Analysis of the impact of readily available AI on Academic Text: Lexical Diversity & Syntactic Complexity

FELIX JAN BENEDIKT VAN DELDEN, University of Twente, The Netherlands

The widespread adoption of AI writing tools, such as ChatGPT, has raised questions about their influence on academic writing, particularly among non-native English (L2) speakers. This study examines how these tools have shaped lexical diversity and syntactic complexity in 3223 bachelor's and master's theses from a technical department at a Dutch university, written in english, spanning from 2018–2025.

All theses from the department were analyzed, with three programs highlighted— including two technical programs (theses are completed in half a semester) and a creative technology program (theses are completed in one semester)—as these had the most published theses between 2018 and 2025. Theses from pre-LLM (2018–2022) and post-LLM (2023–2025) periods were contrasted using two lexical diversity metrics (vocd-D, MTLD) and three syntactic complexity metrics (mean sentence length, clauses per sentence, sentence length variation).

Results show a significant rise in lexical diversity after 2022: vocd-D increases by 7.5 % and MTLD by 11.6 %, suggesting broader vocabulary use with AI support. In contrast, syntactic complexity remains stable within narrow bounds across all programs (clauses per sentence: 0.75–0.78; mean sentence length: 15.2–15.6 tokens; sentence length variation: 8–10 words). Program-level patterns persist, creative technology theses exhibit higher subordination and sentence length, possibly influenced by their longer writing period and narrative style, while technical theses stay concise under tighter semester and format constraints.

These findings highlight a nuanced AI impact: richer vocabulary without greater syntactic complexity. Limitations include a single-department, predominantly L2 sample and unknown student backgrounds. Future work should expand to L1 writers, other disciplines, controlled AI- vs. human-written corpora, and model-specific analyses to strengthen AI-content detection methods.

Additional Key Words and Phrases: AI-generated text, lexical diversity, syntactic complexity, thesis analysis, AI-detection

1 Introduction & motivation

The release of ChatGPT 3.5 in November 2022 started an avalanche of unexpected proportions in the Artificial Intelligence (AI) industry. It showed that AI was more than science fiction, but something attainable and usable. In the past two and a half years thousands of Large Language Models (LLMs) and similar AI powered systems have been created to support humans in the completion of various tasks. These range from gene sequencing, code generation and text generation, to personalized learning and more [11].

Author's Contact Information: Felix Jan Benedikt van Delden, University of Twente, Department of Electrical Engineering, Mathematics and Computer Science, Enschede, The Netherlands, f.j.b.vandelden@student.utwente.nl.

TScIT 43, Enschede, The Netherlands

Academia and research are no exception, where AI helps students and teaching staff with the structuring of content, data management and analysis, idea generation and literature reviews [10]. The ethical use of these tools is a major concern in academia and research. As LLMs such as ChatGPT evolve, it is impossible to detect them with 100% accuracy, as they mimic human writing incredibly well and possess the capability to paraphrase, rewrite, and come up with new human-like text [4]. Solving the detection problem is crucial for universities to maintain their integrity and prevent plagiarism.

Meanwhile, the long-term implications of the use of AI are impossible to predict. However, with roughly two and a half years of increasingly widespread adoption of these tools, there is now enough data available to analyze its impacts to date. Looking at bachelor's and master's theses written in English by predominantly L2 speakers in a technical department at a University in the Netherlands, this study assesses the potential influence of LLMs on academic writing. This thesis addresses three research questions:

- **Lexical Diversity:** Has the range of vocabulary used in student theses expanded or contracted?
- **Syntactic Complexity:** Are AI-aided texts more syntactically complex than human-written texts?
- **AI Detection:** Can lexical diversity (LD) and syntactic complexity (SC) be used to detect AI-generated content?

To answer these questions, 3223 theses were used to create a dataset for analysis. It contains the full text of each thesis as well as metadata about the study program, year of publication, and reference ID. The publications range from 2018 to 2025. Henceforth, data from 2018-2022 will be labeled as pre-LLM and data from 2023-2025 as post-LLM.

Before the analysis, a brief discussion of related work will place this paper in the current research landscape. Next, the applied methodology to analyze LD and SC will be introduced and explained. Afterwards, the analysis, a discussion of the results, and their relevance with regard to existing research will follow. Finally, a brief conclusion will summarize the findings and their implications, as well as discuss future research questions that need to be investigated.

2 Background & Related Work

2.1 LLM impact on Lexical Diversity and written text

In the post-LLM years, a growing body of research has explored how LLMs such as ChatGPT, LLaMA, DeepSeek, and Gemini influence language use in human writing. Since this topic has become increasingly more relevant in the past two and a half years, the number of peer-reviewed research articles is limited. Addressing this deficit and finding out more about how AI is influencing writing is important to ensure that educators, writers, and policymakers can develop informed strategies on how to deal with and adjust to the technical advances of LLMs.

