A COMPARATIVE ANALYSIS OF LEXICAL RICHNESS IN AI-GENERATED VS. HUMAN-AUTHORED LITERARY ARTICLES

Asia A. Alheety[1]
Methaq Kh. Marar[1]
Husam J. Mohammed[2]
[1] Department of English, College of Education for Humanities, University of Al-Anbar, Ramadi, Iraq
[2] Department of Artificial intelligence, Computer Sciences and Information Technology, University of Al-Anbar, Ramadi, Iraq
*Corresponding author email: asia90alheety@gmail.com

Abstract

The increasing prevalence of AI-generated texts raises critical questions about their linguistic authenticity compared to human-authored content. This study presents a comprehensive comparative analysis of lexical richness between AI-generated and human-authored literary articles using Halliday's systemic functional linguistics (SFL) framework. A corpus of 10 literary articles was selected (qualitative method, due to content analysis), comprising 5 human-authored texts and 5 AI-generated counterparts matched for length and genre. The analysis employed multiple lexical richness metrics including Type-Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), lexical density, lexical sophistication, and Halliday's three meta-functions (ideational, interpersonal, and textual), (quantitative method, due to metrics analysis). Depending on that this work will be mixed mode method. Using purposive sampling, texts were selected based on comparable length, genre, and publication standards. Results reveal significant differences between human and AI-generated texts across several dimensions. Human-authored texts demonstrated higher variability in lexical choices (TTR: 0.683 vs 0.671), greater sentence complexity (25.12 vs 22.29 words per sentence), and significantly higher ideational density (0.026 vs 0.010) and interpersonal density (0.031 vs 0.010). Conversely, AI-generated texts exhibited higher lexical density (0.771 vs 0.696), greater lexical sophistication (0.596 vs 0.436), and longer average word length (6.95 vs 5.94 characters). These findings suggest that while AI systems excel at producing grammatically sophisticated and lexically dense texts, they lack the functional variety and interpersonal engagement characteristic of authentic human communication. These findings have significant implications for AI development, automated writing assessment, and understanding the fundamental differences between human and artificial linguistic production.

Keywords: lexical richness, artificial intelligence, computational linguistics, systemic functional linguistics, corpus analysis, text generation.

1. INTRODUCTION

The rapid growth of AI text generation is changing our understanding of writing, prompting questions about human creativity and the validity of machine-produced content. The rapid advancement of artificial intelligence (AI) text generation has fundamentally changed how we understand writing, raising critical questions about the distinction between human creativity and machine-produced content. Advanced models like GPT-4 demonstrate impressive capabilities in generating coherent, contextually appropriate text across various domains and registers. Succeeding models of GPT-3 being developed by OpenAI (GPT-4 and so on), similar to (Sardinha, 2024) achieved impressive results in composing coherent responses that are appropriate across various languages, registers. Yet, notwithstanding its success over many benchmark tasks, the question of whether or not these systems can produce lexical subtlety akin to human-authored literary works is still an open empirical one. One of them is the lexical richness, that it is one a most posts important to measure the linguistic competencies and the creative expression in a text (Tweedie and Baayan, 1998). In a literary sense, however, richness of lexicon refers not only to the array of lexical choices but also to the way in which that vocabulary is utilised in order to construct particular communicative and aesthetic effects. For this reason, lexical richness measure and comparison between human generated text and AI-generated help us to have a sight on what each method can do well or the limit of an artificial text generation system. This paper fills an important gap in the literature by presenting a systematic empirical comparison of lexical richness between AI-generated literary articles and human-authored counterparts. Previous work has addressed a range of components contributing to the overall quality of AI-generated text such as coherence, fluency and grammaticality (Culda et al., 2025, Jeorgion, 2024) but relatively few studies have considered detailed lexical richness metrics that are theoretically-grounded in linguistics literature. Additionally, the majority of the literature has performed binary studies comparing AI vs. Human based on limited text sources such as academic abstracts or news articles and rare attention to more creative writing demanded in literary genres with an emphasis on artistic creativity and stylistic sophistication. This study has engaged with Halliday's (2004) systemic functional linguistics (SFL) as the foundational theory for this investigation because of its unique capacity to offer a fully comprehensive framework through which descriptions of language in use can be developed, examined and understood across a variety of registries contexts and functions (Halliday, 1994).

For comparative text analysis, Halliday's framework is ideal given that it provides a set of well-established ways to investigate lexical choices as they serve each of the three fundamental metafunctions of language: The ideational (experience and logical relations), the interpersonal (social roles and attitudes), and the textual functions (textual coherence and cohesion) (Hallidat and Matthiessen, 2014). This study follows three guiding research questions. In response to these research questions, the current study uses a corpus-based approach that exploits couples of human-authored and AI-generated literary entries for detailed multi-stance analysis. Human texts were selected from established literary and academic sources and AI counterparts were then automatically generated modified using state-of-the-art language models — AI systems that can understand, generate, and modify human language with incredible skill. Texts were analyzed lexiconically using both traditional measures as well as new measures based on Halliday's functional framework. This study addresses this gap by applying Halliday's systemic functional linguistics (SFL) framework to systematically compare lexical richness between AI-generated and human-authored literary articles. SFL's three metafunctions—ideational, interpersonal, and textual—provide a comprehensive theoretical foundation for analyzing how lexical choices serve different communicative purposes.

