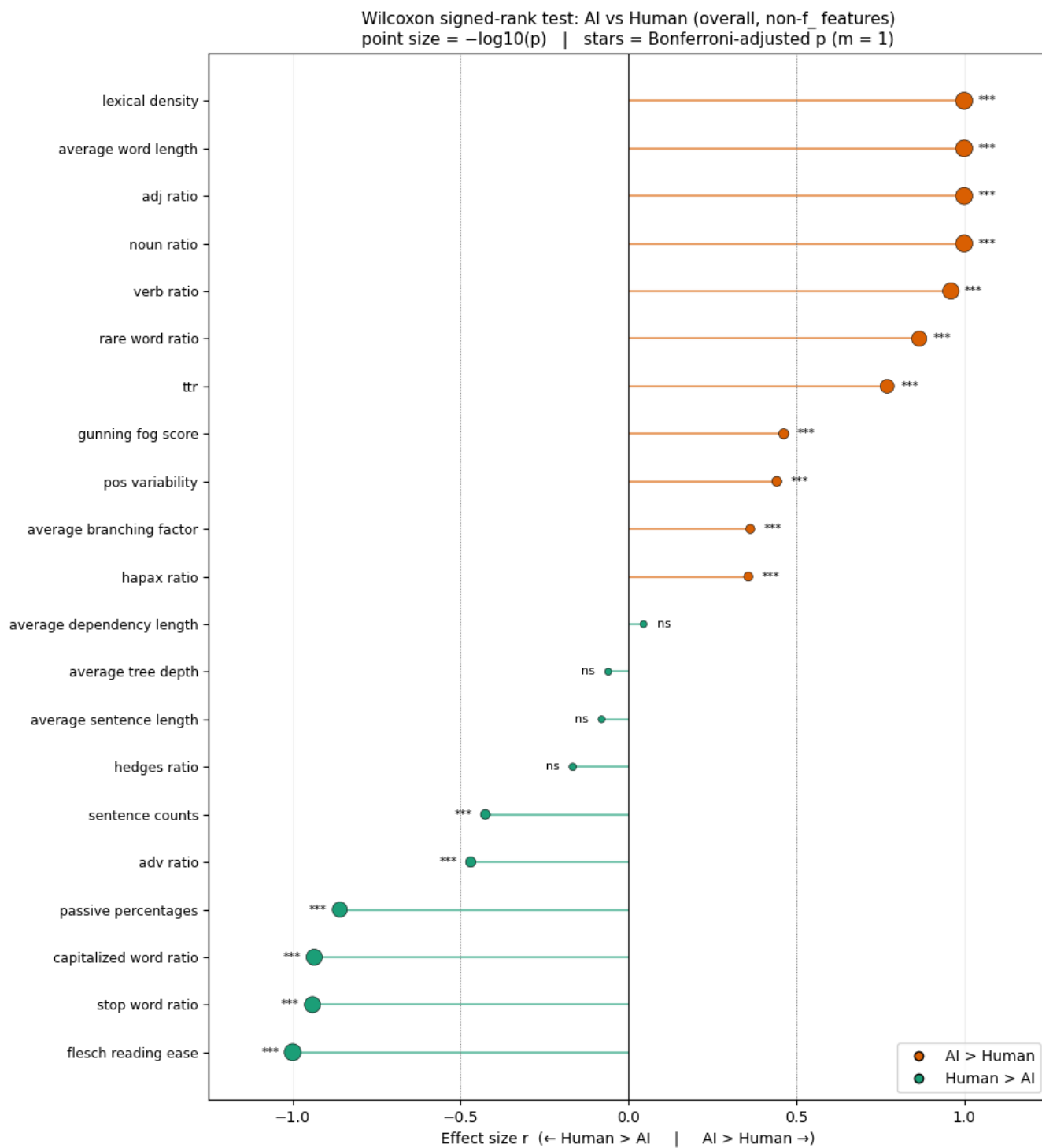
The Economics and Finance program was found to have lower values and was discovered to be significantly differing from the other two courses in the human-only set (Environmental Sciences: $p < .001$; Language Sciences: $p = .0464$). As to the AI-only group, as it can also be seen from the figures, the values of each course significantly differ from one another ($p < .01$ for all pairwise comparisons).

The statistical analysis of the values across sources (i.e. human and AI-combined group) found the human-written theses from Environmental Sciences ($p < .05$ in four out of five pairwise comparisons) and LLM-generated theses from Economics and Finance ($p < .001$ in four out of five comparisons) to be the most divergent groups.

4.4.5.3.Source-Based Comparison

The paired Wilcoxon signed rank test demonstrated that human-authored theses have a higher ratio of passive voice than GPT-generated theses ($W = 916.0$, $p < .001$, $r = -0.86$, Direction: Human $>$ AI).

A plot that visualizes and summarizes the source-based differences between the linguistic feature values of human-written and AI-generated theses can be found below.



The results of the paired Wilcoxon signed-rank test for the source-based differences of the linguistic features

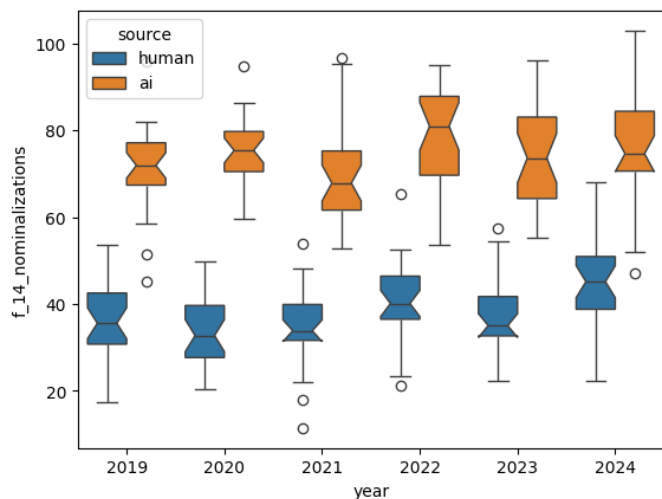
4.4.6. Biber Features

Since the Biber features tag set consists of 67 features, only the top statistically significant differences in the features will be reported.

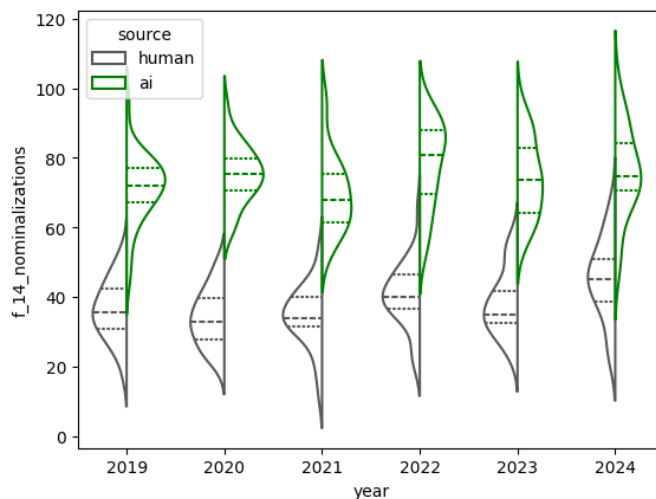
4.4.6.1. Year-Based Comparison

Just like the previous linguistic features, human-written master's theses from 2024 were the ones that yielded the most statistically significant results when compared with the theses from the pre-ChatGPT era. The features that yielded only statistically significant differences between the pre-ChatGPT and ChatGPT years were nominalizations, perfect aspect, “be” as the main verb, wh- clause, predicative adjectives, predictive modals and analytic negation.

The nominalization feature displayed considerable differences between human-written theses from each year in the pre-ChatGPT era (2019, 2020 and 2021) and the ones from 2024 (2019-2024: $p = .0342$; 2020-2024: $p = .00108$; 2021-2024: $p = .0034$). It is of crucial importance to note that the year 2024 in human-authored theses has the highest and thereby the closest median value to AI-generated theses, which have a higher number of nominalizations in every year.

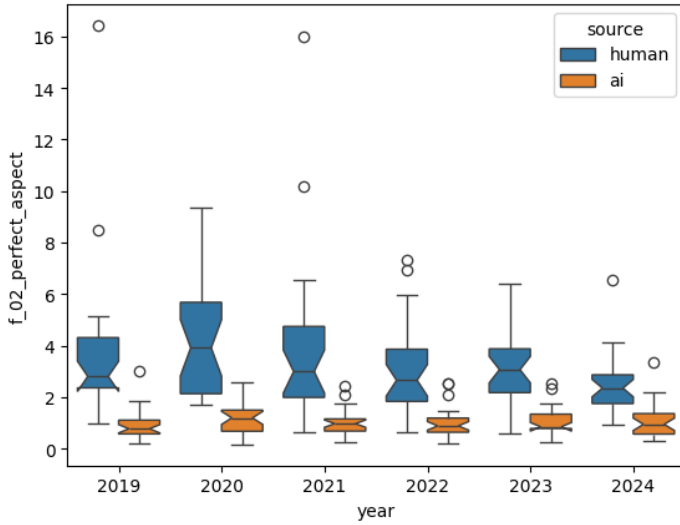


Year-based nominalization values of human- and AI-written theses (box plot)

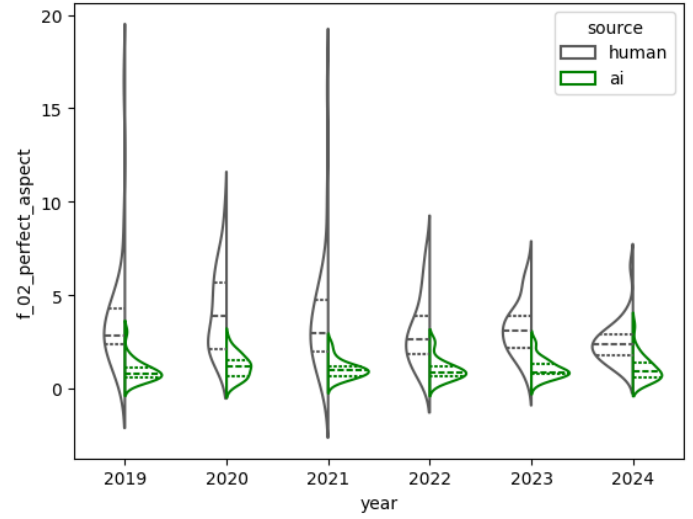


Year-based nominalization values of human- and AI-written theses (violin plot)

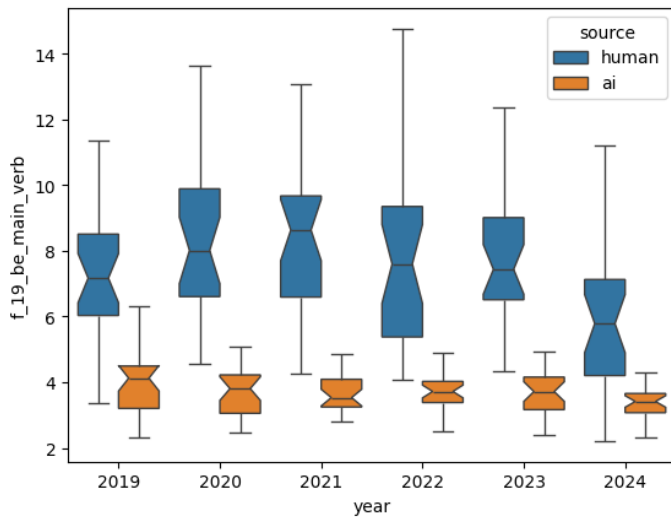
The other features which showed difference between 2020 (pre-ChatGPT) and 2024 (ChatGPT era) are perfect aspect ($p = .0243$), “be” as the main verb ($p = .0103$) and predictive modals ($p = .0196$). The value distributions for these features are displayed in the following figures.



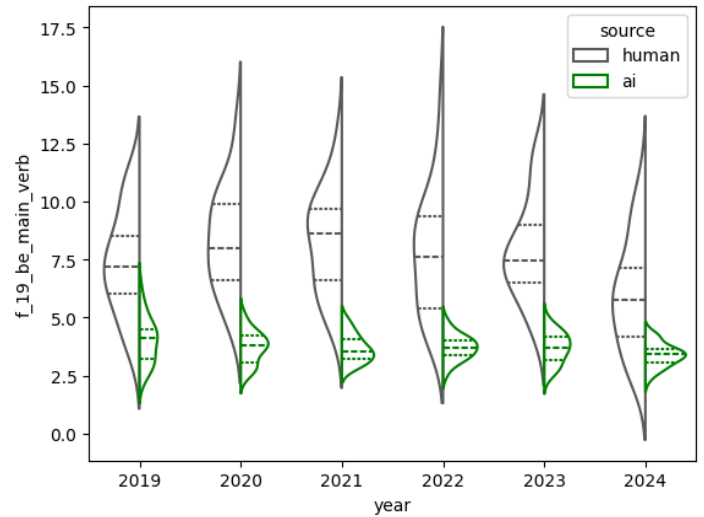
Year-based perfect aspect values of human- and AI-written theses (box plot)



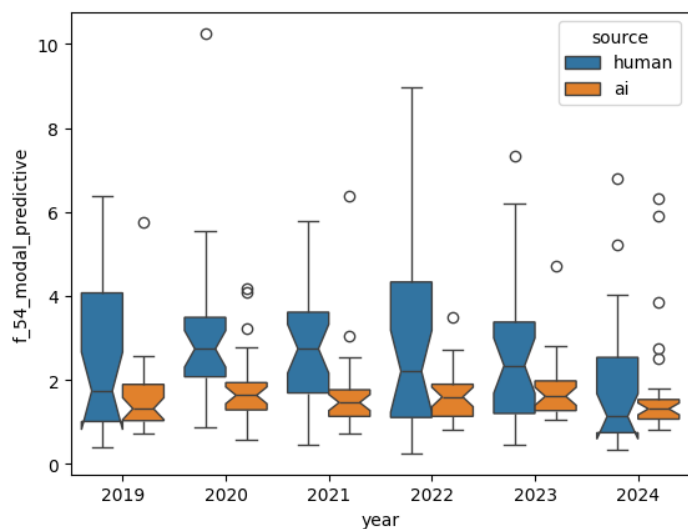
Year-based perfect aspect values of human- and AI-written theses (violin plot)