One study, that is yet to be peer reviewed, conducted a large-scale analysis of over one million arXiv abstracts to quantify the linguistic impact of LLM-style revisions. Their findings indicated increases in the frequency of "ChatGPT-favored" vocabulary and estimated LLM impact rates of up to 35% in computer science papers by 2024 [7].

Another study that has been peer reviewed compared LD between human-written and ChatGPT-generated texts. They showed that older models such as GPT-3.5 produced lower LD scores than human writers, while GPT-4 performed more similarly to humans, indicating that the models are improving in writing more like humans [25]. While these studies offer relevant insights, more research is required to understand how LLMs influence written expression and LD in texts.

2.2 LLM impact on Syntactic Complexity

Beyond the variety of vocabulary, changes in SC have also been investigated. As shown by recent work [22], which compared six LLM-generated and human-written newspaper articles, AI-generated texts often consist of more balanced sentence lengths, smoother syntax, and fewer disfluencies. This paper will analyze whether a similar trend can be observed in primarily L2 academic writing.

2.3 Detection of AI generated content

Closely linked to the previously discussed research is the task of detecting AI-generated content. As LLMs are increasingly used in academic, professional, and malicious contexts (e.g., phishing, misinformation), detection research has emerged as an important subfield. Several systems have been proposed for this purpose, including binary classifiers like Grover [27], GLTR [6], and TuringBench [26], as well as more recent stylometric approaches that analyze text structure and vocabulary usage [12].

Many detectors rely on statistical features, such as perplexity or surface-level fluency. Research has shown that such methods often misclassify L2-authored texts as AI-generated, since non-native writers tend to use simpler constructions and have lower lexical diversity [14].

To address these limitations, some researchers argue for the inclusion of deeper linguistic attributes (syntactic complexity, sentence variety, vocabulary richness) as inputs for more robust detection systems. A study showed that even when ChatGPT-generated student essays scored highly on quality rubrics, they contained linguistic patterns in sentence structure and lexical uniformity that could be traced to AI usage [9]. This suggests that the very metrics used in this study (LD, SC) may be useful not just for evaluating the impact of LLMs on writing, but also for future applications in AI-content detection.

3 Methodology

Moving on to the methodology, first the creation of the dataset will be discussed. Afterwards, the choice of the selected metrics to measure the LD and SC will be justified and the metrics themselves will be explained.

3.1 Dataset Construction and Preprocessing

The theses papers were scraped, using an automated Python script, from the official website of a Dutch University. Each scraped pdf file was converted to plain text using an automated pipeline based on `pdfminer`. Afterwards the following cleaning steps were applied:

(1) **Removal of non-textual elements:** Figures, tables, equations (in LaTeX or images), headers/footers, footnotes, and bibliography sections were stripped out using regular expressions (regex) and custom heuristics.
(2) **Normalization:** After cleaning, the text was converted to lowercase (hyphens and apostrophes within words were kept), with no additional punctuation removed beyond what the regex rules had already eliminated.
(3) **Tokenization:** When computing lexical-diversity metrics, the cleaned, lowercase text was split on whitespace. Each contiguous run of letters, digits, hyphens, or apostrophes became a single token.

3.2 Lexical Diversity Measures

3.2.1 Type-Token-Ratio. Many metrics have been proposed to calculate LD in speech and text. A basic metric is the Type-Token-Ratio (TTR), defined as the number of unique word types divided by the total token count:

$$\mathrm{TTR} = \frac{V}{N} \tag{1}$$

Here V is the number of distinct tokens and N is the total token count. While TTR was traditionally used to measure lexical diversity, its reliability decreases as text length increases, which is the reason for the development of a variety of alternatives [20]. Among them are vocd-D and Measure of Textual Lexical Diversity (MTLD).

3.2.2 Vocd-D. Vocd-D is a probabilistic measure of lexical diversity that estimates a parameter D representing the rate of vocabulary growth. Unlike TTR, vocd-D accounts for the fact that vocabulary size increases non-linearly with text length, making it more reliable for evaluating longer texts. The metric has been validated in peer-reviewed research and is widely used in computational linguistics [18, 19]. For this paper, vocd-D scores are computed using a python script that implements the standard algorithm. Readers interested in the underlying mathematical formulation are referred to [18] and [19].

3.2.3 MTLD. MTLD builds on TTR, calculating the average length of word sequences (called "factors") that maintain a TTR above a set threshold, typically 0.72 [20]. Once the TTR in a sequence falls below the threshold, a factor is counted and the TTR is reset. This process is performed both forward and backward through the text, and the final MTLD score is the average of both passes. Unlike TTR, MTLD has been shown to remain stable across varying text lengths and demonstrates strong construct validity across multiple domains [13]. It has been validated in many research studies and has been implemented in common text analysis libraries, making it a reliable and accessible tool for LD assessment. For a comprehensive technical description, see [20].

3.2.4 Summary. In sum, both vocd-D and MTLD correct TTR's downward bias in longer texts. Vocd-D does so by modeling vocabulary growth, while MTLD segments the text by a TTR threshold and computes the average segment length. Combined, these metrics provide a solid foundation for judging lexical richness of the academic-length texts in the dataset [18, 21].