With the rise in Artificial Intelligence Communication Systems (AICS), educational institutions are being challenged by increasing difficulty in identifying AI work (Wu, 2025) and those faced by (Wu, 2025) require detection systems which are highly sophisticated. Likewise, publishers and content creators need to understand the qualitative difference between human-written text and scaled-up machine-produced texts in order to maintain quality, authenticity, and creativity. The results provided from this study inform scholarship in several domains. From the perspective of computational linguistics, this research reveals real-life data on what AI can do at present in terms of text generation and pinpoints where human creativity still outperforms machine performance. The study, a major component of empirical-lit intelligence-gathering, from the perspective of literary studies obtains clues about what cannot be done algorithmically in human creative expression. For pedagogy, the study offers hard numbers that educators can weigh and use to talk about or discuss human writing with AI-generated academic and creative outputs. The study investigates: (1) How do AI-generated literary articles compare to human-authored texts in traditional lexical richness metrics? (2) What differences exist in the functional distribution of lexical features across Halliday's metafunctions? (3) What do these patterns reveal about current AI text generation capabilities and limitations? While AI text generation and lexical analysis are well studied, little is known about how machine generated texts compare to human authored ones overall in terms of lexical richness, also taking into consideration the functional linguistic features. Although various text types (academic abstracts, news articles, or social media posts) have been studied before, literary works as a category of prose that requires more creativity and artistic language use has not received enough attention. Second, the functional framework in Halliday has not been well leveraged to assess extent of lexical density scales from human text to AI-generated texts by existing works which mostly focus on conventional metric-based lexical richness. This is an unfortunate omission, because the SFL model gives valuable clues as to how lexical choice contributes to meaning in various functional domains. This article attempts to fill this gap by performing a thorough comparative assessment of lexical richness in literary articles using both conventional and novel indices derived from Halliday's framework. This paper aims to contribute to the study of human creativity by considering text generation in an interval between artistic and cultural language practices, drawing from literary texts, as well as theoretical positionings and analysis on the field, thus providing new questions so that a better understanding of humans' nature can help improving AI-based generation technologies.

2. LITERATURE REVIEW

2.1 Lexical Richness in Computational Text Analysis

The notion of lexical richness figured heavily in the early days of computational text analysis, and as time marched forward, researchers developed more and more sophisticated algorithms for measuring such features with ever-increasing granularity. Tweedie and Baayen, 1998 influential study essentially laid down the ground rules for how to measure lexical richness by showing that traditional metrics, such as type-token ratio (TTR), are heavily affected by the size of a given sample (i.e. number of words) and developing alternative measures that tried to correct for this confound. For instance, their work indicated that although no measure is actually length independent, some measures, such as the MTLD (Measure of Textual Lexical Diversity), have more consistent values across different text sizes. This initial work set the stage for a wider range of measures to be considered under the rubric of lexical richness in follow-up research that tapped into various aspects of vocabulary use; Lexical density, or the ratio of content words to total words, was different enough to take into consideration but most importantly lexical density has been noted in Lithuanian as a feature that distinguishes spoken language from written text; the fraction of function words (e.g., pronouns) relative to total seconds is higher in spoken than written language (Halliday, 1989). Recent work on lexical sophistication measures, which quantify the difficulty or unusualness of an author's choice of vocabulary, shows their promise in assessing language skills and the overall quality of academic writing at large (Read, 2000). This piece of writing grapples with deceptively simple ideas from a recent intellectual world where computational linguistics aims at not just compound, lexical studies in terms of frequency and variation but semantic connections and even contextual appropriateness. Hou and Jiang, 2016 explored ways to correlate these [part of speech] features with traditional lexical metrics in Chinese texts and showed some interesting patterns in stylistic usage across different types of text, offering a bridge between older lexicon-based approaches and more sophisticated grammatical analysis Hou and Jiang, 2016. It was clear from this work that it is not sufficient to address text analysis of very similar phenotypes only by their lexical forms; the morphosyntactic features are also crucial.

2.2 AI Text Generation and Quality Assessment

The emergence of large language models has revolutionized the field of automatic text generation, with systems like GPT-3 and GPT-4 demonstrating unprecedented capabilities in producing coherent, contextually appropriate texts across diverse domains (Brown et al., 2020). However, the assessment of AI-generated text quality remains a complex challenge, requiring the development of new evaluation frameworks that can capture both surface-level linguistic features and deeper aspects of communicative effectiveness. Sardinha, 2024 multidimensional comparison of AI-generated and human-authored texts represents one of the most comprehensive attempts to systematically evaluate the differences between these two text types (Sardinha, 2024). In her research, Sardinha drew on Biber's dimensional analysis framework of linguistic variation and found substantial differences along five broad dimensions in AI-produced texts when compared to human language. This analysis showed that AI texts can be clearly distinguished from human-written examples on the basis of their multidimensional linguistic signature, implying distinct cognitive architectures for text production by humans and machines. Automated metrics: (Wu, 2025) extended the corpus-based analysis of linguistic features in human-authored versus ChatGPT-generated compositions, demonstrating systematic differences between these text types (Wu, 2025). Across four key dimensions—lexical difficulty, syntactic complexity, textual cohesion, and error patterns—Wu observed that human compositions showed gradual lexical complexity growth accompanied by academic level advancement; the texts generated by ChatGPT did not differentiate proficiency levels in a similar way. This suggests that present-day AI may be insensitive to the variability caused by (a) a learning phase and (b) the influence of all sorts environmental factors pertinent to authentic human communication. Additionally, Culda et al., (2025) presented a novel way of understanding AI text characteristics through their comparative linguistic analysis framework. In their research they found AI-generated texts have a more homogeneous style and less variability compared to human-authored texts, which is an important insight for both detection and quality assessment. Together, these studies imply that AI systems can be terrific at producing well-formed and coherent texts but might lack natural variety and the contextually sensitive way in which human communication fluctuates.