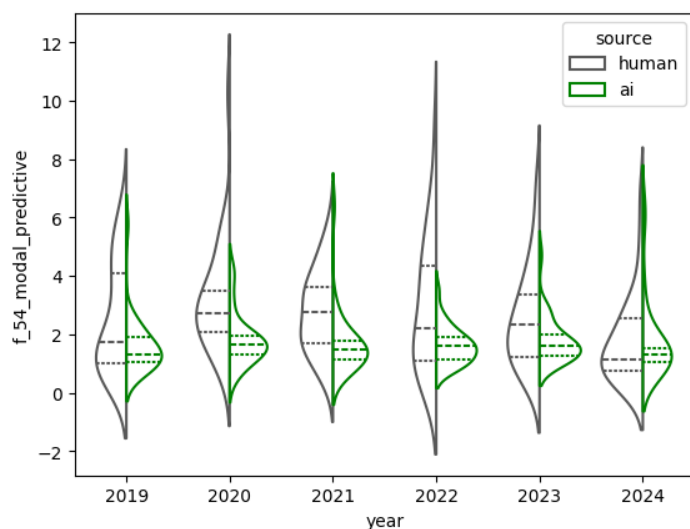
Year-based “be” as main verb values of human- and AI-written theses (box plot)



Year-based “be” as main verb values of human- and AI-written theses (violin plot)

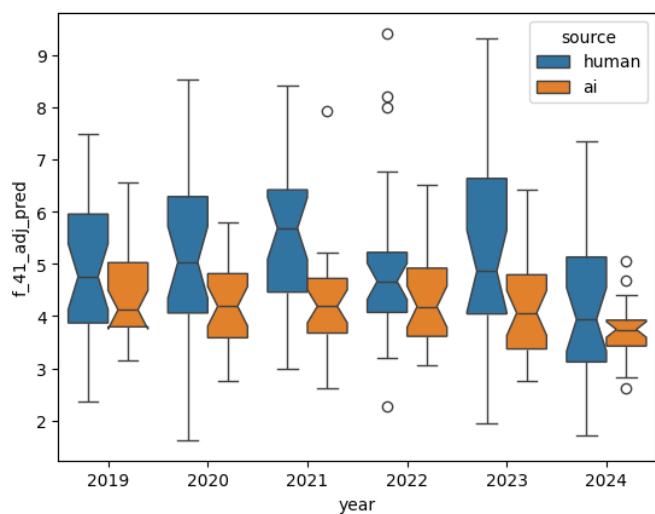


Year-based predictive modal values of human- and AI-written theses (box plot)

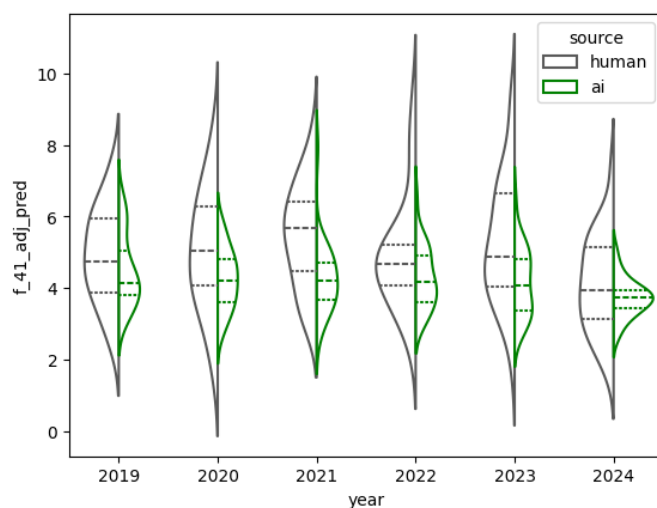


Year-based predictive modal values of human- and AI-written theses (violin plot)

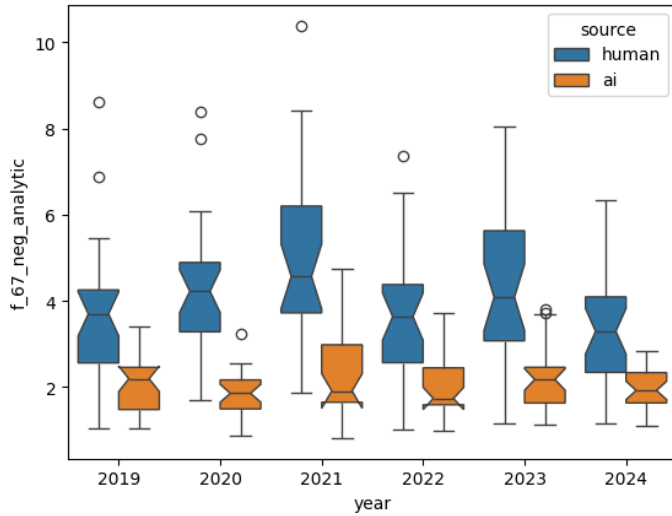
The following figures show the distribution of the features that significantly differ between the human-written theses from 2021 and 2024, which are predicative adjectives ($p = .0192$) and analytic negation ($p = .0111$).



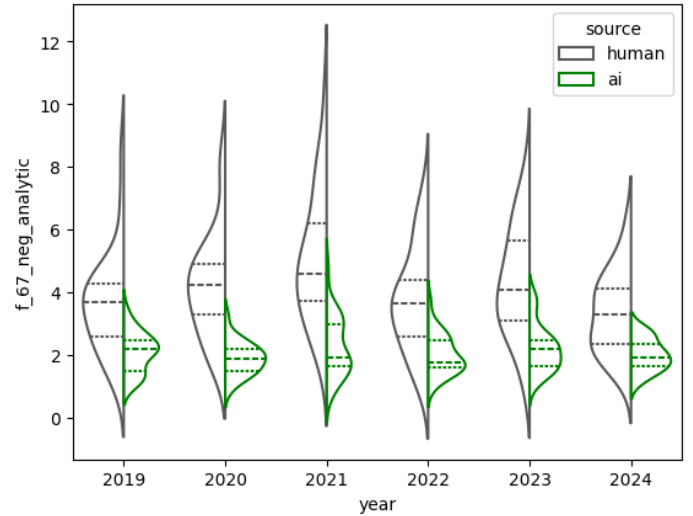
Year-based predicative adjective values of human- and AI-written theses (box plot)



Year-based predicative adjective values of human- and AI-written theses (violin plot)

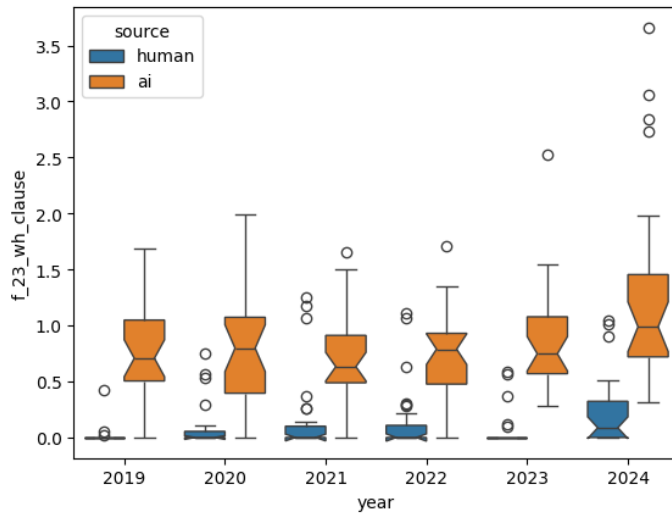


Year-based analytic negation values of human- and AI-written theses (box plot)

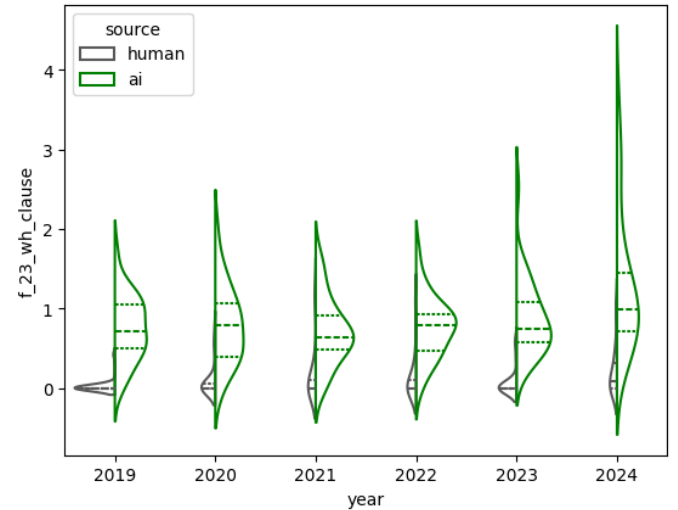


Year-based analytic negation values of human- and AI-written theses (violin plot)

Finally, the wh- clause feature exhibited a statistically significant difference between the year pair 2019 and 2024 in human-written master's theses ($p = .0087$).



Year-based wh clause values of human- and AI-written theses (box plot)

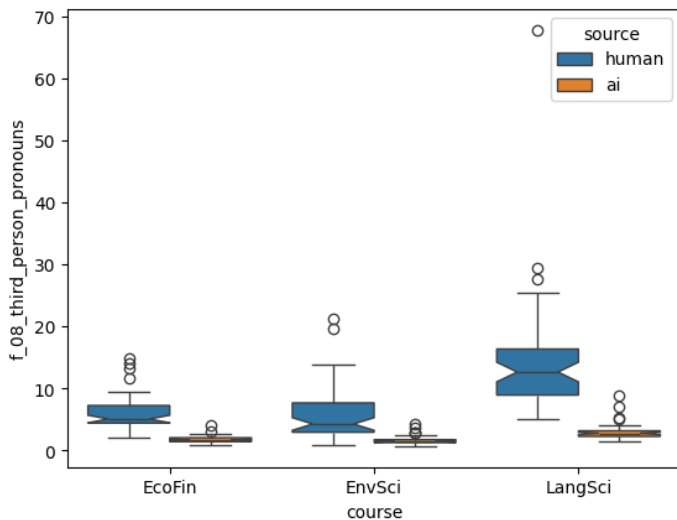


Year-based wh clause values of human- and AI-written theses (box plot)

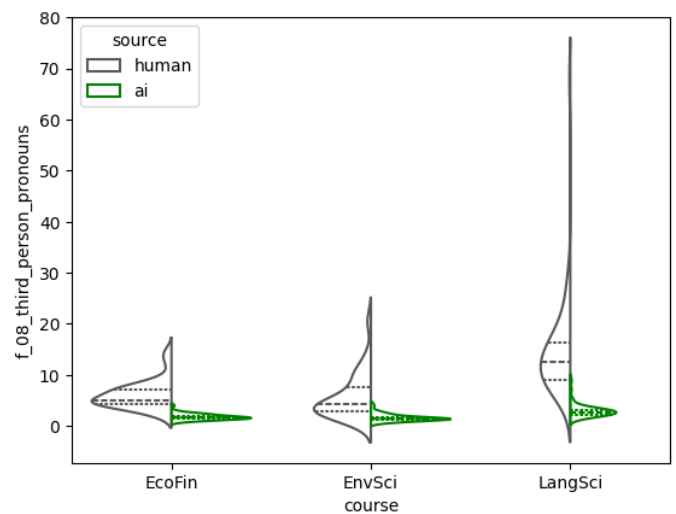
What can be commonly observed from the figures above is that the human-written master's theses in 2024 that were examined show the most similar median values to LLM-generated theses, signifying a change in the stylistic quality of human writing that shows similarity with the stylistic feature values of AI-generated texts.

4.4.6.2.Course-Based Comparison

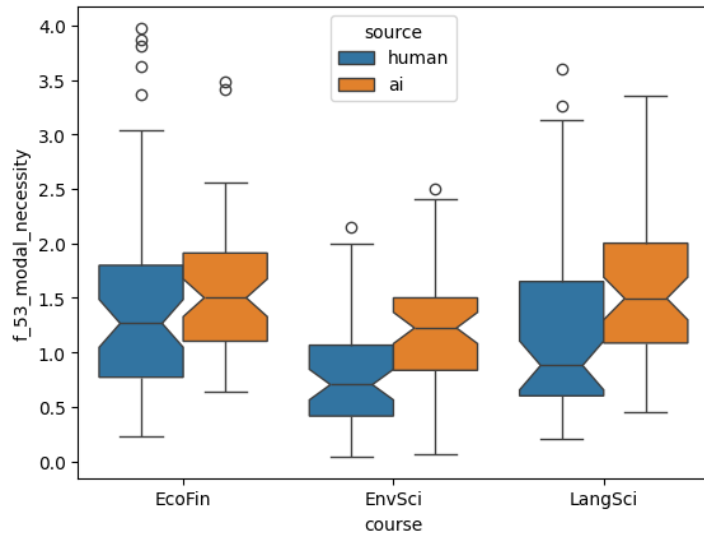
As to the linguistic differences among the human-written theses from different areas, it was exhibited that the features in the tag set with the top five highest significance in the value difference consisted of hedges (Environmental Sciences and Language Sciences, $p < .001$), third person pronouns (Environmental Sciences and Language Sciences, $p < .001$), modals of necessity (Environmental Sciences and Economics and Finance, $p < .001$), the pronoun “it” (Environmental Sciences and Economics and Finance, $p < .001$) and first-person pronouns (Environmental Sciences and Language Sciences, $p < .001$). Since the hedge word ratio was covered as a separate stylistic feature on its own, the remaining four statistically significant features in the human-written course-based comparison are displayed below.