3.3 Syntactic Complexity

Similarly to the LD metrics, the SC metrics were chosen due to their robustness with longer texts. Three established metrics used in L2 writing research will be applied. Firstly, mean length of sentences (MLS); secondly, clauses per sentence (C/S); and lastly, sentence length variation (SLV). The first two reflect an author's proficiency in a given language and have been proven to be reliable across genres and varying text lengths [15, 24]. The third has been used among other contexts, to distinguish human-written from AI-generated text, as AI-generated text tends to have less variation in sentence length [3].

Alternative syntactic complexity measures considered but not selected include mean length of T-unit (MLT) [16], which captures the average number of words per T-unit but is less reliable to automate over long texts, and the subordination ratio—mean number of dependent clauses per clause—[2], which isolates clause embedding yet overlooks coordination and phrasal complexity.

3.3.1 Mean Length of Sentences. MLS sometimes referred to simply as average sentence length, is defined as the total number of words (W) divided by the number of sentences (S) in the text:

$$\text{MLS} = \frac{W}{S} \tag{2}$$

Longer average sentences usually indicate greater syntactic elaboration, as writers combine clauses and phrases within single sentences. MLS has been widely validated as a reliable indicator of overall sentence complexity in both manual and computational studies [15, 24]. MLS was computed automatically by segmenting sentences and tokenizing words using spaCy's `en_core_web_sm` model, and then applying Equation (2) to compute the MLS score.

3.3.2 Clauses per Sentence. C/S measures the mean number of finite (clauses with a tense-marked verb that can stand alone) and non-finite (clauses with an infinitive, participle, or gerund verb form that cannot stand alone) clauses (C) per sentence (S):

$$\text{C/S} = \frac{C}{S} \tag{3}$$

A higher C/S ratio indicates more complex sentence structures with multiple embedded or coordinated clauses. Research in L2 writing has consistently found that more proficient writers use a greater number of clauses per sentence, implying advanced syntactic control [24].

Clauses are identified via spaCy's dependency parser[1], which implements a transition-based arc-eager algorithm: tokens are read into a buffer, shifted onto a stack, and then linked by applying `LEFT-ARC` or `RIGHT-ARC` operations based on learned neural scores, creating a tree where each token has a single head and a dependency label (`token.dep_`) [23]. Each parsed document is scanned and all tokens are counted for which `token.dep_` is one of the following clause categories, `advcl`, `ccomp`, `xcomp`, `csubj`, `csubjpass`, or `relcl`, giving the total clause count C. Dividing C by the sentence count S produces the C/S score [15].

3.3.3 Sentence Length Variation . The SLV also known as the standard deviation of sentence lengths quantifies the variability in the number of words per sentence within a text. It is defined as the population standard deviation of word counts across all sentences in a text:

$$\text{SLV} = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(L_i - \overline{L})^2} \tag{4}$$

Here, S is the total number of sentences, L_i is the word count of the i-th sentence, and $\overline{L}$ is the mean word count across all sentences (i.e., MLS). A higher SLV indicates greater diversity in sentence length, which may reflect a more dynamic or varied writing style, while a lower SLV suggests uniformity in sentence construction. This metric is particularly relevant for distinguishing writing styles, as human-authored texts often exhibit greater sentence length variation compared to machine-generated texts, which often produce more consistent lengths[3]. SLV has been established as an indicator of syntactic complexity in computational linguistics studies[1]. In this thesis, SLV is computed automatically by applying spaCy's `en_core_web_sm` model to segment sentences, extracting each sentence's word count with the regex `\w+`, and calculating the population standard deviation across those counts [8].

3.3.4 Summary. In conclusion, MLS and C/S provide complementary views on SC [15, 24]. SLV also adds to the evaluation of SC, while simultaneously indicating the potential influence of LLMs, yielding more uniform sentence lengths [3].The combination of the three metrics will give an indication whether the SC of L2 writers' in academia has changed in the post-LLM era.

4 Analysis

Throughout the analysis, the three study programs with the largest number of published theses are considered (Technical Study 1 - blue, Technical Study 2 - green and Creative Technology Study - orange) as well as an average of all combined study programs - red.

4.1 Lexical Diversity

4.1.1 VOCD-D Over Time. Figure 1 shows that approximate vocd-D scores have generally increased across all study programs from 2018 to 2025. The y-axis represents the vocd-D lexical-diversity score, where higher values indicate a broader variety of unique word usage within each thesis. All of the programs developed very similarly between 2018 and 2021 with scores ranging between 310 and 345. The low point of Technical Study 1 in 2018 has to be disregarded here, as there was a very limited amount of data available for that year. From 2022 onward an upward trend in all programs can be observed, rising from 322 to 348 in 2025 for all programs. All studies developed close to the average, except for Technical Study 2, which