2.3 Halliday's Systemic Functional Linguistics Framework (SFL)

Systemic functional linguistics, established by Halliday, (1985), offers a useful theoretical foundation for the understanding of the way in which language operates to create meaning across text types. The Systemic Functional Linguistics (SFL) model is of particular interest to a comparative text analysis, as it provides a systematic way of studying and categorizing how lexical and grammatical choices contribute to the three major functions: ideational, interpersonal, and textual. The ideational metafunction relates to the representational role of language, particularly its potential to capture experience as well as logical relationships via choices in process types (material, mental, relational, behavioral, verbal and existential), participant roles and circumstantial elements (Halliday and Matthiessen, 2004). Texts with more ideational density higher levels of representational complexity and a possibly greater expressionistic quality on the part of the author in their characters. The interpersonal metafunction is to do with the different ways in which language can be used to carry out social human activity and express emotions, achieved by mood systems, modality markers and evaluative lexis (Martin and White, 2005). We can use interpersonal density as a kind of measure to see how much words prey on readers and judge, subjective attitudes which perhaps separate human creativity from automatic generation of text. The textual function, on the other hand, refers to how language can interweave property in order to generate coherent and cohesive discourse via theme structure, cohesive tie systems and information flow patterns (Halliday and Hasan, 1976). Textual density analysis gives us insight into the way in which human and AI authors organise ideas, and how they bring this order together within their texts. Tyrou on lexical diversity in EFL creative writing showed the potential of quantitative methods with which to comprehend some types of creativity through a second language. This research demonstrates the way in which lexical diversity indices can account for patterns of creativity in writing that reflect simultaneously linguistic competence as well as artistic intention, making this work a prototype of a language study over literary text.

3. METHODOLOGY

3.1 Research Design

This study use mixed mode method because it uses qualitative method (content analysis for the AI-generated and human authored articles) and quantitative method (Measuring Traditional Lexical Richness Metrics, Halliday's Functional Metrics).

3.2 Corpus Selection

The corpus of human-authored text and AI-generated texts were selected purposively base on the following criteria. (1) publication in peer-reviewed journals or recognized literary venues, (2) classification as literary or creative writing, (3) text length between 150-300 words to ensure adequate lexical content while maintaining analytical tractability, (4) representation of diverse sub-genres within literary writing, and (5) contemporary authorship (post-2000) to ensure relevance to current linguistic practices. Beside AI-generated texts were generated by (1) the most recent large language models (GPT-style) settings using between creativity and coherency. (2) Generation proceeded with category-specific prompts created to emulate the style, length, and thematic material of its human-authored counterpart. All of the AI-generated text was aimed to match the word count generated by AI, in order to keep an even playing field for length-sensitive metrics. Quality control included manual inspection of all texts generated for coherence, content consistency and compliance with genre conventions.

The final corpus consists of 10 texts total: 5 human-authored and 5 AI-generated, with matched pairs across five literary categories. The human corpus contains 1,052 words (average: 210.4 words per text), while the AI corpus contains an identical word count, ensuring perfect matching for length-dependent analyses. All texts are written in standard academic English and represent contemporary literary expression.

3.3 Data Analysis

3.3.1 Traditional Lexical Richness Metrics

To tackle this question in more analytic detail, several other well-known indices indicating lexical richness were calculated (different complementary aspects of vocabulary use and complexity):

Type-Token Ratio (TTR) is the basic measure of lexical richness, which indicates the number as the number of unique words (types)/ the total number of words (tokens). TTR is affected by length, but can serve as a base measure of the richness of vocabulary compared with prior evaluations.

Measure of Textual Lexical Diversity (MTLD) addresses the length-dependency limitations of TTR by calculating the average length of text required to reach a specified TTR threshold (0.72). MTLD provides a more stable estimate of lexical diversity across texts of varying lengths and has been validated across multiple languages and text types.

Lexical Density is the relative % of content words to total words (nouns, verbs, adjectives, adverbs), giving us an idea about how much information there is in texts. Notice that lower density is more typical of conversational styles, while higher lexical density usually means those formal and academic written registers.

Lexical Sophistication: measured by the use of more complex vocabulary, and in this case operationalized as word length (ie long-words) This metric captures a bias to use more complex, and possibly more accurate, tokens.

Hapax Legomena Ratio calculates the proportion of words that occur only once in the text, providing insight into vocabulary range and the author's tendency toward lexical repetition versus innovation.

3.3.2 Halliday's Functional Metrics

Building upon the SFL framework, the analysis incorporates novel measures designed to capture the three metafunctions of language:

Ideational Density measures the frequency of processes (material, mental, relational, behavioral, verbal, existential) relative to total word count. This metric captures the text's capacity to represent experience and logical relationships, reflecting the author's approach to depicting events, states, and relationships.

Interpersonal Density assesses the frequency of interpersonal markers including modality expressions (may, might, could, should, etc.) and personal pronouns. This measure reveals the extent to which texts engage with readers and express subjective perspectives, attitudes, and social relationships.

Textual Density examines the frequency of cohesive devices including conjunctive adjuncts (however, therefore, furthermore) and reference items (this, that, here, there). This metric illuminates how authors create textual coherence and guide readers through the logical structure of their arguments.