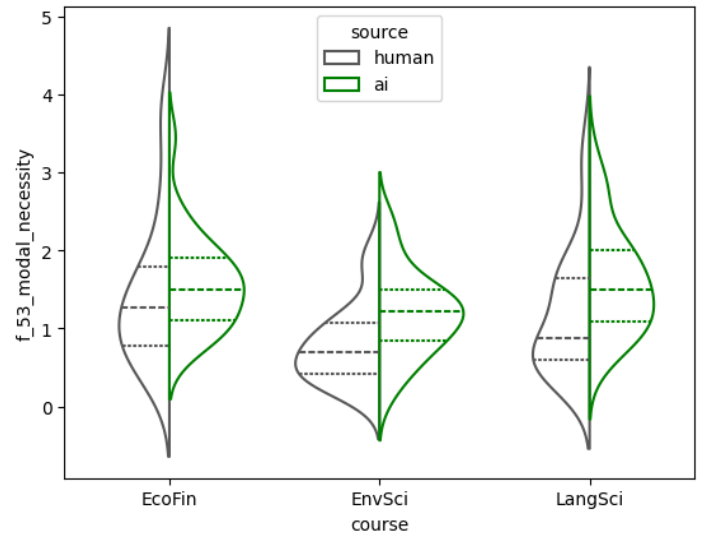
Course-based third person pronoun values of human- and AI-written theses (box plot)



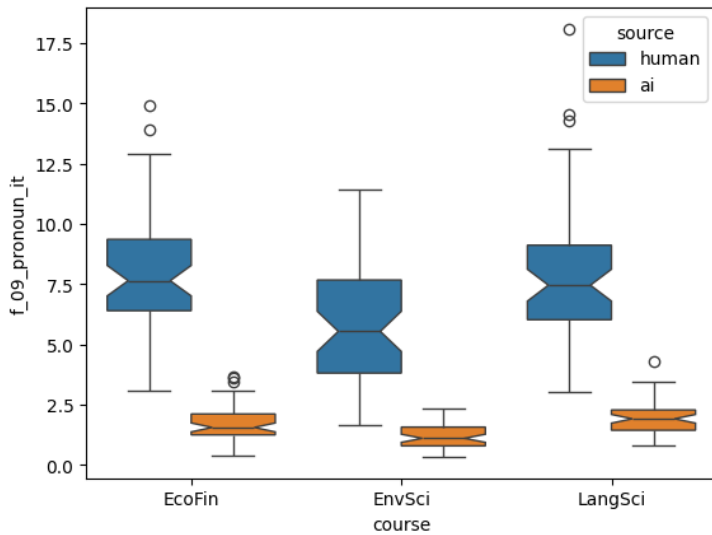
Course-based third person pronoun values of human- and AI-written theses (violin plot)



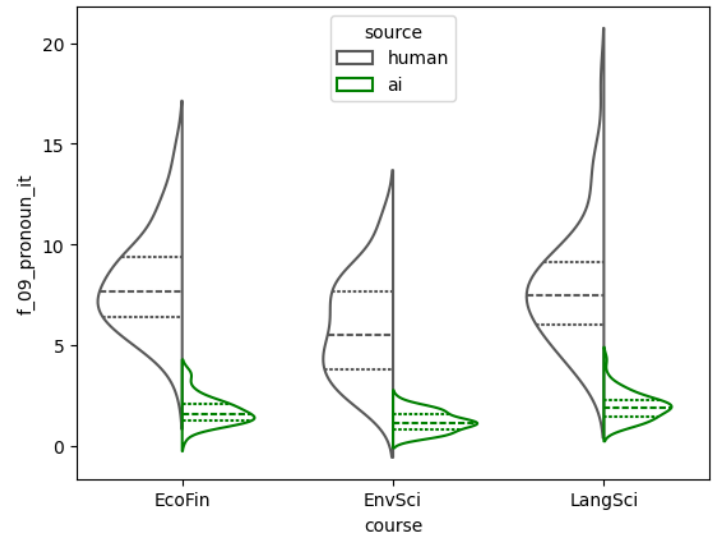
Course-based modals of necessity values of human- and AI-written theses (box plot)



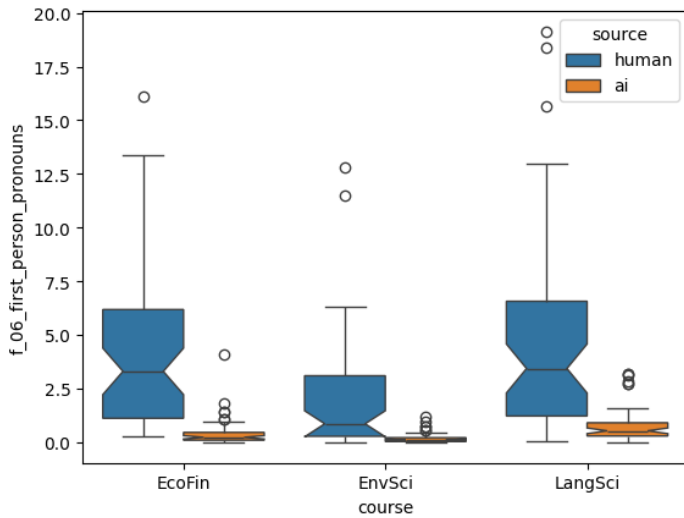
Course-based modals of necessity values of human- and AI-written theses (violin plot)



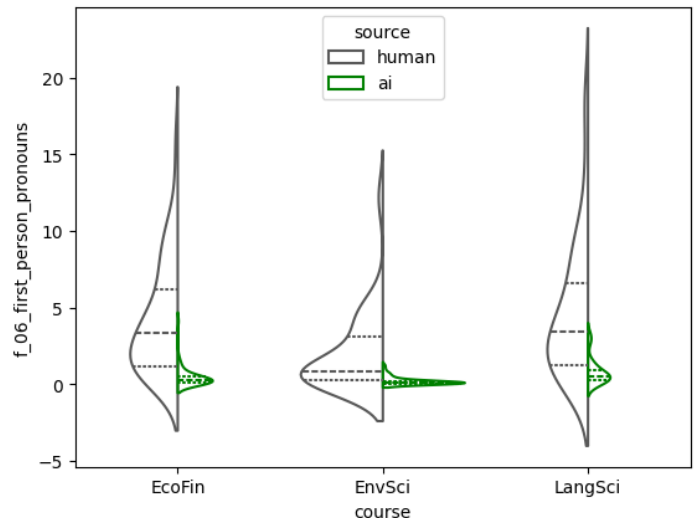
Course-based pronoun “it” values of human- and AI-written theses (box plot)



Course-based pronoun “it” values of human- and AI-written theses (violin plot)



Course-based first person pronoun values of human- and AI-written theses (box plot)



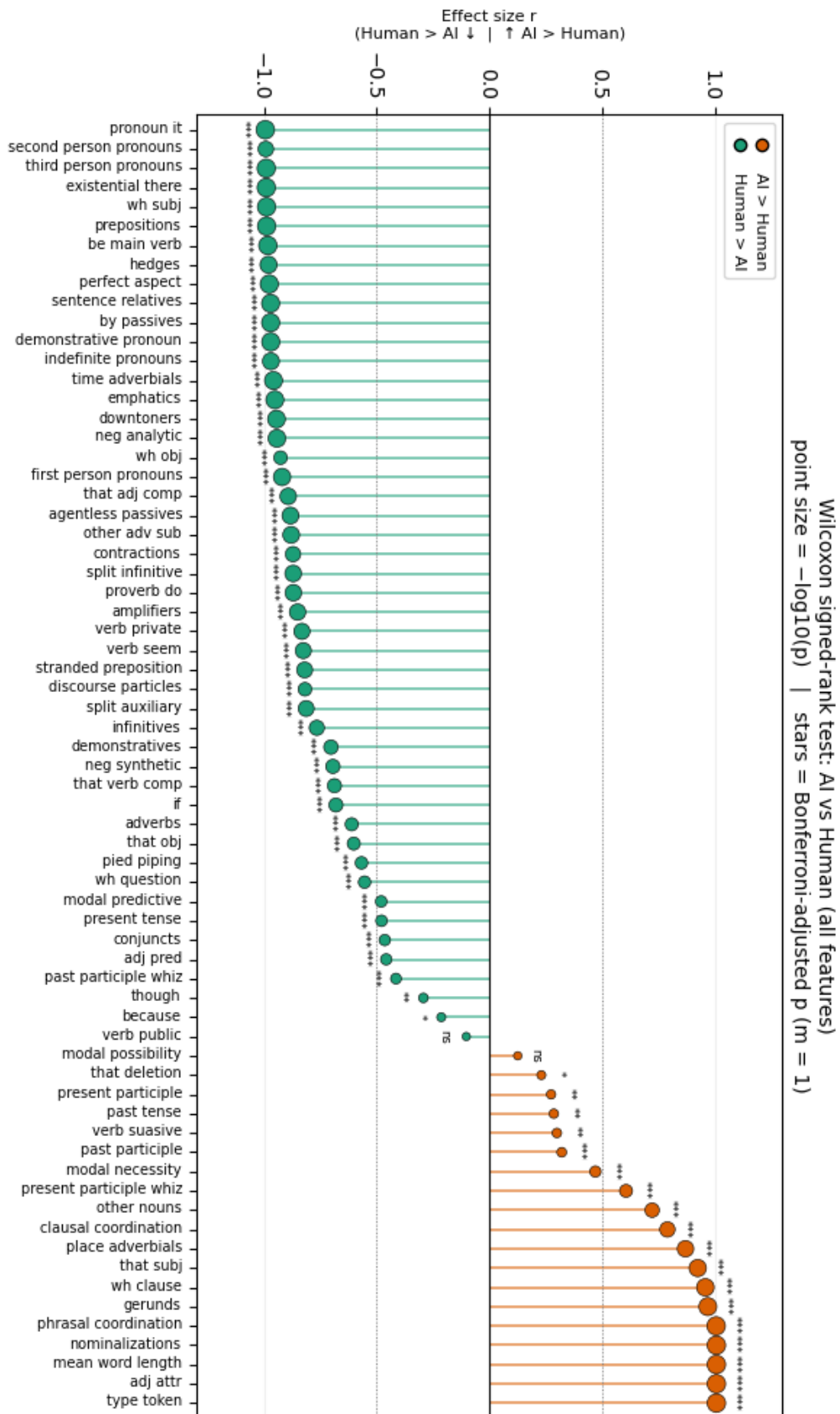
Course-based first person pronoun values of human- and AI-written theses (violin plot)

4.4.6.3. Source-Based Comparison

Finally, in the overall source-based comparison that was conducted through the paired Wilcoxon signed rank test, 63 features out of 67 in the tag set yielded statistically significant differences between AI and human-written texts. The only four features that were non-significant are modals of possibility, public verbs, the use of “because”, and “that” deletion. ($p > .05$).

Moreover, the features for which AI-generated theses demonstrated considerably higher values are TTR, attributive adjectives, mean word length, nominalizations, and phrasal coordination ($p < .001$), whereas the human-authored theses show significantly higher values for the following features: the pronoun “it”, second person pronouns, third person pronouns, existential “there” and “wh” as subject ($p < .001$).

The plot below displays the source-based differences for each feature in the Biber tag set.



5. Discussion

After the detailed reporting of all the linguistic and statistical data, this section will focus on the most significant findings, under the broad categorization of lexical, syntactic, readability and stylistic features.

5.1. Lexical Features

Hewa Julige (2026) observed that lexical density of abstracts showed a tendency to go upwards after the introduction of GenAI, which indicates a text quality that is “more compressed and densely packed with information” (p. 53). Similarly, Andre et al. (2023) found that human-written abstracts include a higher number of function words, which denotes that humans make a more frequent use of function words while AI has a higher percentage of content words. The current thesis confirms both of these findings; in that, the lexical density of AI-generated theses were found to be higher than human-written theses in every year, and the highest lexical density score in human-authored theses was observed in 2024, which is the last year of the ChatGPT era. The specific increase in lexical density was especially seen from the year 2023 to 2024. The higher values observed in lexical density for the human-written theses in 2024 could have been caused by numerous factors, one of which could possibly be interpreted as GenAI influence.

The current findings also confirm the results of Zindela (2023), who found a slightly higher lexical density in AI-generated content as well, and Bao et al. (2025), who showed that the lexical density values of human-written abstracts increased following the introduction of ChatGPT.

Regarding average word length, most of the previous studies in the literature point to the same direction. Hewa Julige (2026) observed an increase in average word length in modern abstracts, which include significantly longer and more morphologically complex words. Bao et

al. (2025) also discovered an increase in average word length values of human-authored texts after the release of ChatGPT. Moreover, in Andre et al.'s (2023) study on fully AI-generated and human-written texts, average word length was found to be higher in LLM texts. The current thesis validates these previously found results in the literature. It was discovered that fully AI-generated master's theses in every year had a higher average word length than the human-written ones, and a considerable increase was observed in 2024 in the human-written theses group. Andre et al. (2023) interpreted the rise in average word length and complexity as the increasing prevalence of technical academic vocabulary, which is considered as one of the characteristics of AI-assisted texts.

The course-based comparisons of the average word length in the human-written theses further ensure that the technicality in words is not brought about by area-specific variances, because no statistically significant course-based differences were found for the average word length feature in the human-only group ($p = .254$), which strengthens the probability of LLM assistance in this increase instead of the introduction of field-specific novel vocabulary.

Regarding the ratios of adjectives, adverbs, verbs and nouns separately, Tudino and Qin (2024) demonstrated that LLM-generated texts have a higher ratio of verbs, adjectives and nouns, while adverbs are attested more frequently in human-authored texts (confirmed also by Chong et al., 2023; Guo et al., 2023; Liao et al., 2023, cited in Terčon & Dobrovoljc, 2025). The current thesis presented the exact findings, where AI-generated master's theses displayed higher adjective, noun and verb ratios while human-written theses have higher adverb ratios. The present study contradicted Culda et al. (2025), who reported that LLM-authored texts feature a fewer number of adjectives, adverbs, nouns and verbs, thereby having a simpler structure. This great contradiction in the results could be due to the focus texts in the current study and Tudino

and Qin (2024) which are solely academic, while Culda et al. (2025) analyzed blogs, Wikipedia entries and abstracts, where the academic genre is only one among the other options.