[1] https://spacy.io/api/dependencyparser

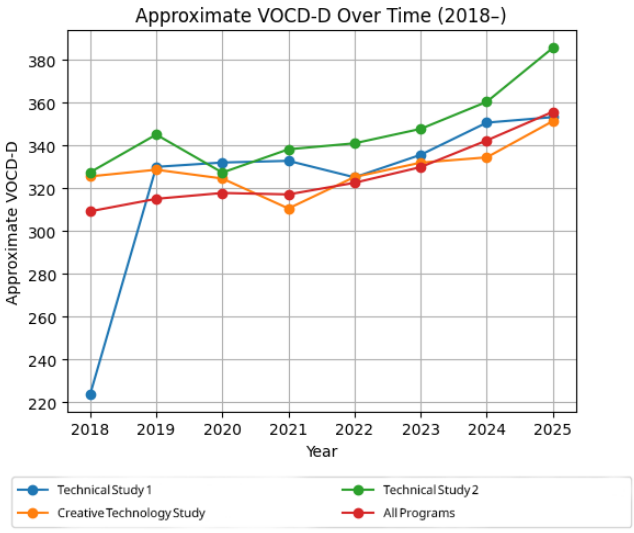

Fig. 1. VOCD-D over time (2018–2025) by program.

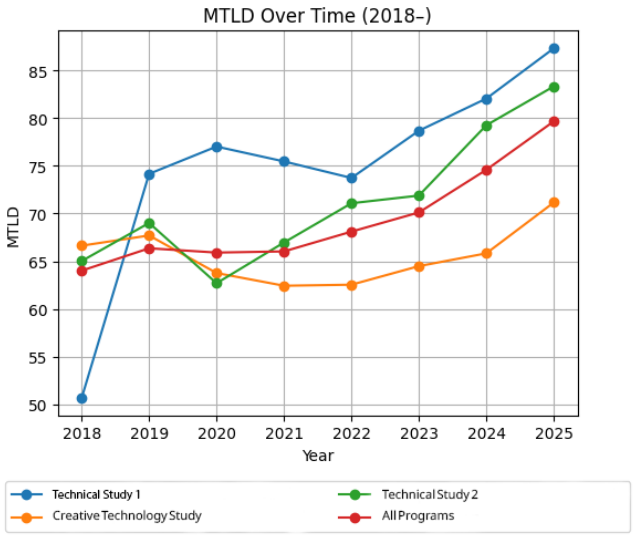

Fig. 2. MTLD over time (2018–2025) by program.

reached a global maximum in 2025 with 386. There are various possible reasons for this peak. It could be due to complex topics requiring a wider range of vocabulary to be researched, outside factors such as writing courses offered to students, or an increasing use of AI tools to aid in writing. It is likely that ultimately LLMs were used more frequently, as the vocd-D increased by 26 points between 2022-2025, marking a significant jump that goes beyond regular fluctuations observed in previous years.

4.1.2 Statistical Significance. To validate the increase as statistically significant across all programs, the distribution of vocd-D scores was tested for normality using the Shapiro–Wilk test. This indicated a significant departure from a normal distribution ($W = 0.9572$, $p = 1.7267 \times 10^{-31}$, $p < .05$).

Consequently, a Welch's two-sample t-test comparing vocd-D scores from pre-LLM theses to post-LLM theses was conducted. The mean vocd-D rose from $X_{\text{pre}} = 316.34$ ($n = 1800$) to $X_{\text{post}} = 339.86$ ($n = 1423$). Welch's t-test yielded $t = -11.091$, $p = 4.864 \times 10^{-28}$, indicating a statistically significant increase in lexical diversity ($p < 0.05$). Although the raw distributions departed from normality, with over 1,400 observations per group the Central Limit Theorem (CLT) ensures that the sampling distribution of the mean is essentially normal, so Welch's t-test remains valid [5]. This provides an indication that the rise in vocd-D after the introduction of LLM tools is unlikely to be due to chance, but rather the overarching influence of readily available AI-tools across all study programs.

4.1.3 MTLD Over Time. Figure 2 shows that for MTLD the trends are more variable across programs. The y-axis represents the MTLD score, where higher values indicate longer average token-sequence "factors" before the repetition threshold is reached, reflecting greater vocabulary richness. The Creative Technology Study consistently maintains low MTLD scores compared to the other programs, peaking in 2025 with a score of 71. In contrast, Technical Study 1 remains relatively high throughout the period (again disregarding 2018 due to limited data for that year). It peaks in 2025 similar to the vocd-D measure. While all programs showed fluctuations between 2018-2022, they increased continually from 2022-2025, reaching their peak in 2025. Thus, the global peak for all programs lies at almost 80 in 2025. Overall, MTLD seems to align with the implications indicated by vocd-D.

4.1.4 Statistical Significance. Again, to validate statistical significance for MTLD, the results were assessed for normality with the Shapiro–Wilk test. The test showed a clear deviation from a normal distribution ($W = 0.9143$, $p = 2.0648 \times 10^{-41}$, $p < 0.05$), similar to the non-normality observed for vocd-D.