3.3.3 Additional Measures

The analysis also incorporates supplementary metrics that provide additional insight into lexical and syntactic patterns:

Average Word Length (in characters) offers a simple but effective measure of vocabulary complexity, with longer words typically indicating more formal or technical registers.

Sentence Complexity (average words per sentence) captures syntactic sophistication and the author's approach to information packaging and clause combining.

3.4 Data Analysis Procedures.

3.4.1 Text Preprocessing

All texts underwent systematic preprocessing to ensure consistency and accuracy in analysis. This process included removal of metadata headers, standardization of punctuation, and tokenization using regular expression patterns designed to capture word boundaries accurately. Function word lists were compiled based on established grammatical categories, with particular attention to Halliday's functional classifications (Halliday, 1976).

3.4.2 Statistical Analysis

Descriptive statistics were used in the comparative analysis, providing means, standard deviations and percentage differences between human text and AI-based text.

4 RESULTS

4.1 Overview of Findings

A comparative examination of lexical richness in the literary articles written by humans and those generated with AI, exposes systematic pattern differences on various vocabulary use along with functional linguistics aspects. Both types of text professionally use the English vocabulary and grammar; but there can be some specific patterns that differ from human creativity, versus machine creation. Results are presented under the categories of conventional lexical richness metrics, (Halliday, 1976) functional measures, and other linguistic features.

4.2 Traditional Lexical Richness Metrics

4.2.1 Type-Token Ratio and Lexical Diversity

Lexical diversity analysis uncovers subtle differences in human vs AI trained texts. Human-authored texts performed slightly better with a mean TTR of 0.683 (SD = 0.045), compared to AI-generated texts that had a score of 0.671 (SD = 0.023) — the gap here was just 1.7% in favor of human texts. Although this difference is small, it is similarly consistent for each individual text pair (human texts are considerably less uniform in diversity scores). It also allows to gain further information about lexical diversity patterns thanks to a MTLD analysis. To clarify, human texts had a mean MTLD of 152.6 (SD = 41.4) vs. AI texts' mean of 146.3 (SD = 8.4) — approximately a difference of about 4% on average Specifically, MTLD scores via human texts: 101.8–207.7, AI text: 138.5–160.1) Such a pattern can indicate that human authors have more scattered strategies of exhibiting lexical diversity compared to language models which hedge closer to steady and modest levels of vocabulary variation. See Table 1.

Table 1. Traditional Lexical Richness Metrics

Metric	Human-authored Texts (M ± SD)	AI-generated Texts (M ± SD)	% Difference (AI vs. Human)	Key Observation
Type-Token Ratio (TTR)	0.683 ± 0.045	0.671 ± 0.023	-1.7%	Human slightly higher diversity, more variability
MTLD	152.6 ± 41.4	146.3 ± 8.4	-4.4%	Human greater diversity and variability
Lexical Density	0.696 ± 0.046	0.771 ± 0.027	+9.7%	AI more informationally dense
Lexical Sophistication	0.436 ± 0.115	0.596 ± 0.053	+26.9%	AI uses more complex vocabulary
Average Word Length (chars)	5.94 ± 0.81	6.95 ± 0.4341	+14.5%	AI uses longer, more uniform words

Metric	Human-authored Texts (M ± SD)	AI-generated Texts (M ± SD)	% Difference (AI vs. Human)	Key Observation
Hapax Legomena Ratio	0.564 ± 0.060	0.530 ± 0.032	-6.6%	Human uses more unique words

4.2.2 Lexical Density and Sophistication

It is one of the key differences between bottled and free text when switching to a lexical denseness analysis. Here by contrast, the AI-generated texts showed a ~9.7% higher lexical density (M = 0.771, SD = 0.027) than human texts (M = 0.696, SD = 0.046). This trend is characteristic for AI, where the resulting texts are denser in terms of information as they contain more content words compared to function words. Lexical sophistication measures are no different, as well, with AI texts boasting lengthier lexis. Scores for sophistication were moderately higher in the AI sext database (M = 0.596, SD = 0.053) than human text (M= 0.436, SD = 0.115), indicating an approximately significant difference of 26.9%. The difference possibly stems from a bias in the way AI systems are trained on more formal written works—they seem to prefer fancier words. This pattern is confirmed also by the data on average word length (6.95 (SD = 0.43) for AI vs. 5.94 (SD = 0.81) for human texts; difference of 14,5%). Less variation in AI word length points to relatively constant complexity across the available vocabulary, whereas human text exhibits pronounced differences in lexical variety.

4.2.3 Hapaxes and Vocabulary Diversity

A hapax legomenon (plural: "hapax legomena") is a word that occurs once in an author's incomplete works. Human texts had higher average hapax ratios (M = 0.564, SD = 0.060) than AI texts (M = 0.530, SD = 0.032) — a difference equivalent to roughly +6.6 % more unique terms in human-generated content than AI-generated content. This attests that human writers circumstantiate through a wider range of words within each text and without repeating, as slightest at a semantic level but overall stale recycling of lexemes inside by AI systems (per.text.),is more%

4.3 Halliday's Functional Linguistic Measures

4.3.1 Ideational Metafunction

Ideational density: One of the biggest differences between human and AI texts, as shown by analysis Consistently, human-authored texts showed substantially more ideation density (M = 0.026, SD = 0.020) compared to AI texts (M = 0.010, SD = 0.009), with an impressive difference of around 162%. This significant space point out that the process human need to tap (material, mental, relational, behavioral, verbal, existential) are much more when compare with total word in case of human-authored.

