ISELT 204
Interdisciplinary Studies in English Language Teaching

Pages 204 to 222

Lexico-Syntactic Complexity in Grok 4-Generated and Human-Written Essays: Implications for Writing

Farzad Mahmoudi Largani [1], Hooshang Khoshsima [1]*
[1] Chabahar Maritime University, Chabahar, Iran

10.22080/iselt.2026.32081.1201

Received
June 10, 2026
Accepted
July 16, 2026
Available online
July 21, 2026

Keywords:
Artificial Intelligence, Grok 4, Lexical Complexity, Syntactic Complexity, Essays.

Abstract

An increasing number of studies examining Artificial Intelligence (AI)-generated writing have produced conflicting findings, which is largely due to simplistic complexity measures and an overreliance on a single AI model. This study investigated differences in lexico-syntactic complexity between AI-generated and human-written argumentative essays under controlled conditions. Seventy International English Language Testing System (IELTS) Writing Task 2 essays were analyzed, comprising 35 written by Iranian English-as-a-foreign-language (EFL) learners and 35 generated by Grok 4, all responding to identical prompts targeting IELTS Band 8. Lexical and syntactic complexity were measured using Lu's (2010) framework through the Lexical Complexity Analyzer (LCA) and the L2 Syntactic Complexity Analyzer (L2SCA). Independent-samples t-tests with Bonferroni adjustments revealed that Grok 4 essays demonstrated significantly higher lexical diversity, sentence-level complexity, and coordination-based syntactic complexity, whereas human essays exhibited greater lexical variation across parts of speech. No significant differences emerged in length-based, dependent-clause, or phrasal complexity. These findings show AI's structural elaboration and human writers' lexical flexibility, highlighting the pedagogical value of AI-generated essays as contrastive models for syntactic revision and lexical refinement in argumentative writing.

1. INTRODUCTION

The quick rise and improvement of Artificial Intelligence (AI) models, with their prominent potential to generate texts resembling human-authored writings, has considerably changed the field of writing studies (Dergaa et al., 2023; Godwin-Jones, 2022). Accordingly, AI-generated writing has become a center of attention and a growing body of comparative studies attempted to examine the role of AI in the writing process, including its supportive role in feedback provision and revision practices (Ahmad et al., 2025; Bahari, 2025; Bozorgian & Rahimi, 2025; Wang, 2023), its impact on discourse-level features such as cohesion and coherence (Ismail, 2023),

* **Corresponding Author:** Hooshang Khoshsima, Ph.D., English Department, Chabahar Maritime University, Chabahar, Iran, **Email**: khoshsima2002@yahoo.com

benchmarking the linguistic features of AI-generated essays (AlAfnan & MohdZuki, 2023), and cross-rhetorical-stage variation in writing (Liu & Liu, 2025). In academic writing courses, AI tools are especially valued for improving surface-level accuracy such as grammar and vocabulary range, which are crucial for non-native English learners who intend to meet the academic requirements (Rehman & Farooq, 2024). Among the various academic genres, argumentative writing has received particular attention because of its central role in academic success and professional communication (Kuhn & Moore, 2015). Nevertheless, producing effective argumentative texts remains a demanding task for many English-as-a-foreign-language (EFL) learners (Sulistyo & Heriyawati, 2017), making this genre an important context for examining the linguistic characteristics of AI-generated and human-authored writing.

Within this context, linguistic complexity has emerged as a major focus of comparative AI writing research because it captures multiple dimensions of linguistic sophistication and is widely used to evaluate genre effects (Bulté & Housen, 2012). As a multidimensional construct, linguistic complexity encompasses both syntactic and lexical complexity (Ortega, 2012). Bulté and Housen (2018) maintained that syntactic complexity, demonstrated by the use of diverse and hierarchically organized sentences, serves as an accurate and valid indicator of writing proficiency and quality. Also, lexical complexity has been associated with greater rhetorical accuracy and better writing (Csomay & Prades, 2018; Fredrick & Craven, 2025). Accordingly, examining linguistic complexity in argumentative essays provides valuable insights into learners' writing proficiency and the rhetorical effectiveness of AI-assisted writing practices (Zindela, 2023). Several recent comparative studies have examined lexical and syntactic complexity in essays generated by AI and authored by humans (Fredrick & Craven, 2025; Georgiou, 2025; Hadji Jamel, 2025; Herbold et al., 2023; Nkhobo & Chaka, 2023; Verma & Arora, 2025; Wu, 2025; Zindela, 2023).