According to Tudino and Qin (2024), the higher frequency of nouns in AI-generated texts indicate a more scientific writing style that is more characterized by the more frequent employment of nominal structures, and the high number of adjectives simply corresponds to this increasing tendency in nominalization. In the present study, specifically the adjective ratio of human-written theses show higher values in 2024, which signifies a more LLM-like linguistic behavior considering the higher adjective ratios consistently observed in GPT-generated texts. Moreover, the majority of the pre-ChatGPT human-theses differ significantly from the adverb ratio values of LLM texts, denoting less LLM-like lexical values in human-authored theses written before the introduction of ChatGPT. Finally, the similarity of verb ratios in Language Sciences theses across human and AI groups could be interpreted as the LLMs' increased capacity in imitating this lexical feature of human-written Language Sciences theses better than the other micro areas.

As to lexical diversity, Hewa Julige (2026) observed that the lowest values of lexical diversity was found in the AI-Assisted Writing era, specifically from 2018 to 2022, which suggests that although the word length and complexity is higher in the abstracts from this period, the same set of words are used repetitively instead of a more diverse lexical selection. Muñoz-Ortiz et al. (2024) and Andre et al. (2023) also revealed that humans use a wider range of vocabulary than LLMs. This increasing trend in word repetitions and including words with lower lexical diversity was interpreted as the LLMs' main working principle, which is the tendency towards featuring vocabulary with the highest probability of occurrence (Hu & Zhou, 2024, cited in Tudino & Qin, 2024) rather than making use of a more diverse selection of vocabulary.

On the other hand, contradicting results were reported by Zhang and Crosthwaite (2026), Yeadon et al. (2023), Bao et al. (2025) and Zindela (2023), who found that LLM-generated texts have a higher lexical diversity, evidenced from the occurrence of a broader range of vocabulary. The TTR results of the current study are more in line with these researchers; in that, GPT-generated master's theses were found to have significantly higher TTR values than human-written theses. In the case of the present study, the results could be attributed to how advanced and lengthy the target texts are, in which humans could be facing more lexical limitations, thereby inclining more towards the repetition of the same sets of words, unlike the human-written texts in previous studies which had higher TTR values. Aside from this overall human and AI-generated text comparison, it should be indicated that the TTR levels of human-written master's theses show a larger range of value distribution than AI-generated texts, which might signify a more natural, human-like and heterogenous writing behavior, unlike the LLM writing, whose TTR values are compact and less varying, though higher.

Another feature which measures lexical diversity is hapax legomena ratio, a higher value of which corresponds to a higher lexical diversity. The current study found that although AI-written texts have higher median values of hapax legomena overall, this difference was not statistically significant. Therefore, it could be argued that the year variable does not play a determinant role in the changes observed for the hapax legomena ratio. Rather, the course variable demonstrates significant differences, which were observed with lower median values in human-written Language Sciences theses, and higher median values noted in LLM-generated Environmental Sciences theses.

These overall findings on hapax legomena do not consistently replicate the results of Culda et al. (2025), Opara (2024), Hamat (2024) and Schepens et al. (2023), who found that

machine-generated texts exhibit lower values of hapax legomena ratio. The higher ratio of unique words were previously connected to humans' ability to integrate their personal experiences to their texts, (Babiker, 2023; Jones, 2022; Kojima, 2024; Maamuujav et al., 2021; Verma & Srinivasan, 2019, cited in Culda et al., 2025) and their effort to avoid using the same lexical structures repeatedly (Simon et al., 2023; Yanagita et al., 2024, as cited in Terčon & Dobrovoljc, 2025). Although the current study did not find stable and interpretable differences between human and AI-generated theses, it should be noted that, as observed in TTR results, the overall data distribution is much more heterogenous in human-written theses, which could still be an implication of less predictable and more personal nature of human writing. Although humans appear to be resorting to using a more limited set of words in master's theses, they still show a certain level of lexical diversity through the wider range of unique words used, which is still not easily mirrored by AI.

Furthermore, rare word ratio results, which also provide insights into the lexical diversity of a text, demonstrate that GPT-generated texts have a higher proportion of rare words than human-authored master's theses in each year. Previous literature found conflicting results, where human-written poems exhibited a higher rare word ratio than LLMs (Hamat, 2024). Nonetheless, it could be argued that Hamat (2024) examines a very distinct text genre, in which GenAI may have encountered more challenges in imitating humans due to the creative nature of those texts.

LLMs' inclination towards featuring more rare words to display a more sophisticated writing style, which leads to unnatural texts with excessive ratios of rare words was indicated previously in the literature (Opara, 2024). The present study, therefore, confirms GPT's tendency to use a higher number of rare words in order to reach the intricate nature of human writing that is difficult to replicate. But still, lower values that were attained in the rare word ratios of human-

written master's theses further confirm the results of other lexical diversity features analyzed (i.e. TTR, hapax legomena, rare word ratio), denoting, overall, a lower lexical variety in human-authored master's theses, which is new information for this specific text genre.

According to Hewa Julige (2026), the lexical features showed the clearest differences and diachronic changes for the AI-Assisted Writing Era. Although AI-generated abstracts have specific linguistic qualities that are still clearly distinct from that of human-authored abstracts, Hewa Julige (2026) confirms that the presence of lexical and readability-related features that characterize fully AI-generated abstracts are gradually increasing in the modern-day human-written abstracts. The current thesis confirms this new phenomenon, demonstrating that some lexical features of human-written master's theses from the post-ChatGPT period gradually is becoming closer to the values in GPT-generated master's theses, especially evident in features such as adjective ratio, average word length, lexical density and some LLM-associated word frequencies.

Hewa Julige (2026) observed the most significant diachronic changes between 2021 and 2022. The initial research hypothesis of the current thesis also predicted that the year 2022 would start showing diachronic differences in terms of linguistic features, since it is the final year of the pre-ChatGPT era. However, in the analysis, the linguistic features which displayed statistically significant year-based differences mostly aggregated in 2024 and 2023. This relatively delayed observation of LLM-like features in human-written master's theses could be originating from a slow adaptation of GenAI tools in university contexts. Recognizing that master's degree dissertations are highly advanced final works, where authenticity plays a major role, the integration of LLMs in such texts may have occurred much more slowly than abstracts, or

argumentative essays, which were investigated more often in the previous literature (Bao et al., 2025; Yildiz Durak, et al., 2025).

Hewa Julige (2026) showed the part-of-speech diversity in human-written abstracts to be the lowest in the AI-Assisted Writing Era, specifically reporting a critical decrease in 2022, which, according to the author, implies that AI-assisted abstracts include the repetition of the same set of grammatical patterns and syntactic frames. However, the current study did not replicate these results. In fact, there was no statistically significant year-based difference among the theses in the human-only group and across AI and human-authored theses group for this feature. The absence of a substantial diachronic change regarding this feature could be signifying that this specific feature might be the most stable one, displaying a lower strength and role in distinguishing AI-generated and human-written texts. The scarcity of previous literature that found significant results in this measure might be further evidence for its relatively less considerable role in the comparison of LLM and human-written texts.

When it comes to the frequency of LLM-associated words in human-authored texts, the findings of Bao et al. (2025) revealed that there was a considerable diachronic rise in LLM-associated adjectives and adverbs in the human-written pre-print arXiv abstracts. The present study confirms their findings, since 38 LLM-associated words out of the predetermined list of 233 were exhibited to have statistically significant diachronic differences in frequency values among human-written theses. The fact that the usage of these AI-associated vocabulary especially increases in the year 2024 demonstrates the possibility that GenAI may have become gradually more integrated into human-authored academic writing after the introduction of ChatGPT. Although the initial research hypothesis of this thesis was that the LLM-associated word frequencies would start showing a boost in 2022, the interpretation to this delayed rise,

again, could be the slow integration, knowledge and awareness of AI tools in academic writing. These findings point out the potential for further increases in the AI-associated words in human-authored texts in the upcoming years, highly likely to be supported by the ever-growing popularity of GenAI tools.

5.2.Syntactic Features

Regarding syntactic features, Hewa Julige (2026) detected a decrease in average sentence length in the human-authored abstracts from the AI-Assisted Writing Era. Furthermore, as to the fully AI-generated content, Zindela (2023) and Culda et al. (2025) found that AI-generated texts have shorter sentences than human-written content. Bao et al. (2025) specified in their findings that the most significant decrease in syntactic complexity was observed in the Computer Science area, which reflects its close ties with LLMs.

The current study, on the other hand, has varying results in the year-based comparisons, because the LLM-generated master's theses were not consistently found to have higher average syntactic complexity values in each year. Rather, year-based differences were observed in human-written theses for the average tree depth and average branching factor features. Both these measures yielded statistically significant differences between two pre-ChatGPT and ChatGPT era human-authored thesis pairs, which are 2021 (pre-ChatGPT) and 2024 (ChatGPT) for the former, and 2020 (pre-ChatGPT) and 2023 (ChatGPT) for the latter. It needs to be noted that the different values in the ChatGPT era human-written theses for these syntactic complexity measures do not converge with LLM-generated thesis values. While the average tree depth observed in human-written theses in 2024 surpass the LLM-generated values in the same year, the average branching factors in 2023 are considerably lower than their AI-counterparts. In the scope of this thesis, the higher average tree depth values attested in human-authored theses from

the year 2024 could be an implication of how valuable and advanced results a possible AI-human collaboration in master's theses can yield, possibly overperforming the syntactic depth of fully LLM-generated writing.

The course-based comparison for all syntactic complexity measures, on the other hand, yielded more valuable findings. The Environmental Sciences theses had significantly lower values in terms of sentence counts, which were statistically similar across human and AI text condition. Moreover, the average sentence lengths, average dependency lengths and average tree depths were revealed to be considerably higher in LLM-generated Environmental Sciences theses, which might be denoting an overall higher syntactic complexity attributed by LLMs to the theses written in the scope of this master's program. Finally, the average branching factor measure, on the other hand, demonstrated higher values in the GPT-generated theses in the Language Sciences area, which were found to be statistically significant.

Overall, most of the syntactic features revealed higher values in the GPT-generated theses from the Environmental Sciences area, which might be attributed to a higher difficulty and complexity of field-specific academic writing characteristics that GPT-5 mini might have identified yet exceeded, possibly implying that LLMs are not completely competent in being sensitive to area-related syntactic characteristics. Even though there were no considerable diachronic changes in the human-written theses in terms of most syntactic complexity measures, it could be argued that these features have revealed area-specific discrepancies, as Bao et al. (2025) also observed in the texts they examined from different fields.

Because higher values observed in syntactic features show higher complexity and structural nesting, it could be argued that LLMs build sentences with more subordination and embeddedness than humans in the Environmental Sciences are. The current results, hence,

partially confirm the previous researchers who found higher syntactic complexity in AI-generated content (Culda et al., 2025; Herbold et al., 2023; Liu et al., 2023; Nkhobo & Chaka, 2023, cited in Terčon & Dobrovoljc, 2025).

The present study also revealed that the average sentence length values spread to a higher and larger range in human-written texts, resulting also in higher maximum values (though the differences in the medians were not statistically significant), while LLM-theses values are much narrower in distribution. This difference in data spread confirms the results of Muñoz-Ortiz et al. (2024), who found that the sentence lengths in AI-generated texts display a lower variation, having more uniform sentence lengths. This finding, like the lower variability in the lexical feature results of AI-generated texts that was reported in the previous subsection, confirms the notion that humans employ more varied linguistic resources that result in a more heterogeneous distribution, while LLM-written texts are much more consistent and homogenous. This type of concentrated value distribution in GPT-generated theses is no surprise, taking into account the main principle of LLMs, which is to imitate human language with the most predictable and consistent way, choosing the more probable characteristics over extreme or outlier options.

5.3. Readability Metrics

Regarding readability, Culda et al. (2025) observed that the readability scores achieved by both Flesch Reading Ease and Flesch-Kincaid Grade Levels metrics denote a higher readability ease in GPT-generated texts, when compared to their human-authored counterparts. On the other hand, Markowitz et al. (2024) reported that AI-authored texts have lower readability scores. Similarly, Pan, Oliveira and Mello (2025) argue that GPT-4 has lower readability ease due to employing more rare, complex and longer words. The current thesis exactly replicates these results; in that, LLM-generated master's theses have significantly lower values in the

Flesch Reading Ease metric. For this readability feature, there is a score range from 0 to 100, where low values indicate that the target text is difficult to read, whereas high values denote a higher ease in readability. In line with this scoring logic, it can be said that AI-generated master's theses were revealed to be more challenging to read than human-written theses. In addition, Pan et al.'s (2025) claim is also supported with the current results, because GPT-generated theses were indeed shown to have a higher rare word ratio and average word length (see Section 5.1), which leads to a lower ease in the readability of these texts.

Finally, the Flesch readability ease score of human-authored theses showed the lowest median values in 2024, which suggested closer scores to those of GPT theses. As a result of the signed gap calculation of the values over time, the differences between the Flesch readability ease scores of human-authored and LLM-generated theses were indeed shown to be lowest in 2024, which is a year from the ChatGPT era. In other words, it could be argued that human-written master's theses has the lowest and most LLM-like readability scores two years after the introduction of ChatGPT.

The interpretation of the Gunning Fog Index scores corresponds to the results of Flesch readability ease values, because the LLM-generated texts were again found to be more challenging to read, evidenced by the higher grade level requirement for the comprehension of these texts. Moreover, the year-based changes in the human-written master's theses exhibit that these texts became more difficult to read starting from 2022, which corresponds to the final year of the pre-ChatGPT period. However, since the Dunn's post-hoc test did not reveal pairwise year-based differences in the human-only condition, the decreasing values are merely showing a trend. On the other hand, the signed gap results for this feature demonstrated significant results for the year 2024, where human- and AI-written theses showed the lowest value difference,

which could be an indicator of a higher similarity between the texts from the two sources in the ChatGPT era in terms of readability.