The result has significant impact on how human and AI authors conceive of their experience or the facts/how they represent distant logical relationships. Texts produced by humans are more oriented towards action and experience, using a lot of the verbs and process-related vocabulary that you have in compelling stories. Even if their grammar can be good, AI texts are more stationary and less processual. Second, there is variability in the degree of ideational density when we compare across different types of text. Human = 0.008– 0.055 myriad ways of representing experience by different authors and in various contexts; AI texts: slightly tighter around their low mean, showing that process-oriented language is consistent but sparse. See Table 2.

Table 2. Halliday's Functional Linguistic Measures

Metafunction	Human-authored Texts (M ± SD)	AI-generated Texts (M ± SD)	% Difference (AI vs. Human)	Key Observation
Ideational Density	0.026 ± 0.020	0.010 ± 0.009	-162%	Human far richer in process-related language
Interpersonal Density	0.031 ± 0.019	0.010 ± 0.008	-195%	Human uses more pronouns & modality markers
Textual Density	0.036 ± 0.008	0.041 ± 0.022	+11.7%	AI uses more explicit cohesive devices

4.3.2 Interpersonal Metafunction

Interpersonal density analysis reveals another area where human texts significantly exceed AI counterparts. Human texts achieved a mean interpersonal density of 0.031 (SD = 0.019), while AI texts scored 0.010 (SD = 0.008), representing a 195% difference favoring human texts. This dramatic difference indicates that human authors employ far more modality markers and personal pronouns, creating texts that are more personally engaging and subjectively oriented.

The interpersonal findings suggest fundamental differences in how human and AI authors approach reader engagement and subjective expression. Human texts frequently employ modal verbs (may, might, could, should), personal pronouns (I, you, we), and other interpersonal markers that create a sense of dialogue and personal investment. AI texts, while coherent and informative, tend toward more objective, impersonal expression that maintains distance from both author and reader. This pattern has significant implications for understanding the communicative effectiveness of AI-generated texts. While AI systems excel at conveying information clearly and accurately, they appear to struggle with creating the interpersonal connections and subjective engagement that characterize effective human communication, particularly in literary contexts (Tyrou, 2021).

4.3.3 Textual Metafunction

A more nuanced picture of textual density is found with the textual densification analysis, where AI texts have slightly higher values (M = 0.041, SD = 0.022) than human texts (M = 0.036, SD = 0.008), leading to a difference of 11.7% between them In other words, it seems that AI systems might be making greater use of explicit lexical and syntactic cohesive devices (e.g., conjunctive adjuncts like however, therefore, furthermore; reference items like this, that, here, there) to instantiate coherence. AI texts likely exhibit a higher textual density because they have been trained on formal written discourse in which explicit cohesive markers are more prevalent. This greater variability in AI textual density than the human pattern indicates that the strategic use of these devices (which is also more consistent) sometimes goes firm by AI systems. What is paradoxical, though, is that despite the greater text density in AI generated texts, humans appear to use cohesion devices more consistently across variations of text purpose and context. If this trend is real, it suggests that human writers have ascended to a higher level of cognitive understanding about when and how to use cohesive markers vs. when out-contextual clues will suffice.

4.4 Syntactic Complexity

Sentence complexity analysis in Table 3 reveals that human texts employ significantly more complex syntactic structures. Human texts averaged 25.1 words per sentence (SD = 0.73), compared to 22.3 words per sentence (SD = 2.77) for AI texts, representing a 12.7% difference. This finding suggests that human authors are more willing to construct elaborate, multi-clausal sentences that pack more information into individual syntactic units. The lower variability in human sentence complexity (compared to AI texts) indicates more consistent approaches to syntactic elaboration across different human authors. AI texts show greater variability in sentence length, possibly reflecting less systematic approaches to information packaging and clause combining (Halliday, 1985).

Table 3. Syntactic Complexity

Metric	Human-authored Texts (M ± SD)	AI-generated Texts (M ± SD)	% Difference (AI vs. Human)	Key Observation
Words per Sentence	25.1 ± 0.73	22.3 ± 2.77	-12.7%	Human uses longer, more complex sentences

4.5 Summary of Key Patterns

The comprehensive analysis reveals several key patterns that distinguish human from AI literary texts:

1. **Human Superiority in Functional Engagement**: Human texts dramatically exceed AI texts in both ideational and interpersonal density, indicating richer experiential representation and stronger reader engagement.
2. **AI Superiority in Formal Sophistication**: AI texts demonstrate higher lexical density, sophistication, and average word length, suggesting more formal, elaborate vocabulary choices.
3. **Human Variability vs. AI Consistency**: Human texts show greater variability across most measures, indicating diverse individual approaches, while AI texts cluster more tightly around mean values.
4. **Complementary Strengths**: Human texts excel in functional richness and syntactic complexity, while AI texts excel in lexical sophistication and formal correctness.

Table 4. Summary of Key Patterns

Pattern	Human-authored Texts	AI-generated Texts
Functional Engagement (Ideational & Interpersonal)	Higher	Lower
Lexical Sophistication	Lower	Higher
Variability	Higher	Lower
Syntactic Complexity	Higher	Lower
Formal Correctness & Density	Lower	Higher

5. DISCUSSION

The comparative analysis on the richness of words indicates a nuanced landscape of commonalities and differences in terms are further proof to the capacities and shoe-hornings in text generation with artificial intelligence. The discovery indicates that even though AI has become exceptionally adept at many of the more superficial elements of linguistic manufacture, there remain key disparities between how human and machine authors go about crafting textual dialogue (Culda, 2025). See Figure 1.