Despite the wealth of comparative studies in the field, there are gaps in the operationalization of syntactic and lexical complexity, inconsistencies in results, and the use of AI models, all of which necessitate further scrutiny of the topic. In particular, several studies measured syntactic complexity through specific metrics, such as mean sentence length (Zindela, 2023) and measures of subordination (Fredrick & Craven, 2025; Nkhobo & Chaka, 2023). Considering lexical complexity, studies utilized narrowly integrated metrics, such as type-token ratio, lexical diversity, vocabulary density, or lexical sophistication (Fredrick & Craven, 2025; Wu, 2025; Zindela, 2023). Lu (2010) suggested that syntactic and lexical complexity are multidimensional; therefore, operationalizing these measures by incorporating broader indices would yield deeper understanding of the distinctive features of AI-generated essays. In terms of the contradictory nature of the findings, some research indicated that AI-generated essays demonstrated greater syntactic complexity (Fredrick & Craven, 2025; Herbold et al., 2023; Nkhobo & Chaka, 2023), while others found that human-authored essays exhibited greater structural variety, syntactic sophistication, and clausal embedding (Verma & Arora, 2025; Wu, 2025; Zindela, 2023). Some showed no significant disparity between AI-generated and human-authored essays (Nkhobo & Chaka, 2023; Wu, 2025). Moreover, the research generally utilized various versions of ChatGPT for essay generation, with little attention paid to alternative AI models. Only one study has integrated DeepSeek (Sun & Li, 2025). The growing number of AI models, coupled with their unique stylistic characteristics derived from training data and architectural design (Fedoriv et al., 2023; Jaashan & Bin-Hady, 2025), underscores the need to examine further AI models.

To address the identified gaps and enhance the knowledge of lexical and syntactic complexity in AI-generated essays, this study used Grok 4, a recently released AI model that has achieved widespread popularity among users for content generation (xAI, 2025), and adopted Lu's (2010) multidimensional model to operationalize the two measures. Accordingly, two research questions were formulated:

1. Are there statistically significant differences in lexical complexity between human-authored and Grok 4-generated argumentative essays?
2. Are there statistically significant differences between human-authored and Grok 4-generated argumentative essays concerning syntactic complexity?

2. LITERATURE REVIEW

Grok as an Emerging AI Model

Along with the rapid change in AI-related technologies, the launch of Grok 4 by xAI demonstrated outstanding progress in the developing relationship between AI and higher education (xAI, 2025). Grok 4, which is a highly advanced AI model, represents a new generation of generative models designed to help users in research, problem-solving, and knowledge generation within different academic contexts (Al-Zubaidi, 2025). Reports indicated that Grok 4 can analyze and reason through complex academic and scientific texts conventionally managed by researchers while also presenting quicker and more structured answers (Smythos, 2025; xAI, 2025). Grok, developed as a major competing platform to ChatGPT, highlights a different design objective. Unlike ChatGPT, which considers flexibility across educational, professional, and everyday uses, Grok concentrates on the broader goals of enhancing scientific discovery and deepening understanding of intelligence (Al-Zubaidi, 2025). In particular, Grok differs from well-known systems such as ChatGPT by emphasizing real-time data processing, interactive communication, and advanced mathematical reasoning, underlining its capability as a strong tool for dynamic, data-driven academic and educational use (Wangsa et al., 2024). Despite the continued importance of ChatGPT, the advent of novel models such as Grok has drawn scholarly attention to investigate their pedagogical advantages, limitations, and educational impacts, especially given that research on Grok is still in its early stages (Mohammed et al., 2025).