5.4. Stylistic Features

Culda et al. (2025) showed that AI-generated texts feature a fewer ratio of passive voice usage when compared to human-written texts. Moreover, Reinhart et al. (2025) demonstrated that the agentless passive ratio in AI-generated texts is approximately half of the ratio in human-written theses. The findings of both studies were confirmed in the current thesis. It was demonstrated that in each of the six years analyzed, the passive voice ratio is considerably higher in human-written master's theses. An interesting finding is that the human-authored theses in 2024 was the only year in which the human-written set did not significantly differ from the passive percentages of GPT-generated theses, which suggests that the final year in ChatGPT era (2024) human-authored theses might be showing similarities to ChatGPT theses in terms of passive voice usage. Considering the scarcity of previous literature on the comparative proportion of passive voice usage in LLM and human-written texts, the current thesis presents valuable results not only in terms of the linguistic comparison of AI and human-written texts, but also regarding the year-based changes of these values in human-written master's theses.

With respect to the use of hedges, Almulla (2025) reported that LLM-generated texts had a considerably higher ratio of hedge words. On the other hand, in other studies, the use of hedging devices and modal verbs were found to be fewer in AI-generated content, because these linguistic tools are more frequently and typically used by humans due to their effort for reflecting their level of certainty and attitude, which LLMs lack (Mo & Crosthwaite, 2025, cited in Wu, 2025). The findings of the current thesis replicate this line of research by revealing that humans employ a higher ratio of hedge words in their texts when compared with AI-generated theses.

Moreover, the year-based changes in the hedge word ratio of human-written master's theses show similar results with passive voice; in that, the human source texts show the lowest hedge ratio value in 2024 (ChatGPT era), which was found to be significantly similar to the lower values attested in the LLM-generated theses. It should be noted that this similarity between the hedge word ratios of human-written theses in 2024 year and GPT-generated texts marks the introduction of ChatGPT as merely one of the possible reasons, instead of drawing a direct connection. Overall, it can be argued that the hedge word ratio could be one of the stylistic features that can distinguish between AI-generated and human-written texts.

About the capitalized word ratio, which aims to calculate the number of proper nouns, it was observed in the previous literature that AI-generated texts feature fewer proper nouns than human-authored texts (Chong et al., 2023; Desaire et al., 2023; Petukhova et al., 2024, cited in Terčon & Dobrovoljc, 2025). The results of the current thesis validate these findings, because AI-generated master's theses were shown to feature a considerably lower capitalized word ratio than human-written theses. This finding indicates GenAI's current incapacity in integrating real-world knowledge ("encyclopedic knowledge", as De Cesare, 2026 puts it) into its textual outputs, considering that proper nouns refer to people, entities or concepts that originate from the real world that the humans experience, while LLMs are only fed the data, which they are not completely able to successfully and realistically integrate in writing due to their prompt-dependent nature. Similarly, Sardinha (2024) observed that the LLM-authored texts in the academic genre do not employ a lot of narrative elements and explicit references, which also aligns with the results of the current thesis, since capitalized words (proper nouns) normally denote explicit and provable referents in the real world, which GenAI naturally avoids.

As to the use of stop words, the results are a mirror image of lexical density values. In other words, while lexical density (i.e. the ratio of content words) was higher in LLM-generated master's theses, the proportion of stop words is lower in these texts than human-written theses. This finding indicates that humans make a higher use of function words than LLMs, which might be signifying the natural quality attested in human speech and writing. Considering that LLMs generate outputs in seconds providing all the information that is requested in the prompt, the linguistic and textual features of such content has to be much more densely structured (Hewa Julige, 2026), whereas humans can convey their messages to the reader in a less informationally dense way, thereby using more stop words and fewer content words than GenAI. Hence, due to this significant difference in the nature of the two sources, a discrepancy in the stop word ratio was observed in the present study. It is also of major importance to report that the stop word ratios in human-written master's theses, just as observed in passive voice ratio and hedge word ratio, show the closest value to LLM-generated theses in the year 2024, which once again indicates a possible trace of ChatGPT in human-authored theses, among other probable factors.

Sprung (2025) found that the em-dash, which is the most typical punctuation mark associated with LLMs, showed an increase between 2023 and 2025 in human-written texts, which is the ChatGPT era. Surprisingly, no considerable diachronic increase in the use of em-dashes was observed in human-written theses in the current study. This finding may have a possible interpretation, which is the probability that human authors may have noticed the overuse of this punctuation symbol in LLM-generated texts, which may have discouraged them from frequently using this mark.

On the other hand, the dramatic decrease in the frequency of the question mark towards 2024 in human-authored master's theses and the overall lower values of this symbol in LLM-

texts could be signaling a possible GenAI trace or a diachronic influence of LLMs in human writing. The fact that no significant difference was found in terms of question mark frequency between human-authored and AI-generated thesis pairs in 2024 could be signifying a gradually increasing similarity with LLM-texts. Since the values are specifically getting closer in 2024, this could be implying that values that might be achieved in the future analyses (i.e. studies that can be conducted on theses written in 2025 and 2026) could yield even closer results between human and AI texts.

Finally, as to the Biber features, the higher ratio of nominalization in the LLM-generated theses that was demonstrated exactly confirms the findings of the seminal study of Reinhart et al. (2025). Moreover, the lower ratio of the perfect aspect usage in AI-written theses could be possibly signaling LLMs' avoidance from intermediary temporal features such as the perfect aspect. As to the year-based comparison, the Biber features "be" as main verb, predictive modals, predicative adjectives and analytic negation demonstrated more LLM-like values (which were shown to be typically lower) with dramatic decreases in 2024, while the nominalization and wh-clause features had values that are closer to those of LLM-generated theses (which were shown to be higher) in 2024 by showing an increase in that specific year.

The consistent observation of higher values of third person and first-person pronouns and the pronoun "it" in the course-based comparison values of human-authored theses rather than AI-written ones was not surprising, considering the previous findings which denote the personal experience factor that human writing has (Sardinha, 2024; Zhang & Crosthwaite, 2026), and LLMs' adoption of a more impersonal style than human-authored content (Goulart et al., 2024, cited in Terčon & Dobrovoljc, 2025). Furthermore, the higher use of other stylistic devices such

as specifying referents (using “it”, “this”) exhibits humans’ endeavor to ensure textual coherence (Halliday & Hasan, 1976, cited in Wu, 2025, p. 104).

Aside from these specific features, it was shown in the present study that AI and human-written texts substantially differ in terms of 63 features out of 67 in the tag set. This finding implies that although AI and human-written texts may be gradually becoming blended with one another, the stylistic choices they make can still help in distinguishing the texts from these two sources, much better than the values of syntactic complexity features.

6. Conclusion

This thesis aimed to answer two research questions. The first objective was to explore whether LLM-associated words displayed an increasing frequency in the Ca’ Foscari University master’s theses after the introduction of ChatGPT in 2022. In order to discover the diachronic changes in AI-associated vocabulary frequencies, the current study collected master’s theses from the Ca’ Foscari University of Venice database from six years, ranging from 2019 to 2024. To reveal the influence of ChatGPT in the vocabulary used in master’s theses, the six years were separated into two periods, which are pre-ChatGPT (i.e. 2019, 2020, 2021 and 2022) and ChatGPT (i.e. 2023 and 2024) eras. It was found that 38 words out of a pre-established LLM-associated frequency list of 233 words showed a statistically significant increase in 2024 in the human-written master’s theses, demonstrating pre-and post-ChatGPT variances in word usage. These results partly confirm the present hypothesis; in that, a notable number of LLM-associated words indeed had a higher frequency in the human-written theses from the ChatGPT era. However, the expected start of increase in the occurrences was actually observed to be mostly around 2023, reaching the peak values in 2024. It could be argued that this finding is natural,

considering that ChatGPT was introduced at the end of 2022, which may not yet have produced significant effects on the word choices of students.

Secondly, the RQ2 in the current study aimed to make an in-depth analysis and comparison of the linguistic features in fully AI-generated and human-written texts. The main objective for this detailed examination was to explore whether statistically significant differences between the pre-and post-ChatGPT human-written texts exist in the master's thesis genre, and if so, to report which measures differed the most, which were hypothesized to help in distinguishing the texts created by LLMs and humans. The findings confirm the hypothesis, because the values of average word length, lexical density, Flesch reading ease score, Gunning fog index, stop word ratio, hedge word ratio, passive voice ratio and several Biber features indeed provided different results in the pre-ChatGPT and ChatGPT era human-authored theses.

The average word length, adjective ratio, lexical density and Gunning fog values were exhibited to be consistently higher in LLM-generated texts. These same features in human-written texts in year 2024 demonstrated the closest values to AI-generated theses, showing a considerable increase in this specific year of the ChatGPT era. Likewise, a consistently lower Flesch readability ease score, stop word ratio and passive voice ratio in LLM-theses, and the more similar scores of human-written theses in 2024 with these low values implies a further distinction between pre- and post-ChatGPT human-authored theses, and between AI-and human-written texts. Lastly, the hedge word ratio led to important findings in terms of the diachronic distinction between human-authored theses, showing lower values in 2024 (ChatGPT era) when compared to the pre-ChatGPT era thesis values. Since GPT-generated theses demonstrated lower values overall, the decrease in the hedge word ratios of human-authored master's theses could be

pointing to more LLM-like values in the ChatGPT era of human-authored theses, further confirmed with the signed gap results.

The specific increases in some feature values observed in 2024 for human-written theses, which show similarity with the values of AI-generated theses, however, should not be directly interpreted as evidence of GenAI use. Rather, the possibility of increasing AI use in master's theses could only be one of the other relevant factors that may be causing the increase in the values of LLM-like linguistic features, some of which are differences in student backgrounds (both academic and linguistic), textual requirements of each master's program, topic preferences and the overall complexity of works. Since these contextual elements also exist and their effects may not be eliminated easily, it appears to be challenging to make a clearer analysis of AI influence in human writing, and to draw a direct connection between higher LLM-like linguistic values in human writing and higher ChatGPT use.

6.1.Limitations

In addition to the very comprehensive linguistic features set, a detailed analysis of textual cohesion may have also yielded valuable results. However, due to the considerable length of texts in the current dataset, a manual annotation approach could not be adapted. Since there are not any widespread automatized methods for measuring textual cohesion that follow the comprehensive frameworks provided in the literature (e.g. cohesion theory by Halliday & Hasan, 1976, cited in Wu, 2025), the current thesis did not cover the employment of cohesive devices in human-written and AI-generated texts. Therefore, this thesis shows a clear need for the computational automatization of the extraction of cohesive devices given in widely used frameworks, which can pave the way for and ease the methodological implementations of further research.

Also, due to the lack of computational tools that remove all table and figure contents from a PDF file, such contents from human-written master's theses were not removed. This is a limitation of the current thesis, because the AI-generated theses did not include any table contents, which does not allow for a full comparability between the two sources of text. Similarly, although a Python function was used to clean the .txt files in the corpus from lines that only consist of numbers and irrelevant symbols, the function did not work consistently for all files, possibly due to the varying nature of the fields from which the texts were retrieved, and the overall formatting. Future studies can adapt a thoroughly comprehensive and consistent approach for data cleaning (numbers, symbols, tables, figures, etc.) and carry out analyses from fully cleaned human-written theses.

6.2. Practical Implications of the Current Thesis

The results that were received from this thesis could have significant implications for the future of academic writing. Being aware of the distinction between the features of AI-generated and human-written texts could significantly assist academic writing instructors in detecting machine-generated work (Jiang & Hyland, 2024, cited in Zhang & Crosthwaite, 2026). The findings of this thesis could also call for further modifications in the guidelines for the use of AI in university contexts. Considering the finding that 30% of researchers used AI tools for assistance while writing manuscripts since late 2023, and that Wiley allows the use of such tools as long as the researchers explain in which parts they used them, it is crucial that detailed guidelines be constructed, and AI-usage be tracked more thoroughly in academic publications (Gray, 2024). Moreover, according to Geng and Poibeau (2025), it is gradually becoming frequent to display the ways through which GenAI tools are used for drafting, post-modifications

and paraphrasing sections in academic publishing contexts, which implies that guidelines could be modified to assist better in regulating a more ethical use of LLMs in academic works.

References

- Agrahari, S., Bisht, S., & Sanasam, R. S. (2024, December). Text authorship attribution: Stylometric insights into human and llm-generated text. In *Proceedings of the 8th International Conference on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD)* (pp. 344-346).
- Almulla, N. (2025). The Use of Hedging Devices and Engagement Markers in AI-Generated and Human-Written Essays: A Corpus-Based Comparison. *Open Journal of Modern Linguistics*, 15(05).
- Amirjalili, F., Neysani, M., & Nikbakht, A. (2024, March). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. In *Frontiers in Education* (Vol. 9, p. 1347421). Frontiers Media SA.
- André, C. M., Eriksen, H. F., Jakobsen, E. J., Mingolla, L. C., & Thomsen, N. B. (2023). *Detecting AI authorship: Analyzing descriptive features for AI detection. Paper submitted at the seventh workshop on natural language for artificial intelligence, Rome, Italy.*
- Anthropic. (2026). *Claude (Claude Opus 5.5, September 2026 Version)* [Large language model]. <https://claude.ai/>
- Bagdasarov, S., & Alves, D. (2025, September). Like a human? a linguistic analysis of human-written and machine-generated scientific texts. In *Proceedings of the First Workshop on Natural Language Processing and Language Models for Digital Humanities* (pp. 38-47).
- Banat, M. (2024). Investigating the linguistic fingerprint of GPT-4o in Arabic-to-English translation using stylometry. *Journal of Translation and Language Studies*, 5(3), 65-83.