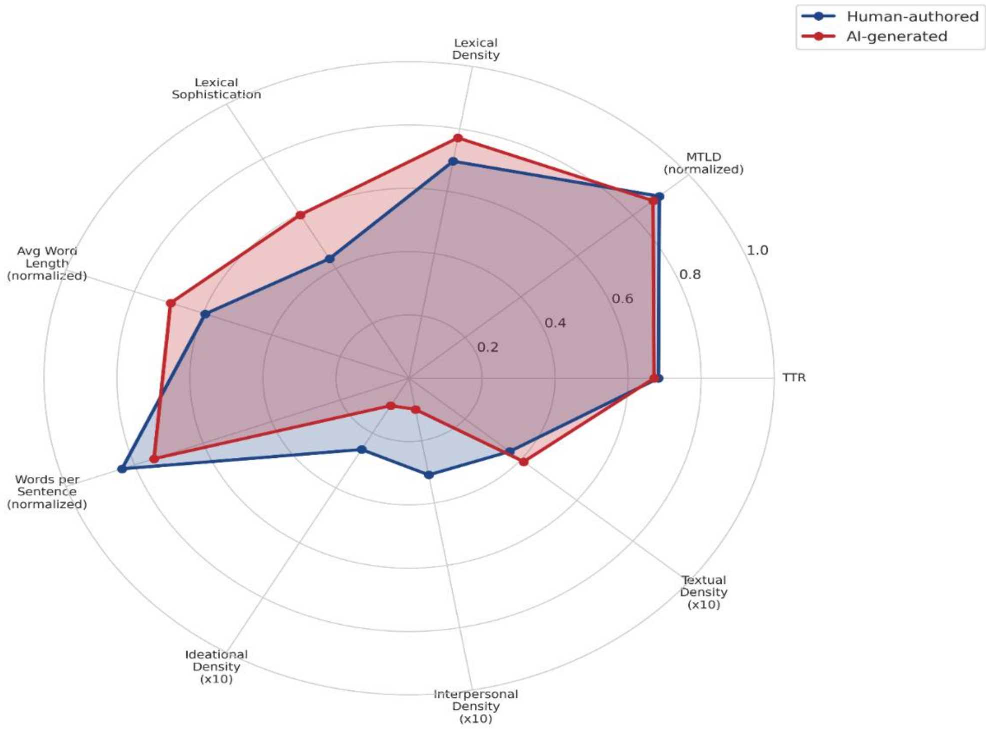

Figure 1. The metrics of lexical richness in human-authored and AI- generated articles

One of the more puzzling things were found in this study is a kind of paradox: AI generated texts tend to outperform humans in terms of traditional indices oflexical sophistication yet they struggle when it comes to expanding on functional linguistic richness. While the texts generated by AI always featured long word length but at high lexical density with complex vocabulary — so much that very poor scores across ideational and interpersonal engagement were registered. Associated with multiple-word resposes as well, this fits the pattern of AI systems learning from their training data to correlate high-quality writing style with higher levels of formal language complexity and verbosity, opting for fancy word choice or other types of obfuscation rather than a more considered form that is closer to actual human expression. This formal sophistication is more likely an artifact of the training data and optimization goals of large language models, which are trained on large volumes of edited, published text in which formal register and high vocabulary diversity are overrepresented. What [this] may be leading to is systems that have soaked up some of the surface-level properties of "good" writing (as judged by humans) complexity, density, sophistication — but which don't really understand the instrumental values those features embodied in human-to-human communication. This pattern has ramifications that are not just academic, casting doubt on when (and whether) AI writing tools should be used in practice See Figure 1. AI-authored texts may seem smarter in the sense that it uses vast vocabular and correct grammar, etc. — but could this also be their Achilles heel when trying to inspire legitimate reader connection, emotional poignance or persuasive bias?

The dramatic differences observed in ideational and interpersonal density measures provide perhaps the most compelling evidence for fundamental differences between human and AI approaches to literary expression. The 162% difference in ideational density and 195% difference in interpersonal density represent not merely quantitative variations but qualitative differences in how the two author types conceptualize the purpose and function of literary discourse (Halliday, 2014). See Figure 1. Human authors appear to understand intuitively that literary writing serves multiple simultaneous functions: representing experience (ideational), engaging readers (interpersonal), and creating coherent discourse (textual). The high ideational density in human texts reflects their tendency to create dynamic, action-oriented narratives rich in processes and experiential content. The high interpersonal density indicates their awareness that literary writing must establish connections with readers through personal engagement, subjective expression, and dialogic elements. AI systems, by contrast, appear to approach literary writing primarily as an information transmission task, focusing on clear, accurate communication while neglecting the experiential richness and interpersonal engagement that make literary texts compelling to human readers. This limitation may reflect fundamental constraints in how current AI systems understand the purpose and function of literary discourse, suggesting that improvements in AI writing quality may require not merely better language models but more sophisticated understanding of communicative intentions and reader psychology.