Thus, a Welch's two-sample t-test was used to compare MTLD scores from pre-LLM to post-LLM theses. The mean MTLD rose from $X_{\text{pre}} = 65.59$ ($n = 1800$) to $X_{\text{post}} = 73.16$ ($n = 1423$). Welch's t-test yielded $t = -11.266$, $p = 7.977 \times 10^{-29}$, indicating a statistically significant increase in lexical diversity ($p < 0.05$). This parallel pattern to the vocd-D results suggests the post-LLM rise in MTLD is unlikely to be the result of random fluctuation and underscores the existence of a department-wide pattern.

4.1.5 Comparison and Interpretation. The differences between vocd-D and MTLD offer a wider perspective of how language seems to have evolved over the past 7 years than either of them alone. Although both metrics measure LD, they analyze slightly different aspects of it:

- **VOCD-D** estimates the variety of vocabulary, i.e., how many different words are used in a text, adjusted for the length of the text. It tends to reward variation and richness, even if concentrated in parts of the text.
- **MTLD** measures the consistency of this diversity throughout the text. It penalizes repetition or stylistic inconsistency, especially if certain sections reuse similar vocabulary.

This difference helps to explain why, for instance, Technical Study 2 shows a higher LD than Technical Study 1 measured by vocd-D, while for MTLD the opposite is the case. A possible interpretation is that students from Technical Study 2 used a broader vocabulary overall (increasing vocd-D), but this richness was not sustained evenly throughout the thesis (lowering MTLD). This unevenness could be due to modular writing practices. Another reason may be the use of AI-tools in a variety of ways. For example, to let AI write a paragraph, paraphrase existing text or provide structured bullet points, which are converted to text by the writer afterwards.

In summary, both measures indicate an increase in LD after 2022 across all programs. It should be mentioned that this is not to be considered conclusive evidence that LLMs were used by all students. There are many factors that are not considered due to a lack of available data. Among them are the backgrounds of the students. While there is data available to show that the majority of students at the observed University are L2 speakers, the exact numbers are not publicly available for each year.

Furthermore, these metrics can be influenced by many additional factors, such as complexity of a given topic, outside writing help offered to students, or a specific effort to use a greater variety of vocabulary by students, searching for synonyms or sophisticated adjectives to improve the overall quality of their texts. Nevertheless, the average increase does provide a meaningful indication that a broader impact has to have taken place, likely due to the advent of LLMs and AI-tools.

4.2 Syntactic complexity

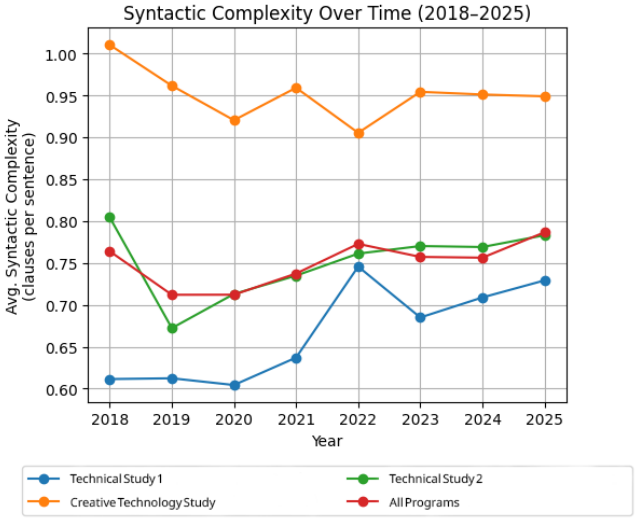

Fig. 3. Average clauses per sentence over time (2018–2025) by program.

4.2.1 Clauses per sentence Over Time. Figure 3 shows the average SC over time, measured by clauses per sentence. A value of 0.75 would mean, for example, that out of four sentences, three have a sub-clause. What is striking is the large difference between programs. While the Creative Technology Study leads with values between 0.9 and 1.01, Technical Study 1 only varies between 0.6 and 0.75. The high sub-clause count in the Creative Technology Study could be caused by a more narrative or discursive writing style, while the other purely technical studies are focused on brief descriptions of facts and statements. Furthermore, the Creative Technology Study is undertaken over the course of one semester, while the two technical studies only take half a semester to be completed. Also, while the technical studies have rigid guidelines on format and length of the final papers, this is not the case for the Creative Technology study, which could be an additional reason for the observed differences. The average across all programs fluctuates around 0.75, with Technical Study 2 developing very similarly to the average. Even though the mean of subclauses across all programs reaches a peak in 2025 with 0.78, this is only slightly higher than previous years such as 2018 and 2022 with 0.76 and 0.77 respectively. There is no clear indication that LLMs have had a notable impact on the S/C.