- Bao, T., Zhao, Y., Mao, J., & Zhang, C. (2025). Examining linguistic shifts in academic writing before and after the launch of ChatGPT: a study on preprint papers. *Scientometrics*, 130(7), 3597-3627.
- Batane, T. (2010). Turning to Turnitin to fight plagiarism among university students. *Journal of Educational Technology & Society*, 13(2), 1-12.
- Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511621024>.
- Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
- Biber, D. (1993). Using register-diversified corpora for general language studies. *Computational linguistics*, 19(2), 219-241.
- Biber, D., & Conrad, S. (2019). *Register, genre, and style*. Cambridge University Press.
- browndw. (2026). browndw/pybiber: v0.3.0 (Version v0.3.0) [Computer software]. Zenodo. <https://doi.org/10.5281/zenodo.18209886>
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021, August). All that's 'human' is not gold: Evaluating human evaluation of generated text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7282-7296).
- Culda, L. C., Nerişanu, R. A., Cristescu, M. P., Mara, D. A., Bâra, A., & Oprea, S. V. (2025). Comparative linguistic analysis framework of human-written vs. machine-generated text. *Connection Science*, 37(1), 2507183.

De Cesare, Anna-Maria (2026). *L'italiano sintetico dell'intelligenza artificiale generativa*.

Franco Cesati.

Fröhling, L., & Zubiaga, A. (2021). Feature-based detection of automated language models:

tackling GPT-2, GPT-3 and Grover. *PeerJ Computer Science*, 7, e443.

Geng, M., & Poibeau, T. (2025). On the detectability of LLM-generated text: What exactly is

LLM-generated text?. *arXiv preprint arXiv:2510.20810*.

Georgiou, G. P. (2026). What distinguishes AI-generated from human writing? A rapid review of

the literature. *Big Data and Cognitive Computing*, 10(2), 55.

Giray, L. (2026). AI humanizers and the crisis of information integrity: implications for scientific

writing. *Naunyn-Schmiedeberg's Archives of Pharmacology*, 399(8), 11181-11193.

Gray, A. (2024). ChatGPT" contamination": estimating the prevalence of LLMs in the scholarly

literature. *arXiv preprint arXiv:2403.16887*.

Guo, B., Zhang, X., Wang, Z., Jiang, M., Nie, J., Ding, Y., ... & Wu, Y. (2023). How close is

ChatGPT to human experts? comparison corpus, evaluation, and detection. *arXiv preprint*

arXiv:2301.07597.

Guo, Y., Dou, Z., Nguyen, H. H., Chang, C. C., Sugawara, S., & Echizen, I. (2025, June).

Measuring human involvement in AI-generated text: A case study on academic writing.

In *2025 International Joint Conference on Neural Networks (IJCNN)* (pp. 1-10). IEEE.

Habib, F., Shafi, M. S. R., & Islam, M. U. (2026). AI vs Human Writers: A Comparative

Analysis of Story Rewriting through Readability Metrics. *International Journal of*

Technology in Education and Science, 10(2), 360-378.

Hamat, A. (2024). The language of AI and human poetry: A comparative lexicometric study. *3L,*

Language, Linguistics, Literature, 30(2), 1-20.

- Hashemi, A., Shi, W., & Corriveau, J. P. (2025). AI-generated or AI touch-up? Identifying AI contribution in text data. *International Journal of Data Science and Analytics*, 20(4), 3759-3770.
- Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative science studies*, 1(1), 387-395.
- Hewa Julige, M. (2026). *Linguistic Evolution of Education Research Writing: A Comparative Analysis Across Technological Eras and AI-Generated Abstracts* (Master's thesis, Itä-Suomen yliopisto).
- Honnibal, M., Montani, I., Van Landeghem, S., & Boyd, A. (2020). spaCy: Industrial-strength Natural Language Processing in Python. Zenodo.
<https://doi.org/10.5281/zenodo.1212303>
- Ippolito, D., Duckworth, D., Callison-Burch, C., & Eck, D. (2020, July). Automatic detection of generated text is easiest when humans are fooled. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1808-1822).
- Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, Daniel A. & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. *arXiv preprint arXiv:2403.07183*.
- Lin, M., & Liu, D. (2026). Intersubjectivity as the distinguishing feature or common ground: A contrastive study between human-written abstracts and LLM-generated abstracts. *Research Methods in Applied Linguistics*, 5(1), 100297.

- Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., ... & Hu, H. (2023). ArguGPT: evaluating, understanding and identifying argumentative essays generated by GPT models. *arXiv preprint arXiv:2304.07666*.
- Markowitz, D. M., Hancock, J. T., & Bailenson, J. N. (2024). Linguistic markers of inherently false AI communication and intentionally false human communication: Evidence from hotel reviews. *Journal of Language and Social Psychology*, 43(1), 63-82.
- Maurer, M. (2026, March). elfen: A python package for efficient linguistic feature extraction for natural language datasets. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 3: System Demonstrations)* (pp. 61-74).
- Muñoz-Ortiz, A., Gómez-Rodríguez, C., & Vilares, D. (2024). Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review*, 57(10), 265.
<https://doi.org/10.48550/arXiv.2308.09067>
- Nowacki, L., & Wrochna, A. E. (2025). ChatGPT theses. Identifying distinctive markers in AI-generated versus human-created texts: A multimodal analysis in university education. *E-Learning and Digital Media*, 20427530251331083.
- Opara, C. (2024, July). StyloAI: Distinguishing AI-generated content with stylometric analysis. In *International conference on artificial intelligence in education* (pp. 105-114). Cham: Springer Nature Switzerland.
- OpenAI. (2026). *ChatGPT* (September 2026 Version) [Large language model].
<https://chat.openai.com/>
- OpenAI. (2026). *gpt-5-mini-2025-08-07* (API) [Large language model]. <https://openai.com/api/>

- Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, CJ Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E.A. Quintero, Charles R Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors.
- (2020) **SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python**. *Nature Methods*, 17(3), 261-272. DOI: [10.1038/s41592-019-0686-2](https://doi.org/10.1038/s41592-019-0686-2).
- Pan, H., Araujo Oliveira, E., & Ferreira Mello, R. (2025, March). Exploring human-ai collaboration in educational contexts: Insights from writing analytics and authorship attribution. In *Proceedings of the 15th International Learning Analytics and Knowledge Conference* (pp. 903-909).
- Reinhart, A., Markey, B., Laudénbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, 122(8), e2422455122.
- Sardinha, T. B. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1), 100083.
- Schepens, J., Marx, N., & Gagl, B. (2023). Can we utilize large language models (LLMs) to generate useful linguistic corpora? A case study of the word frequency effect in young German readers. *Preprint from PsyArXiv*. https://doi.org/10.31234/osf.io/gm9b6_v5

- Sepehri, A., Mirshafiee, M. S., & Markowitz, D. M. (2023). PassivePy: A tool to automatically identify passive voice in big text data. *Journal of Consumer Psychology*, 33(4), 714–727. <https://doi.org/10.1002/jcpy.1377>
- Shalevska, E. (2024). Hedges and boosters in AI and human writing: A comparative analysis. *Knowledge-International Journal*, 65(5), 505-510.
- Singh, A., Sharma, D., Nandy, A., & Singh, V. K. (2024). Towards a large sized curated and annotated corpus for discriminating between human written and AI generated texts: A case study of text sourced from Wikipedia and ChatGPT. *Natural Language Processing Journal*, 6, 100050.
- Sprung, G. (2025). Estimating AI Usage by Comparing Changes in Wording and Occurrence of Phrases over Time. In *EDULEARN25 Proceedings* (pp. 5252-5261). IATED.
- Sudharshan, R., Shen, A., Gupta, S., & Zhang-Nunes, S. (2024). Assessing the utility of ChatGPT in simplifying text complexity of patient educational materials. *Cureus*, 16(3).
- Terčon, L., & Dobrovoljc, K. (2025). Linguistic characteristics of AI-generated text: A survey. *arXiv preprint arXiv:2510.05136*.
- Terpilowski, M. A. (2019). scikit-posthocs: Pairwise multiple comparison tests in Python. *Journal of Open Source Software*, 4(36), 1169, <https://doi.org/10.21105/joss.01169>
- Tian, Y. (2025). Syntactic complexity recognition and analysis in Chinese-English machine translation: A comparative study based on the BLSTM-CRF model. *PLoS One*, 20(6), e0325721.
- Toney, A., Bode, L., Ventura, T., Wilcox, E., & Singh, L. (2026). Comparing the Humanness of Machine-Generated and Human-Authored Text. *ACM Computing Surveys*, 58(12), 1-28.

- Tudino, G., & Qin, Y. (2024). A corpus-driven comparative analysis of AI in academic discourse: Investigating ChatGPT-generated academic texts in social sciences. *Lingua*, 312, 103838.
- Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31), 1026, <https://doi.org/10.21105/joss.01026>
- Varga, E. B., & Baksa, A. (2025). Exploring lexical variation through synonym sets in human and AI-written scientific texts. *Multidiszciplináris Tudományok*, 15(4), 48-57.
- Wu, J. (2025). A corpus-based multidimensional analysis of linguistic features between human-authored and ChatGPT-generated compositions. *International Journal of Linguistics, Literature and Translation*, 8(5), 102-110.
- Yao, G., & Liu, Z. (2025). Can AI simulate or emulate human stance? Using metadiscourse to compare GPT-generated and human-authored academic book reviews. *Journal of Pragmatics*, 247, 103-115.
- Yeadon, W., Inyang, O. O., Mizouri, A., Peach, A., & Testrow, C. P. (2023). The death of the short-form physics essay in the coming AI revolution. *Physics Education*, 58(3), 035027.
- Yildiz Durak, H., Eğin, F., & Onan, A. (2025). A comparison of human-written versus AI-generated text in discussions at educational settings: investigating features for ChatGPT, Gemini and BingAI. *European Journal of Education*, 60(1), e70014.
- Zelege, B., Soni, A., & Manikonda, L. (2025, May). Human or GenAI? Characterizing the linguistic differences between human-written and LLM-generated text. In *Companion Publication of the 17th ACM Web Science Conference 2025* (pp. 34-37).
- Zhang, M., & Crosthwaite, P. (2026). More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2

writers. *International Review of Applied Linguistics in Language Teaching*, 64(2), 733-760.

Zindela, N. (2023). Comparing Measures of Syntactic and Lexical Complexity in Artificial Intelligence and L2 Human-Generated Argumentative Essays. *International Journal of Education and Development Using Information and Communication Technology*, 19(3), 50-68.

Zou, Y., Kuek, F., Ng, K. H., & Cheng, X. (2026). Comparative analysis of text readability and writing styles in AI-generated vs. Human-written academic abstracts. *Plos one*, 21(4), e0343163.

Appendix

A. Prompts for Thesis Generation

Asking for a Summary of the Main Ideas of the Human-Authored Thesis:

I am giving you a master's thesis below. Please give me a short summary of its main ideas in a paragraph: {human-written thesis}

Starter Prompt:

You are generating a {thesis_word_count}-master's thesis. I need you to follow my guidelines.

Firstly, I want the thesis to be on the following summary which gives the main ideas:

Main Ideas: {main_ideas}

Then, I am providing you with a master's thesis blueprint which you should strictly follow, and which I will repeat in every prompt I give you. You will generate the thesis section by section. But, while generating sections separately, you must pay attention so that the thesis results in a coherent single document.

In each prompt, I will give you the blueprint. Then, I will ask you to generate the first section of the blueprint. When you generate it, I will save the output externally. Then, you will have the section you generated as the input, and give me its summary as the output. After, I will give you the blueprint again, along with the last chapter summary. You will follow the blueprint, and generate the next section coherently with the chapter summary you receive as the input. You will continue the loop until the whole thesis is generated section by section.

Prompts for Chapter Generation:

Title + Abstract

The master's thesis blueprint:

- Thesis title
- Abstract
- Introduction
- Literature Review
- Research question(s)
- Hypotheses
- Methodology
- Results
- Discussion
- Conclusion

Pay attention to the following criteria while writing the thesis:

- Make terminology decisions appropriate to the research topic.
- Use the APA 7th edition citation style.
- Do not give a references list at the end of the sections.
- Writing tone should be academic/formal.

The topic should be based on the following main ideas:

{main_ideas}

In light of all this information, generate the {title} and the {abstract}.

Introduction

The master's thesis blueprint:

- Thesis title
- Abstract
- Introduction
- Literature Review
- Research question(s)
- Hypotheses
- Methodology
- Results
- Discussion
- Conclusion

Pay attention to the following criteria while writing the thesis:

- Make terminology decisions appropriate to the research topic.
- Use the APA 7th edition citation style.
- Do not give a references list at the end of the sections.
- Writing tone should be academic/formal.

The topic should be based on the following main ideas:

{main_ideas}

In light of all this information, generate the {introduction} section of the thesis. Pay attention to the length. Since the whole thesis should be around {thesis_word_count} words, this section should be around {section_word_count} words.

Literature Review

The master's thesis blueprint:

- Thesis title
- Abstract
- Introduction
- Literature Review
- Research question(s)
- Hypotheses
- Methodology
- Results
- Discussion
- Conclusion

Pay attention to the following criteria while writing the thesis:

- Make terminology decisions appropriate to the research topic.
- Use the APA 7th edition citation style.
- Do not give a references list at the end of the sections.

- Writing tone should be academic/formal.

The topic should be based on the following main ideas:

{main_ideas}

Make it coherent with the previous chapter based on the following summary:

{previous_chapter_summary}

In light of all this information, generate the {literature_review} section of the thesis. Pay attention to the length. Since the whole thesis should be around {thesis_word_count} words, this section should be around {section_word_count} words.

Research Questions + Hypotheses

The master's thesis blueprint:

- Thesis title
- Abstract
- Introduction
- Literature Review
- Research question(s)
- Hypotheses
- Methodology
- Results
- Discussion

- Conclusion

Pay attention to the following criteria while writing the thesis:

- Make terminology decisions appropriate to the research topic.
- Use the APA 7th edition citation style.
- Do not give a references list at the end of the sections.
- Writing tone should be academic/formal.

The topic should be based on the following main ideas:

{main_ideas}

Make it coherent with the previous chapter based on the following summary:

{previous_chapter_summary}

In light of all this information, generate 3-4 {research_questions_and_hypotheses}. It is better if the number of research questions are the same with the hypotheses.