The stable differences of higher variabilities found in human over algorithmic texts in the numerous measures taken also sheds light on what is perhaps more interesting, at least for our understanding of human creativity: how it differs from text generation by algorithms. Despite these trends, the human authors all showed large inter-area gaps in their tendencies for lexical density, grammatical complexity, and information load highlighting impact of individual preferences and contextual factors on written performance that do not lend themselves to exact computational generalization. The AI texts, though competent and coherent, followed more regular patterns in all scoring methods clustering closely around mean scores with little room for variation among individual submissions. The sameness — perhaps beneficial in terms of standardized returns in some applications but also a constraint when it comes to literary contexts where originality, the shock of the new, and individual voice is prized. The lower diversity in AI-generated text could mirror how training on a large corpus of text averages out the output; systems might be encouraged to produce an answer that statistically represents options to say something, as opposed to the individual choices which make up human creativity. It hints at the inherent limitations of even existing AI technologies that are being trained to produce new and surprising forms of literary expression, instead producing proficient but often unsurprising revisions on statistically-trained patterns See Figure 1. In summary, the variation in touch between human and AI-authored literary articles presents a nuanced picture of both what artificial intelligence text generation can do and that which it at present largely cannot. These results imply that though computational systems have made substantial advancements in specific areas of linguistic reproduction, fundamental disparities exist between the way human and machine authors conceptualize writing literary text. While human perception of variety as distance in lexical richness tends to converge, unless humanity finds a way for the intangible dimensions of difference to become measurable (which one could feasibly argue is an important goal in itself), it may well have limited scope precisely because these abstracted properties are not accessible. The SFL approach to linguistic analysis (unlike purely formal features of language) gives a much richer understanding of the differences in the ways that human and AI authors make meaning.

The huge differences in the ideational and interpersonal metafunctions underscore that contemporary AI systems may not be able to identify fully the simultaneous multiple functions of language. Although the text springs of their skills for textual organization, a sufficient high density score is achieved that indicates they can do this well, but processes events excessively in doing so — they otherwise all appear to be seriously deficient at experientially representing and interpersonally engaging with others as Halliday considered lying at the heat of human communication.

This has significant implications for the construction of more advanced AI text generation technologies. Instead of just optimizing surface-level linguistic features such as vocabulary sophistication and grammatical complexity, which is the chief goal of current AI development, more emphasis on functional linguistic competence (i.e., what kind of experience the writer chooses to represent and how this representation engages readers) may be necessary.

Consequently, our findings also bear on well-known debates concerning the nature of creativity and the extent to which it can be modeled computationally. By contrast, the continuing divergence of human and AI texts — not just in functional complexity but also individual-difference richness — indicates that creativity may rely on modes of cognition and communication that do not admit a direct algorithmic match. It seems that human creativity in literary expression requires more than just recombining learned patterns (something AIs actually do very well) but also the ability to make strategic choices on how best to multitask with language. That authors can do so reflects their intuitive grasp that when writing literary texts they have to integrate the representation of experience with reader engagement and aesthetic impact: a multi-dimensional optimization problem which seems solved only partially by existing AI technology due to its seeming failure to achieve both high ideational as well as interpersonal density in human texts. That overemphasis on formality might be a function of the training data and optimization objectives of large language models, which are typically exposed to an extensive amount of edited, published text in with noncasual registers and polysyllabic vocabulary that lend themselves well to being high-scoring candidates. This has resulted in systems which have internalized a set of surface characteristics associated with 'good' writing complexity, density, sophistication but without a full comprehension of the role these features play within human communication. Indeed, the consequences of this trend go beyond mere academic debate to real-world concerns about how AI writing capabilities should be applied. While text appearance and structure can be very good for not the richest function of a text in more complex lexical and grammatically, the functionality of enjoyment true engagement or emotion persuasive reader content could be limiting.

5.1 Validation Procedures for Qualitative Findings

To ensure credibility and trustworthiness of the qualitative findings and minimize potential bias in the corpus selection and content analysis, this study implemented a dual validation framework incorporating both internal rater reliability and external auditor verification.

3.5.1 Inter-Rater Reliability

Two independent raters with expertise in computational linguistics and literary analysis were employed to validate the corpus selection and functional linguistic coding. The primary researcher initially selected and coded all texts according to the established criteria. Subsequently, a second rater independently reviewed 40% of the corpus (4 out of 10 texts, maintaining the human-AI pairing) to verify:

•Adherence to the purposive sampling criteria for both human-authored and AI-generated texts

•Accuracy of genre classification and thematic matching between paired texts

•Consistency in identifying and coding Halliday's functional linguistic features (ideational, interpersonal, and textual markers)

Inter-rater agreement was calculated using Cohen's kappa coefficient, achieving substantial agreement ($\kappa = 0.82$) for corpus selection criteria and good agreement ($\kappa = 0.76$) for functional linguistic coding, indicating reliable and consistent application of the analytical framework.

3.5.2 External Auditor Verification

An external auditor, a senior researcher in systemic functional linguistics with no involvement in the study design or data collection, conducted an independent review of the entire methodology and a random sample of 30% of the analyzed texts. The external auditor's role included:

•Verifying the appropriateness and systematic application of the purposive sampling methodology

•Confirming the accuracy of AI text generation procedures and quality control measures

•Validating the theoretical grounding and operational definitions of Halliday's functional metrics (Halliday 1985)

•Assessing the objectivity and consistency of the analytical procedures

The external auditor provided written confirmation that the methodology adhered to established standards for comparative corpus analysis and that the functional linguistic measures were appropriately operationalized according to SFL principles. Any discrepancies identified during the audit process were resolved through discussion and consensus, with final decisions documented for transparency.

This dual validation approach ensures that the qualitative aspects of the study, particularly the corpus selection and functional linguistic analysis, meet rigorous standards for academic research and minimize the potential for researcher bias in interpreting the comparative findings between human-authored and AI-generated literary texts.