Dimensions of Syntactic Complexity

Syntactic complexity is widely regarded as a core concept in second language (L2) writing research and an important indicator of writing quality, linguistic proficiency, and language development (Bulté & Housen, 2014; Wolfe-Quintero et al., 1998). It captures the range, variety, and sophistication of the grammatical patterns writers employ to express ideas (Lu, 2010; Ortega, 2003). Despite extensive investigation, no consensus has been reached regarding its operationalization, leading researchers to increasingly adopt multidimensional frameworks that account for its diverse structural dimensions (Lu, 2011).

Early approaches to measuring syntactic complexity relied primarily on global or large-grained indices, including mean length of sentence (MLS), mean length of T-unit (MLT), and mean length of clause (MLC) (Hunt, 1966). These length-based measures provide broad estimates of structural sophistication by assuming that longer production units reflect greater syntactic elaboration. Although they have been shown to correlate with overall proficiency (Wolfe-Quintero et al., 1998), they often overlook qualitative differences in how complexity is realized across linguistic structures (Bulté & Housen, 2014).

Advances in corpus linguistics and natural language processing have shifted attention toward finer-grained indices that capture structural variation at clausal and phrasal levels (Kyle, 2016; Lu, 2011). This multidimensional perspective differentiates complexity achieved through subordination, coordination, and phrasal elaboration (Biber et al., 2011). Measures such as coordinate phrases per clause (CP/C), dependent clauses per T-unit (DC/T), and complex nominals

per T-unit (CN/T) provide more detailed insights into the structural sophistication of L2 writing (Lu, 2010; Staples & Reppen, 2016). Accordingly, global measures offer a broad view of syntactic development, whereas finer-grained indices reveal the specific grammatical mechanisms underlying syntactic sophistication.

The Multidimensional Nature of Lexical Complexity

According to Lu (2012), lexical complexity represents a writer's ability to convey ideas efficiently in written language. It is a multidimensional construct comprising three principal dimensions: lexical diversity, lexical sophistication, and lexical density (Lu, 2012). Lexical diversity (also referred to as lexical variation) concerns the range and variety of words employed in a text (Johansson, 2008). Lexical diversity indices are widely used to assess vocabulary range and are regarded as indicators of a writer's depth of lexical knowledge (Kyle, 2020).

Lexical sophistication reflects the complexity of the vocabulary employed and is generally measured by the proportion of less frequent or more advanced words in a text (Read, 2000). Various approaches have been adopted to assess lexical sophistication, including frequency-based word lists (Laufer & Nation, 1995), indices such as the Advanced Guiraud (AG) index (Daller et al., 2003), and computational tools including Tool for the Automatic Analysis of Lexical Sophistication (TAALES) and the Lexical Complexity Analyzer (LCA) (Lu, 2010). Previous research has consistently shown that greater use of low-frequency vocabulary is associated with more advanced L2 writing development and higher lexical sophistication (Laufer, 1994).

Lexical density, another key dimension, is commonly defined as the proportion of content words to the total number of words in a text (Gregori-Signes & Clavel-Arroitia, 2015; Lu et al., 2019). It is widely regarded as an indicator of informational density and text complexity, with denser texts typically conveying more information per unit of language (Johansson, 2008). Halliday and Matthiessen (2014) further argued that lexical density reflects the extent to which lexical meaning is condensed within grammatical structures.

Although lexical diversity, sophistication, and density are conceptually distinct, they are closely interconnected and jointly contribute to overall lexical complexity (Read, 2000). For instance, the use of more sophisticated vocabulary may enhance lexical diversity, while lexical density is influenced by both the range and complexity of lexical choices (Laufer & Nation, 1995; Lu, 2012).