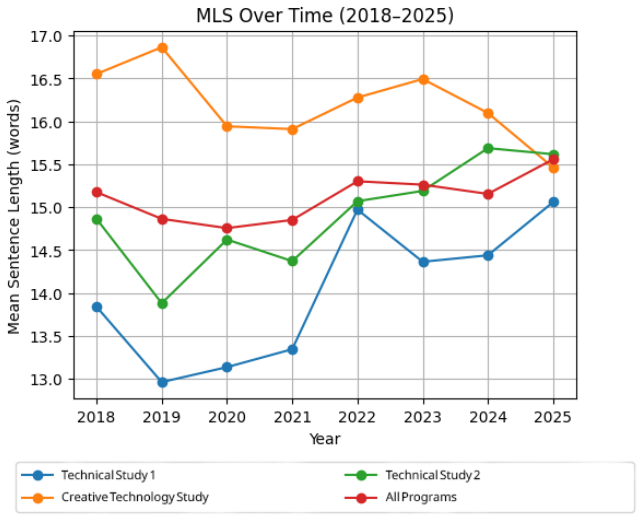

Fig. 4. Mean sentence length (tokens) over time (2018–2025) by program.

4.2.2 Mean Length Sentence. Figure 4 shows the MLS between 2018 and 2025. Similar to the clauses per sentence, the Creative Technology Study has the longest sentences on average. Varying between 15.48 tokens in 2025 and 16.8 tokens in 2019 the results align with S/C scores, which makes sense as MLS and S/C are correlated to each other [17]. Thus Technical Study 1 comes out with the shortest sentences on average, ranging between 12.98 tokens and 15.10 tokens in 2019 and 2025 respectively. Technical Study 2 develops closest to the average across programs, reaching a peak of 15.70 tokens in 2024. Across all programs, the longest sentence average was measured in 2025 with 15.57 tokens. Similar to S/C, peak values of 15.20 tokens in 2018 and 15.30 tokens in 2022 were recorded. The fluctuations between the years fall within the expected range, indicating that there is no significant shift in MLS post-LLM.

4.2.3 Sentence Length Variation. Figure 5 shows the year-to-year standard deviation of sentence length (in words) from 2018 to 2025. Technical Study 1 starts with relatively low variation (8.7 words)

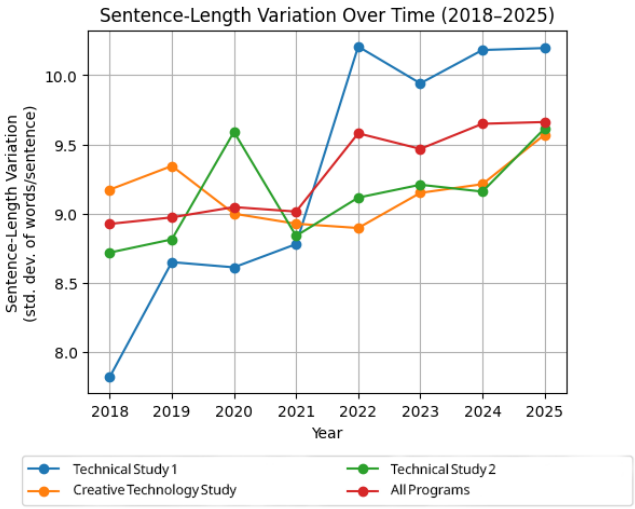

Fig. 5. Sentence Length Variation over time (2018–2025) by program.

in 2019, then climbs steadily, peaking at just over 10.2 words in 2022, before settling around that level up until 2025. The Creative Technology Study remains the most consistent, fluctuating modestly between 8.9 and 9.6 words, suggesting a stable narrative style across the observed timeframe. Technical Study 2 jumps in 2020 to about 9.6 words but then dips back to around 8.9 in 2021, gradually rising again to 9.6 by 2025. The aggregate trend mirrors these lower-variance programs, going up from 8.9 in 2018 to 9.7 in 2025. Although all programs exhibit a slight uptick in sentence-length variation after 2021, most pronounced in Technical Study 1, the fluctuations remain within a narrow range (8–10 words), showing no indication of post-LLM changes.

4.2.4 Comparison and Interpretation. The three measures (subordinate-clause rate (S/C), mean sentence length (MLS) and sentence-length variation (SLV)) together show a consistent picture: despite a clear rise in lexical diversity after 2022, sentence-structure complexity and variability have remained essentially stable across programs. The Creative Technology Study consistently ranks highest on both S/C (around 0.9–1.0 clauses/sentence) and MLS (roughly 15.5–16.8 tokens). Among other reasons, it could be reflecting its more discursive, narrative emphasis, whereas Technical Study 1 remains lowest (S/C = 0.6–0.75; MLS = 13.0–15.1 tokens), hinting at a preference for concise and fact-driven writing styles. Technical Study 2 and the overall average develop closely to each other, with only minor peaks in 2018, 2022 and 2025 that fall within expected year-to-year fluctuation.

SLV supports this pattern: variation stays in a narrow band (8–10 words) in all programs, with only a modest rise after 2021 and no sustained jump post-LLM. For example, the Creative Technology Study hovers between 8.9 and 9.6 words, while Technical Study 1 peaks near 10.2 words in 2022 before returning to values closer to the average again.

Taken together, these results imply that generative AI's suspected influence on vocabulary has not translated into more complex or varied sentence structures. Rather, program-specific conventions (narrative versus technical styles), assignment formats, editorial guidelines, and individual writing habits remain the primary drivers of syntactic complexity and variability.