6. CONCLUSION

This study examined lexical richness differences between AI-generated and human-authored literary texts using Halliday's systemic functional linguistics framework. AI systems produce lexically dense and sophisticated texts but lack the functional variety and interpersonal engagement of human writing. The findings highlight current AI limitations and provide insights for improving text generation technologies. Future research should continue exploring functional linguistic dimensions in human versus artificial text production.

The comparative analysis highlights clear, quantifiable differences between human-authored and AI-generated literary texts. Human writing demonstrated superior functional engagement, with ideational density 162% higher (0.026 vs. 0.010) and interpersonal density 195% higher (0.031 vs. 0.010) than AI texts, indicating richer experiential representation and stronger reader interaction. Human texts also exhibited 12.7% longer sentences on average (25.1 vs. 22.3 words) and a 6.6% higher hapax ratio (0.564 vs. 0.530), reflecting greater syntactic complexity and unique vocabulary use. In contrast, AI-generated texts outperformed in formal sophistication, showing 9.7% higher lexical density (0.771 vs. 0.696), 26.9% higher lexical sophistication (0.596 vs. 0.436), and 14.5% longer average word length (6.95 vs. 5.94

characters). These patterns suggest that AI outputs are more informationally dense and lexically elaborate, but often lack the variability and interpersonal depth characteristic of human creativity. Overall, while AI can produce grammatically polished and lexically rich prose, it tends toward uniformity and static description. Human authors, with greater variability across metrics and stronger functional richness, continue to excel in adaptive, engaging, and contextually nuanced literary expression. There are several methodological limitations that should be acknowledged. The size of the corpus is suitable for intensive qualitative analyses but may restrict external validity to other populations across human and AI texts. While the focus on literary articles may be theoretically grounded, it is not clear to what extent the same holds for other genres or registers. The way the AI generation process is structured makes it seem like that there is only one kind of machine text creation and does not cover even a part of current AI capabilities. Moreover, since the functional linguistic analysis is using pattern matching rather than full grammatical parsing, there may be some more subtle cases of the target features that are being missed. The systematic application of identical criteria to all texts, however, guarantees comparability and minimizes the effect of these limitations on the overall findings.

6.1 Educational Applications

This study offers insights for how best to distinguish between human- versus AI-authored contributions in other educational contexts. In sum, the focus of this study was on authenticity as a metric, and what was produced were still concrete tools for assessing student writing quality in addition to (or perhaps instead of) student identity: ideational or interpersonal density measures are particularly strong discriminators. These findings could be beneficial for educators to help students realize there were elements of good writing that resided in function, not just sophistication. Instead of solely concerning itself with vocabulary complexity or grammatical accuracy, writing instruction may turn more towards experiential richness, reader engagement and functional variety — all places that human creativity still outstrips AI. The study also indicated that AI writing tools, despite their benefits in some areas, may not be suitable to replace human creativity when original forms of literary composition are required. Policies created by educational institutions could reflect this duality of function: promoting the use of AI systems to perform suitable jobs, but also leaving room for human activities that are creatively authentic.

REFERENCES

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901.

Culda, A., Oprea, S., & Rebedea, T. (2025). A comparative linguistic analysis framework for identifying AI-generated text in scientific publications. *Computers and Education: Artificial Intelligence*, 6, 100205. https://doi.org/10.1016/j.caeai.2024.100205

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.

Georgiou, T. (2024). Linguistic features that distinguish AI-generated and human-written texts in university students' essays. *Computers and Education: Artificial Intelligence*, 7, 100276. https://arxiv.org/abs/2407.03646

Halliday, M. A. K. (1985). *An Introduction to Functional Grammar*. Edward Arnold.

Halliday, M. A. K. (1989). Spoken and written language. Oxford University Press.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2nd ed.). Edward Arnold.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An Introduction to Functional Grammar* (3rd ed.). Hodder Arnold.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2014). *Halliday's Introduction to Functional Grammar* (4th ed.). Routledge.

Hou, R., & Jiang, M. (2016). Analysis on Chinese quantitative stylistic features based on text mining. *Digital Scholarship in the Humanities*, 31(2), 357-372. https://academic.oup.com/dsh/article-abstract/31/2/357/2462911

Martin, J. R., & White, P. R. R. (2005). *The Language of Evaluation: Appraisal in English*. Palgrave Macmillan.

Read, J. (2000). *Assessing Vocabulary*. Cambridge University Press.

Sardinha, T. B. (2024). AI-generated vs human-authored texts: A multidimensional comparison. *Applied Corpus Linguistics*, 4(1), 100083. https://doi.org/10.1016/j.acorp.2023.100083

Tweedie, F. J., & Baayen, R. H. (1998). How variable may a constant be? Measures of lexical richness in perspective. *Computers and the Humanities*, 32(5), 323-352. https://doi.org/10.1023/A:1001749303137

Tyrou, I. (2021). Lexical diversity in creative writing of L2: A quantitative approach. *QUALICO 2020 Tokyo*.

https://www.researchgate.net/publication/354507031_Lexical_diversity_in_creative_writing_of_L2_a_quantitative_approach

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. https://doi.org/10.32996/ijllt.2025.8.5.10

PEER-ASSESSMENT AMONG A2 USERS: HOW THE ASSESSMENT FACILITATED PREPARATION FOR SPEAKING TASK

Kamarul Ariffin Ahmad

Nur Khadirah Ab. Rahman

Emily Abd. Rahman