Empirical Studies

Recent studies have increasingly compared AI-generated and human-authored writing across lexical, syntactic, discourse, and pragmatic dimensions, yielding convergent evidence of both the strengths and limitations of generative AI. Early large-scale comparisons by Herbold et al. (2023) showed that ChatGPT-generated argumentative essays, particularly those produced by ChatGPT-4, outperformed human essays in overall writing quality, logical organization, lexical diversity, and syntactic complexity. Nevertheless, human writers employed more epistemic and discourse markers, suggesting greater rhetorical engagement despite lower lexical sophistication. Likewise, Zindela (2023) reported that AI-generated argumentative essays exhibited higher lexical sophistication, whereas human essays demonstrated greater syntactic complexity and richer structural variation. In contrast, Nkhobo and Chaka (2023) found no statistically significant differences in lexical and syntactic complexity, indicating substantial linguistic similarity between AI- and human-generated writing.

Subsequent studies broadened these comparisons by incorporating additional linguistic dimensions. Fredrick and Craven (2025) demonstrated that ChatGPT-generated essays exceeded

human writing in lexical diversity, lexical sophistication, and subordination, but exhibited lower readability because of their greater linguistic complexity. Similarly, Georgiou (2025) found that AI-generated essays relied more heavily on content words, adjectival modifiers, and nouns, whereas human essays employed more function words, verbs, auxiliaries, and prepositional structures, resulting in greater clarity and naturalness. Correspondingly, Etaat (2026) reported that AI-generated essays displayed superior lexical richness, syntactic complexity, semantic organization, and linguistic accuracy, while learner essays remained more readable despite their comparatively simpler language.

Several studies have emphasized discourse and rhetorical characteristics rather than purely structural measures. Hadji Jamel (2025) found that AI-generated persuasive essays exhibited greater sentence complexity, keyword density, and nominalization, whereas human writers demonstrated higher lexical diversity and a broader repertoire of discourse and epistemic markers, contributing to more natural communication. Likewise, Wu (2025) reported only marginal lexical advantages for AI, but showed that human writers consistently produced stronger cohesion, more natural idea development, and greater use of embedded clauses. From a phraseological perspective, Alqurashi and Hamed (2026) demonstrated that AI-generated texts exhibited sophisticated phraseological patterns and immediate fluency; however, human writing reflected gradual developmental growth, contextual adaptation, and authentic cognitive effort.

Beyond traditional complexity measures, recent research has also examined broader pedagogical issues. Verma and Arora (2025) observed that AI-generated literary essays relied on generic vocabulary and predictable sentence structures, whereas human writing displayed greater contextual grounding, originality, and pragmatic engagement. Furthermore, Bozorgian and Rahimi (2025) showed that ChatGPT-4o feedback emphasized linguistic accuracy and organization, while peer feedback promoted interpersonal interaction and self-regulation.

Across the reviewed studies, lexical complexity was measured using a range of indices, including type-token ratio, measure of textual lexical diversity (MTLD), lexical sophistication, vocabulary density, and low-frequency word use, whereas syntactic complexity was operationalized through measures such as mean length of sentence or T-unit, subordination, modifiers per noun phrase, and multidimensional frameworks such as Lu's (2010) model. Despite these methodological advances, the findings remain inconsistent. Some studies reported greater lexical diversity and sophistication in AI-generated essays, whereas others found human-written texts to be lexically richer. Similar inconsistencies were observed in syntactic complexity, with AI-generated texts often exhibiting greater surface-level complexity, while human-authored essays demonstrated richer structural variety, embedding, and discourse-oriented syntactic patterns. Moreover, previous research has predominantly focused on ChatGPT, with comparatively little attention devoted to newly developed AI models such as Grok 4. To address these gaps, the present study compares Grok 4-generated and human-authored International English Language Testing System (IELTS) argumentative essays using Lu's (2010) multidimensional framework of lexical and syntactic complexity under tightly controlled conditions of genre, topic, proficiency, and text length. By controlling these potential sources of variation, the study aims to provide a more comprehensive and reliable account of the lexico-syntactic characteristics of contemporary AI-generated writing.