5 Discussion

5.1 General Discussion

This study was aimed at quantifying the potential impact LLMs may have had on academic writing by predominantly L2 writers in a technical department at a Dutch university. The approach was to analyze metrics for LD and SC over a seven year period and measure if a statistically relevant shift in LD or SC took place, that could only be explained by a major global event, such as the widespread adoption of LLMs following their public release and integration into academic workflows around 2022–2023. The results of the study indicate an increase in LD, while SC remained stable, with differences only showing between different study programs, rather than timeframes within the programs themselves. Answering the research questions posed as follows:

5.2 Research Question 1

Has the range of vocabulary used in student theses expanded or contracted? The range of vocabulary has expanded. Both vocd-D and MTLD show significant post-LLM increases: overall vocd-D rose by approximately 7.5% (from 316.34 to 339.86; $p = 4.86 \times 10^{-28}$) and MTLD by 11.6% (from 65.59 to 73.16; $p = 7.98 \times 10^{-29}$), indicating a broader and more varied vocabulary in recent theses.

5.3 Research Question 2

Are AI-aided texts more syntactically complex than human-written texts? No. Measures of syntactic complexity—clauses per sentence (C/S), mean sentence length (MLS) and sentence length variation (SLV) fluctuate within narrow bands (C/S: 0.75–0.78; MLS: 15.2–15.6 tokens; SLV: 8.9-9.7) when averaging across all programs, with no systematic post-LLM increase. This shows that while the variety of words used has increased, the complexity of sentence structures and how many sentences are expanded still depend more on the rules of the academic field rather than help from AI tools.

5.4 Research Question 3

Can lexical diversity and syntactic complexity be used to detect AI-generated content? The variety of words, measured by vocd-D and MTLD, is useful for detecting changes, as texts written after LLMs were introduced show a significant increase in lexical diversity. SC on the other hand, did not show changes pre-LLM and post-LLM. The results should be treated with caution, as the data covers a very specific area of writing. Investigating L1 writing and differing domains may yield different results. These metrics must be calibrated to the specific characteristics and domains of the texts when applied to AI-generated content detection. Lastly, the approach will need further validation against texts that are known to be generated by AI.

6 Conclusion & Future Work

To better understand the impact of AI on academic writing and improve detection tools, several steps can be taken in future research.

First, using larger and more varied datasets would help. This means collecting theses from different universities and subjects, and comparing texts written mostly by native English (L1) speakers with those by non-native (L2) speakers. This can show how AI affects different groups of writers.

Second, creating standard collections of texts is important. These should include theses written only by humans and others fully generated by AI, using different AI models. Such collections would help identify typical patterns of word variety and sentence complexity for AI and human texts.

Third, studying specific AI models, like GPT-4 or Gemini, would be useful. Each model may leave unique traces in word choice and sentence structure, which can help detect AI use more accurately.

Finally, improving AI-detection tools is necessary. Adding measures of word variety and sentence complexity to existing systems, such as those based on text predictability or writing style, could make them better, especially for detecting AI in L2 writing.

These steps are crucial to move from observing trends to proving the impact of AI and building reliable tools for detecting AI-generated content.

References

[1] Lamia Berriche and Souad Larabi-Marie-Sainte. 2024. Unveiling ChatGPT Text Using Writing Style. *Heliyon* 10, 12 (2024), e32976. https://doi.org/10.1016/j.heliyon.2024.e32976 Accessed: June 20, 2025.

[2] Bram Bulté and Alex Housen. 2012. Defining and operationalising L2 complexity. In *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*, Alex Housen, Folkert Kuiken, and Ineke Vedder (Eds.). John Benjamins Publishing Company, 21–46. https://doi.org/10.1075/lllt.32.02bul Accessed: June 20, 2025.

[3] Heather Desaire, Aleesa E. Chua, Madeline Isom, Romana Jarosova, and David Hua. 2023. Distinguishing Academic Science Writing from Humans or ChatGPT with Over 99% Accuracy Using Off-the-Shelf Machine Learning Tools. *Cell Reports Physical Science* 4, 6 (2023), 101426. https://doi.org/10.1016/j.xcrp.2023.101426 Accessed: June 20, 2025.

[4] Ahmed M. Elkhatat, Khaled Elsaid, and Saeed Almeer. 2023. Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity* 19, 1 (2023), 17. https://doi.org/10.1007/s40979-023-00140-5 Accessed: June 20, 2025.

[5] Morten W. Fagerland, Stian Lydersen, and Per Kristian Laake. 2012. t-tests, non-parametric tests, and large studies—a paradox of statistical practice? *BMC Medical Research Methodology* 12 (2012), 78. https://doi.org/10.1186/1471-2288-12-78 Accessed: June 27, 2025.

[6] Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, Marta R. Costa-jussà and Enrique Alfonseca (Eds.). Association for Computational Linguistics, Florence, Italy, 111–116. https://doi.org/10.18653/v1/P19-3019 Accessed: June 20, 2025.