Methodology

The master's thesis blueprint:

- Thesis title
- Abstract
- Introduction
- Literature Review

- Research question(s)
- Hypotheses
- Methodology
- Results
- Discussion
- Conclusion

Pay attention to the following criteria while writing the thesis:

- Make terminology decisions appropriate to the research topic.
- Use the APA 7th edition citation style.
- Do not give a references list at the end of the sections.
- Writing tone should be academic/formal.

The topic should be based on the following main ideas:

{main_ideas}

Make it coherent with the previous chapter based on the following summary:

{previous_chapter_summary}

Consider the following research questions and the hypotheses:

{research_questions_and_hypotheses}.

In light of all this information, generate the {methodology} section of the thesis. Do not use special characters that the code can't escape. Pay attention to the length. Since the whole thesis should be around {thesis_word_count} words, this section should be around {section_word_count} words.

Results

The master's thesis blueprint:

- Thesis title
- Abstract
- Introduction
- Literature Review
- Research question(s)
- Hypotheses
- Methodology
- Results
- Discussion
- Conclusion

Pay attention to the following criteria while writing the thesis:

- Make terminology decisions appropriate to the research topic.
- Use the APA 7th edition citation style.
- Do not give a references list at the end of the sections.
- Writing tone should be academic/formal.

The topic should be based on the following main ideas:

{main_ideas}

Make it coherent with the previous chapter based on the following summary:

{previous_chapter_summary}

Consider the following research questions and the hypotheses:

{research_questions_and_hypotheses}.

In light of all this information, generate the {results} section of the thesis. Pay attention to the length. Since the whole thesis should be around {thesis_word_count} words, this section should be around {section_word_count} words.

Discussion

The master's thesis blueprint:

- Thesis title
- Abstract
- Introduction
- Literature Review
- Research question(s)
- Hypotheses
- Methodology

- Results
- Discussion
- Conclusion

Pay attention to the following criteria while writing the thesis:

- Make terminology decisions appropriate to the research topic.
- Use the APA 7th edition citation style.
- Do not give a references list at the end of the sections.
- Writing tone should be academic/formal.

The topic should be based on the following main ideas:

{main_ideas}

Make it coherent with the previous chapter based on the following summary:

{previous_chapter_summary}

Consider the following research questions and the hypotheses:

{research_questions_and_hypotheses}.

In light of all this information, generate the {discussion} section of the thesis. Please do not give limitations or concluding remarks yet, because those sections will be included in the conclusion section, not in the discussion section. Pay attention to the length. Since the whole thesis should

be around {thesis_word_count} words, this section should be around {section_word_count} words.

Conclusion

The master's thesis blueprint:

- Thesis title
- Abstract
- Introduction
- Literature Review
- Research question(s)
- Hypotheses
- Methodology
- Results
- Discussion
- Conclusion

Pay attention to the following criteria while writing the thesis:

- Make terminology decisions appropriate to the research topic.
- Use the APA 7th edition citation style.
- Do not give a references list at the end of the sections.
- Writing tone should be academic/formal.

The topic should be based on the following main ideas:

{main_ideas}

Make it coherent with the previous chapter based on the following summary:

{previous_chapter_summary}

Consider the following research questions and the hypotheses:

{research_questions_and_hypotheses}.

In light of all this information, generate the {conclusion} section of the thesis. Pay attention to the length. Since the whole thesis should be around {thesis_word_count} words, this section should be around {section_word_count} words.

Prompts for Chapter Summaries

Give me a 300-word-summary of the following text, including all the key ideas. Your summary will be used in the continuation of the master's thesis, as a groundwork for the next chapter to ensure coherence. Do not start your summary with the structure "this thesis" because it will be a previous chapter summary, not a thesis summary.

{chapter_content}

B. Biber Tag Set

Feature	Description
-	A. Tense and aspect markers
f_01_past_tense	Past tense
f_02_perfect_aspect	Perfect aspect
f_03_present_tense	Present tense
-	B. Place and time adverbials
f_04_place_adverbials	Place adverbials (e.g., <i>above, beside, outdoors</i>)
f_05_time_adverbials	Time adverbials (e.g., <i>early, instantly, soon</i>)
-	C. Pronouns and pro-verbs
f_06_first_person_pronouns	First-person pronouns
f_07_second_person_pronouns	Second-person pronouns
f_08_third_person_pronouns	Third-person personal pronouns (excluding <i>it</i>)
f_09_pronoun_it	Pronoun <i>it</i>
f_10_demonstrative_pronoun	Demonstrative pronouns (<i>that, this, these, those</i> as pronouns)
f_11_indefinite_pronoun	Indefinite pronouns (e.g., <i>anybody, nothing, someone</i>)
f_12_proverb_do	Pro-verb <i>do</i>
-	D. Questions
f_13_wh_question	Direct <i>wh</i> -questions
-	E. Nominal forms
f_14_nominalization	Nominalizations (ending in <i>-tion, -ment, -ness, -ity</i>)
f_15_gerunds	Gerunds (participial forms functioning as nouns)
f_16_other_nouns	Total other nouns

Feature	Description
-	F. Passives
f_17_agentless_passives	Agentless passives
f_18_by_passives	<i>by</i> -passives
-	G. Stative forms
f_19_be_main_verb	<i>be</i> as main verb
f_20_existential_there	Existential <i>there</i>
-	H. Subordination features
f_21_that_verb_comp	that verb complements (e.g., <i>I said [that he went].</i>)
f_22_that_adj_comp	that adjective complements (e.g., <i>I'm glad [that you like it].</i>)
f_23_wh_clause	<i>wh</i> -clauses (e.g., <i>I believed [what he told me].</i>)
f_24_infinitives	Infinitives
f_25_present_participle	Present participial adverbial clauses (e.g., <i>[Stuffing his mouth with cookies], Joe ran out the door.</i>)
f_26_past_participle	Past participial adverbial clauses (e.g., <i>[Built in a single week], the house would stand for fifty years.</i>)
f_27_past_participle_whiz	Past participial postnominal (reduced relative) clauses (e.g., <i>the solution [produced by this process]</i>)
f_28_present_participle_whiz	Present participial postnominal (reduced relative) clauses (e.g., <i>the event [causing this decline]</i>)
f_29_that_subj	<i>that</i> relative clauses on subject position (e.g., <i>the dog [that bit me]</i>)
f_30_that_obj	<i>that</i> relative clauses on object position (e.g., <i>the dog [that I saw]</i>)

Feature	Description
f_31_wh_subj	<i>wh</i> - relatives on subject position (e.g., <i>the man [who likes popcorn]</i>)
f_32_wh_obj	<i>wh</i> - relatives on object position (e.g., <i>the man [who Sally likes]</i>)
f_33_pied_piping	Pied-piping relative clauses (e.g., <i>the manner [in which he was told]</i>)
f_34_sentence_relatives	Sentence relatives (e.g., <i>Bob likes fried mangoes, [which is the most disgusting thing I've ever heard of].</i>)
f_35_because	Causative adverbial subordinator (<i>because</i>)
f_36_though	Concessive adverbial subordinators (<i>although, though</i>)
f_37_if	Conditional adverbial subordinators (<i>if, unless</i>)
f_38_other_adv_sub	Other adverbial subordinators (e.g., <i>since, while, whereas</i>)
-	I. Prepositional phrases, adjectives and adverbs
f_39_prepositions	Total prepositional phrases
f_40_adj_attr	Attributive adjectives (e.g., <i>the [big] horse</i>)
f_41_adj_pred	Predicative adjectives (e.g., <i>The horse is [big].</i>)
f_42_adverbs	Total adverbs
-	J. Lexical specificity
f_43_type_token	Type-token ratio (including punctuation)
f_44_mean_word_length	Average word length (across tokens, excluding punctuation)
-	K. Lexical classes
f_45_conjuncts	Conjuncts (e.g., <i>consequently, furthermore, however</i>)
f_46_downtoners	Downtoners (e.g., <i>barely, nearly, slightly</i>)
f_47_hedges	Hedges (e.g., <i>at about, something like, almost</i>)

Feature	Description
f_48_amplifiers	Amplifiers (e.g., <i>absolutely, extremely, perfectly</i>)
f_49_emphatics	Emphatics (e.g., <i>a lot, for sure, really</i>)
f_50_discourse_particles	Discourse particles (e.g., sentence-initial <i>well, now, anyway</i>)
f_51_demonstratives	Demonstratives
-	L. Modals
f_52_modal_possibility	Possibility modals (<i>can, may, might, could</i>)
f_53_modal_necessity	Necessity modals (<i>ought, should, must</i>)
f_54_modal_predictive	Predictive modals (<i>will, would, shall</i>)
-	M. Specialized verb classes
f_55_verb_public	Public verbs (e.g., <i>assert, declare, mention</i>)
f_56_verb_private	Private verbs (e.g., <i>assume, believe, doubt, know</i>)
f_57_verb_suasive	Suasive verbs (e.g., <i>command, insist, propose</i>)
f_58_verb_seem	<i>seem</i> and <i>appear</i>
-	N. Reduced forms and dispreferred structures
f_59_contractions	Contractions
f_60_that_deletion	Subordinator that deletion (e.g., <i>I think [he went].</i>)
f_61_stranded_preposition	Stranded prepositions (e.g., <i>the candidate that I was thinking [of]</i>)
f_62_split_infinitve	Split infinitives (e.g., <i>He wants [to convincingly prove] that ...</i>)
f_63_split_auxiliary	Split auxiliaries (e.g., <i>They [were apparently shown] to ...</i>)
-	O. Co-ordination
f_64_phrasal_coordination	Phrasal co-ordination (N and N; Adj and Adj; V and V; Adv and Adv)

Feature	Description
f_65_clausal_coordination	Independent clause co-ordination (clause-initial <i>and</i>)
-	P. Negation
f_66_neg_synthetic	Synthetic negation (e.g., <i>No answer is good enough for Jones.</i>)
f_67_neg_analytic	Analytic negation (e.g., <i>That isn't good enough.</i>)

Retrieved from: <https://browndw.github.io/pybiber/feature-categories.html>

C. LLM-Associated Word List

Top 100 Adjectives

commendable innovative meticulous intricate notable versatile noteworthy invaluable pivotal
 potent fresh ingenious cogent ongoing tangible profound methodical laudable lucid appreciable
 fascinating adaptable admirable refreshing proficient intriguing thoughtful credible exceptional
 digestible prevalent interpretative remarkable seamless economical proactive interdisciplinary
 sustainable optimizable comprehensive vital pragmatic comprehensible unique fuller authentic
 foundational distinctive pertinent valuable invasive speedy inherent considerable holistic
 insightful operational substantial compelling technological beneficial excellent keen cultural
 unauthorized strategic expansive prospective vivid consequential manageable unprecedented
 inclusive asymmetrical cohesive replicable quicker defensive wider imaginative traditional
 competent contentious widespread environmental instrumental substantive creative academic
 sizeable extant demonstrable prudent practicable signatory continental unnoticed automotive
 minimalistic intelligent

Top 100 Adverbs

meticulously reportedly lucidly innovatively aptly methodically excellently compellingly
 impressively undoubtedly scholarly strategically intriguingly competently intelligently hitherto
 thoughtfully profoundly undeniably admirably creatively logically markedly thereby
 contextually distinctly judiciously cleverly invariably successfully chiefly refreshingly
 constructively inadvertently effectively intellectually rightly convincingly comprehensively
 seamlessly predominantly coherently evidently notably professionally subtly synergistically
 productively purportedly remarkably traditionally starkly promptly richly nonetheless elegantly
 smartly solidly inadequately effortlessly forth firmly autonomously duly critically immensely
 beautifully maliciously finely succinctly further robustly decidedly conclusively diversely

exceptionally concurrently appreciably methodologically universally thoroughly soundly
particularly elaborately uniquely neatly definitively substantively usefully adversely primarily
principally discriminatively efficiently scientifically alike herein additionally subsequently
potentially

(Liang et al., 2024, pp.16-17)