[7] Mingmeng Geng and Roberto Trotta. 2024. Is ChatGPT Transforming Academics' Writing Style? arXiv:2404.08627 [cs.CL] https://arxiv.org/abs/2404.08627 Accessed: June 20, 2025.

[8] João Gabriel Gralha and André Silva Pimentel. 2024. Gotcha GPT: Ensuring the Integrity in Academic Writing. *Journal of Chemical Information and Modeling* 64, 21 (2024), 8091–8097. https://doi.org/10.1021/acs.jcim.4c01203 Accessed: June 20, 2025.

[9] Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports* 13 (2023). https://doi.org/10.1038/s41598-023-45644-9 Accessed: June 20, 2025.

[10] M. Khalifa and M. Albadawy. 2024. Using Artificial Intelligence in Academic Writing and Research: An Essential Productivity Tool. *Computer Methods and Programs in Biomedicine Update* 5 (2024), 100145. https://doi.org/10.1016/j.cmpbup.2024.100145 Accessed: June 20, 2025.

[11] P. Kumar. 2024. Large language models (LLMs): survey, technical frameworks, and future challenges. *Artificial Intelligence Review* 57 (2024), 260. https://doi.org/10.1007/s10462-024-10888-y Accessed: June 20, 2025.

[12] Tharindu Kumarage and Huan Liu. 2023. Neural Authorship Attribution: Stylometric Analysis on Large Language Models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE, 51–54. https://doi.org/10.1109/CyberC58899.2023.00019 Accessed: June 20, 2025.

[13] Kate Kyle, Scott A. Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity using direct judgments. *Language Assessment Quarterly* 18, 2 (2021), 154–170. https://doi.org/10.1080/15434303.2020.1844205 Accessed: June 20, 2025.

[14] Weixin Liang, Mert Yuksekgonul, Yining Mao, Eric Wu, and James Zou. 2023. GPT detectors are biased against non-native English writers. *Patterns* 4, 7 (2023), 100779. https://doi.org/10.1016/j.patter.2023.100779 Accessed: June 20, 2025.

[15] Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15, 4 (2010), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu Accessed: June 20, 2025.

[16] Xiaofei Lu. 2010. Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics* 15, 4 (2010), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu Accessed: June 20, 2025.

[17] J. Lyu, M. I. Chishti, and Z. Peng. 2022. Marked distinctions in syntactic complexity: A case of second language university learners' and native speakers' syntactic constructions. *Frontiers in Psychology* 13 (2022), Article 1048286. https://doi.org/10.3389/fpsyg.2022.1048286 Accessed: June 20, 2025.

[18] David Malvern, Brian Richards, Ngoni Chipere, and Pablo Durán. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Palgrave Macmillan. https://doi.org/10.1057/9780230511804 Accessed: June 20, 2025.

[19] Paul M. McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing* 24, 4 (2007), 459–488. https://doi.org/10.1177/0265532207080767 Accessed: June 20, 2025.

[20] Paul M. McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods* 42, 2 (2010), 381–392. https://doi.org/10.3758/BRM.42.2.381 Accessed: June 20, 2025.

[21] Graham McKee, David D. Malvern, and Brian J. Richards. 2000. Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing* 15, 3 (2000), 323–337. https://doi.org/10.1093/llc/15.3.323 Accessed: June 20, 2025.

[22] Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. Contrasting Linguistic Patterns in Human and LLM-Generated News Text. *Artificial Intelligence Review* 57 (2024). https://doi.org/10.1007/s10462-024-10903-2 Accessed: June 20, 2025.

[23] Joakim Nivre. 2003. An Efficient Algorithm for Projective Dependency Parsing. In *Proceedings of the Eighth International Conference on Parsing Technologies*. Nancy, France, 149–160. https://aclanthology.org/W03-3017/ Accessed: June 27, 2025.

[24] Lourdes Ortega. 2003. Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics* 24, 4 (2003), 492–518. https://doi.org/10.1093/applin/24.4.492 Accessed: June 20, 2025.

[25] Pablo Reviriego, Javier Conde, Éric Merino-Gómez, Gonzalo Martínez, and Joaquín A. Hernández. 2024. Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. *Machine Learning with Applications* 18 (2024), 100602. https://doi.org/10.1016/j.mlwa.2024.100602 Accessed: June 20, 2025.

[26] Adaku Uchendu, Zeyu Ma, Thai Le, Rui Zhang, and Dongwon Lee. 2021. TURINGBENCH: A Benchmark Environment for Turing Test in the Age of Neural Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (Eds.). Association for Computational Linguistics, Punta Cana, Dominican Republic, 2001–2016. https://doi.org/10.18653/v1/2021.findings-emnlp.172 Accessed: June 20, 2025.

[27] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. 2019. Defending Against Neural Fake News. In *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (Eds.). Curran Associates, Inc., 9054–9065. http://papers.nips.cc/paper/9106-defending-against-neural-fake-news.pdf Accessed: June 20, 2025.