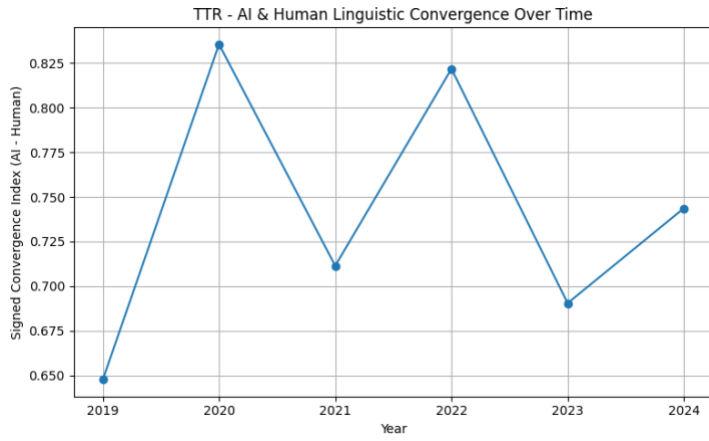
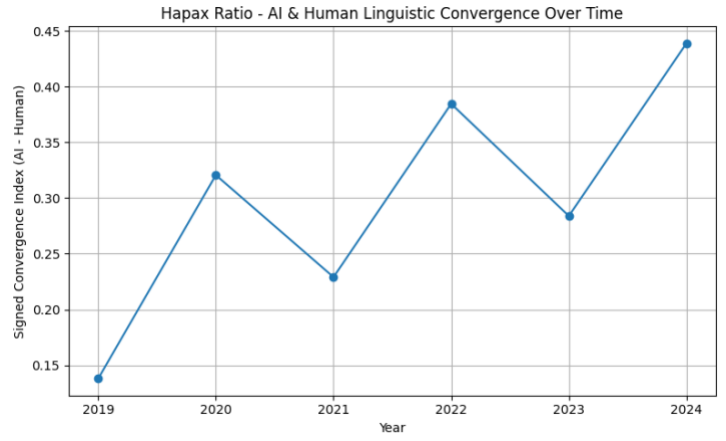
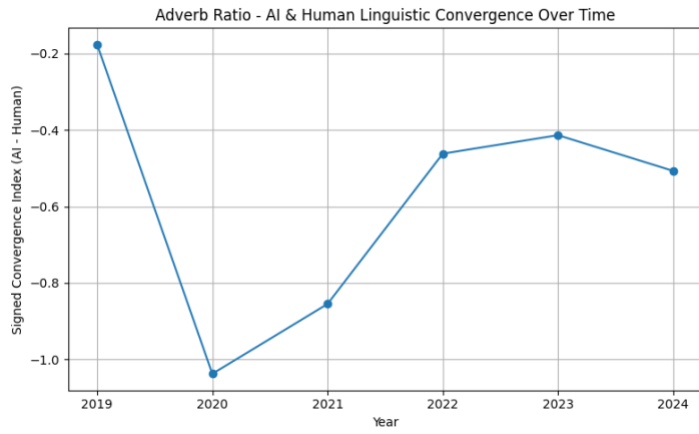
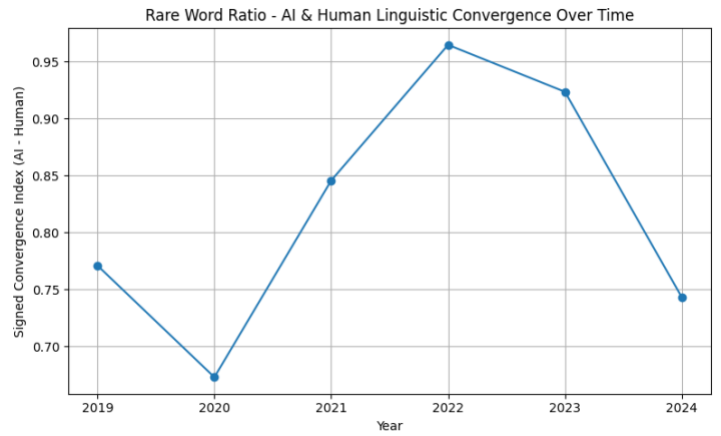
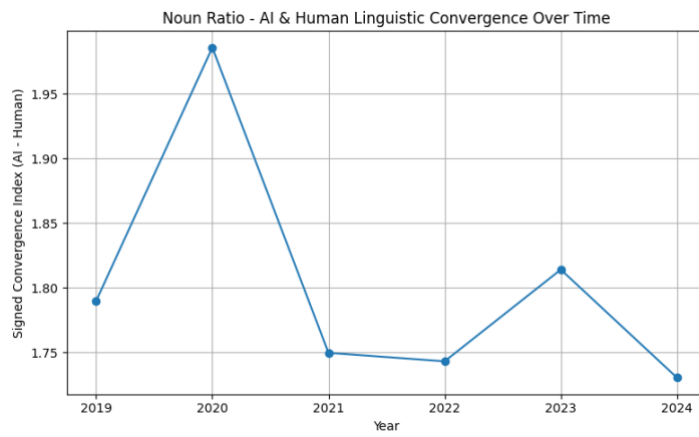
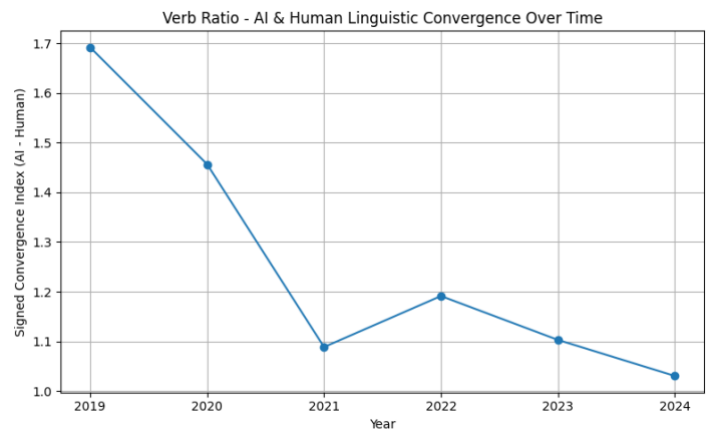
essential, particularly, represent, practices, increase, challenges, including, funding, providing,
primary, changes, related, efficiency, effective, across, complete, comprehensive, making,
towards, funding, mechanisms, successful, level, greater, challenging, objective, least, among,
beyond, insights, refers, effectively, phases, mind, initial, during, though

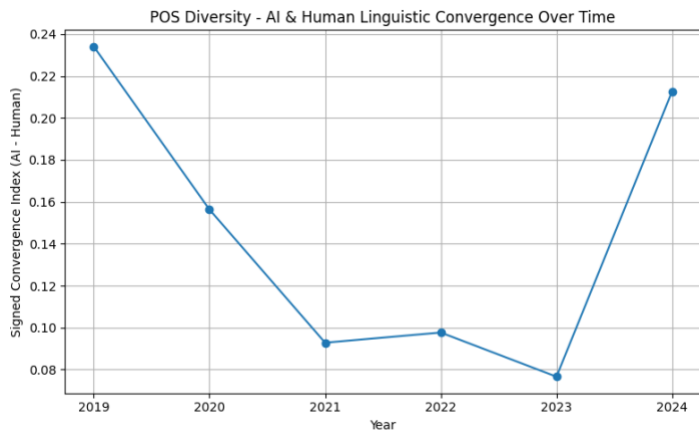
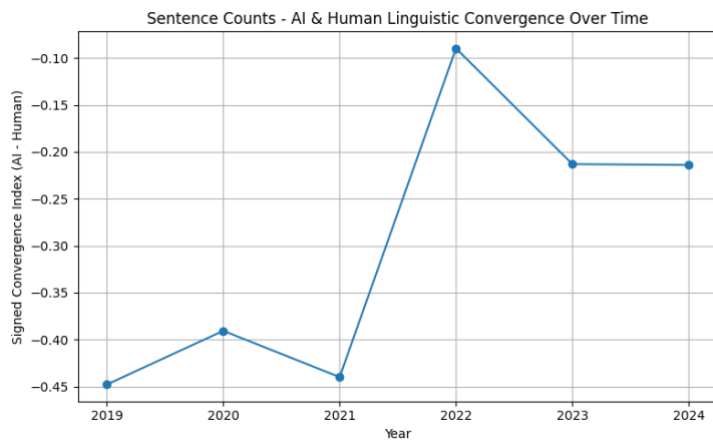
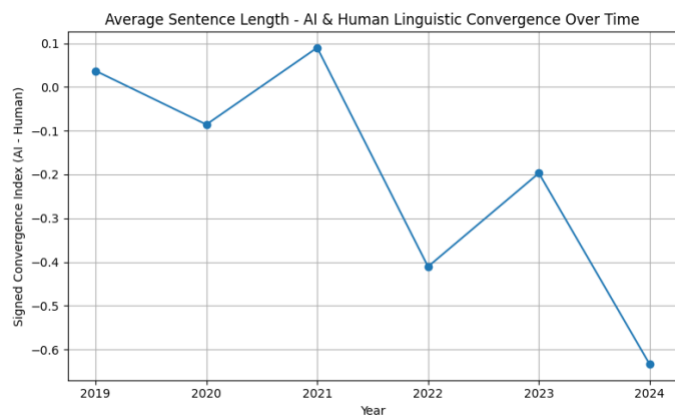
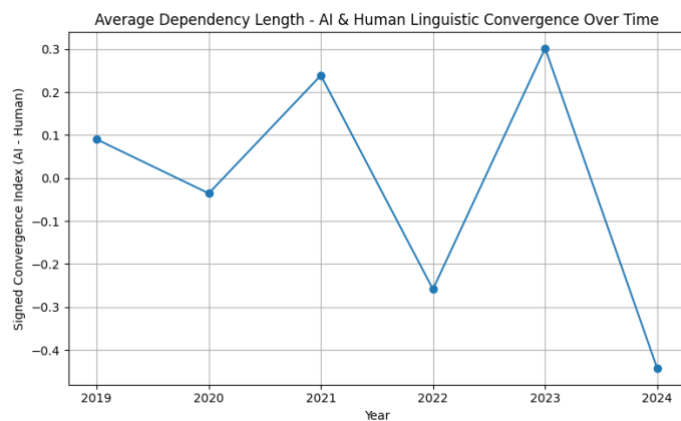
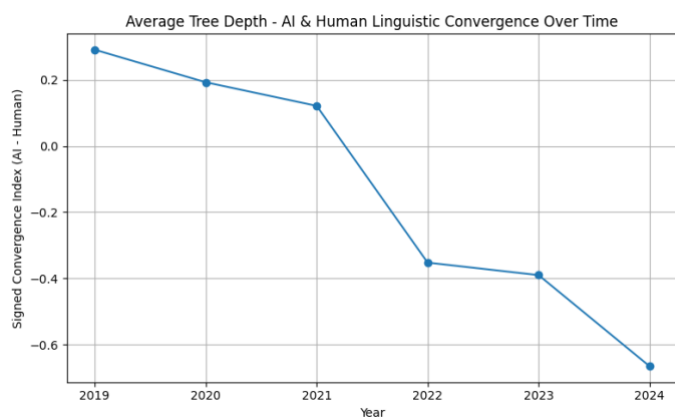
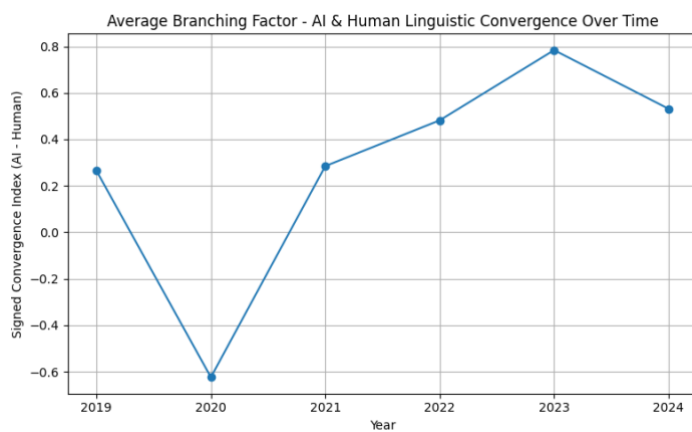
(Sprung, 2025, p.4)

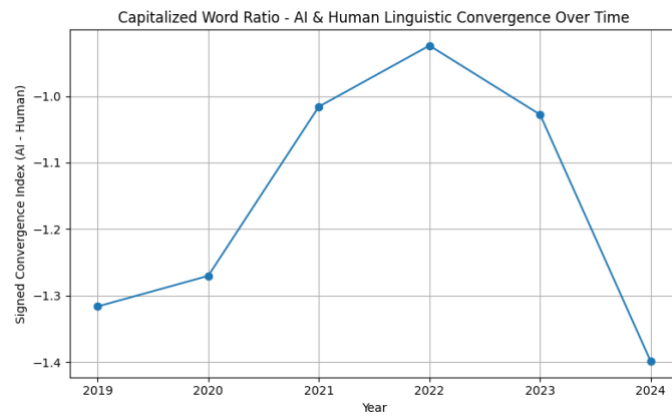
D. The LLM-Associated Words that Returned Zero Values in the Current Corpus

cogent	digestible	optimizable	speedy	demonstrable	minimalistic
lucidly	innovatively	compellingly	impressively	intriguingly	admirably
judiciously	cleverly	refreshingly	purportedly	maliciously	finely
soundly	substantively	usefully	discriminatively		

E. Non-Converging Values from the Signed Gap Calculation







Declaration on the Use of Artificial Intelligence (AI) Tools

The undersigned Sevilnur Bahadır, student ID number 907390, enrolled at Ca' Foscari in the Master's Degree Programme in Language Sciences, author of the final paper/degree thesis "A Comparative and Diachronic Study of Linguistic Features in Human-Written and LLM-Generated Theses: Following the Linguistic Traces of Generative AI in Ca' Foscari University Master's Theses", thesis supervisor Gianluca Lebani, aware of the penalties provided for in Article 76 of the Consolidated Act set out in Presidential Decree no. 445 of 28 December 2000, and of the loss of benefits provided for in Article 75 of the same Consolidated Act in the event of false or misleading statements, under his/her own responsibility pursuant to Articles 46 and 47 of Presidential Decree no. 445 of 28 December 2000,

DECLARES

that she **HAS** used generative artificial intelligence technologies to support various aspects of the research and writing of his/her final paper/degree thesis, in particular:

- AI Models Used: ChatGPT (September 2026 version), Claude Opus 5.5 (version released in September 22, 2026), and GPT-5 Mini (gpt-5-mini-2025-08-07, OpenAI API, released in August 7, 2025).
- ChatGPT and Claude Opus 5.5 were used for systematizing the statistical analyses and data visualization. gpt-5-mini-2025-08-07 was used for constructing the LLM-generated section of the corpus, which was part of the experiment and the research question to be tested.
- ChatGPT and Claude Opus 5.5 were used in the Results chapter for the visualization of the data presented. In addition, these models were also used for assistance in writing the Python code to get the statistical results in the Results chapter. gpt-5-mini-2025-08-07 was only used for corpus construction, since the research aim of the current thesis is to make a comparison between human-written and AI-generated texts. No raw texts generated via this tool are included in any of the sections of the thesis.
- The visual material (plots) was not directly generated through AI image generation. Instead, the Python code that enables the plot generation was assisted by ChatGPT and Claude Opus 5.5, and these codes were used to get plots accordingly. Therefore, all accuracy, quality and reliability checks were conducted by the detailed analysis of the code, paying attention to and ensuring that the values that are visualized on the plots match the actual data. Moreover, the statistical test procedures and results were also checked in detail, and it was ensured that all the tests were conducted accurately by following the actual instructions provided in relevant Python package guidelines. As to the corpus generation through gpt-5-mini-2025-08-07, since the AI-generated texts are only part of the texts that are linguistically analyzed in the scope of

this thesis, they do not pose accuracy, quality or reliability issues, considering that they are only part of the experiment that was conducted.

DECLARES FURTHER

that all material generated by AI has been carefully checked, reviewed and reworked by the author, who assumes full responsibility for the quality, accuracy and scientific integrity of the work.