3. METHODOLOGY

For this study, 35 human-authored and 35 Grok 4-generated argumentative essays were selected as the corpus of the study. As for the human-authored corpus, the argumentative essays were collected from an institute that offers IELTS preparation courses. The essays were written by

Iranian EFL learners for Writing Task 2 of the IELTS examination and were rated by professional and experienced IELTS examiners. To maintain consistency in the proficiency level, only essays with a band score of 8 were selected. In Writing Task 2, test-takers are required to write at least 250 words on a specific topic. In this study, the selected argumentative essays focused on opinion (agree/disagree) topic. To control for topic effects, all essays were elicited using a single IELTS Writing Task 2 argumentative prompt requiring participants to express and justify their agreement or disagreement with the proposition that studying art should be compulsory in schools because of its purported benefits for academic performance.

For the AI-generated corpus, Grok 4 was used. Grok 4 is a prompt-based AI model that can generate human-like texts based on the prompts it receives. The prompt used by the researchers defined the exact topic of the essay written by students, the required word count, and the proficiency level that the output had to reflect. Therefore, for each Grok 4-generated essay, this prompt was used: "*Write a 250-word argumentative essay on [topic]. The essay should demonstrate a proficiency level equivalent to an IELTS band score of 8*". As AI outputs might be influenced by previous conversation, several steps were followed to avoid this. Accordingly, each essay was generated in a single chat conversation. No regenerated, revised, or edited outputs were used, and all generated texts were retained in the corpus. Using the IELTS scoring rubric, a qualified IELTS examiner rated the essays generated by Grok 4 to confirm that they were comparable to the human-authored essays in terms of the proficiency level. After a month interval, the essays were rated again by the same examiner. Intra-rater reliability was calculated using the Pearson correlation coefficient and a desirable reliability coefficient of 0.81 was observed.

Regarding the corpus representativeness, the corpus was designed according to established corpus design principles which emphasize that representativeness depends on principled sampling and a clearly defined target population rather than corpus size alone (Biber, 1993). All texts belonged to the same genre, responded to a single argumentative prompt, were written in English under comparable conditions, and had similar lengths, thereby enhancing the comparability of the two sub-corpora while minimizing sampling bias. As McEnery and Hardie (2011) argued, specialized corpora should be evaluated based on their alignment with the research objectives and target discourse domain. Table 1 summarizes the information about each corpus.

Table 1. Information about the Corpus of the Study

	Human-Authored Corpora		Grok 4-Generated Corpora	
	N of essays	**N of words**	**N of essays**	**N of words**
Essays	35	8771	35	8702

Tools and Measures

Lexical and Syntactic Complexity Analyzer

Lexical and syntactic complexity of the corpus was measured using the LCA and the L2 Syntactic Complexity Analyzer (L2SCA), both developed by Lu (2010). These analyzers are computational tools designed to automatically evaluate texts based on particular indices that represent various complexity features. The web-based L2SCA is one of the widely used tools in this field (Spring & Johnson, 2022). The tool offers a Single Mode, allowing users to analyze a single text or compare two texts based on selected lexical complexity measures, completed with options for graphical representation of the results. Additionally, it offers a Batch Mode, which enables the simultaneous

analysis of up to 200 files, producing results in a Comma-Separated Values (CSV) format appropriate for statistical analysis.

Measures of Lexical Complexity

To operationalize the lexical complexity, Lu's (2010) measures were used. It employs 25 different measures that assess aspects such as lexical variation, density, and sophistication. The detailed information about the measures is provided in Table 2. The instrument allows the researcher to assess the lexical richness of texts through the comparison of the input files inserted by the researcher with predetermined lists of words (Lu, 2012; Spring & Johnson, 2022).

Table 2. Measures of Lexical Complexity Used in the LCA (Lu, 2010)

Measure	Definition
1. Lexical density (LD)	lexical words/N of words
2.Lexical sophistication	
2.1. Lexical Sophistication 1 (LS1)	sophisticated lexical words/N of lexical words
2.2. Lexical Sophistication 2 (LS2)	Sophisticated word types / N of word types
2.3. Verb Sophistication 1 (VS1)	N of sophisticated verb types / N of verbs
2.4. Verb Sophistication 1 (VS2)	variations (corrections) of VS1 measure
2.5. Corrected Verb Sophistication 1 (CVS1)	
3.Lexical variation	
3.1. NDW	
3.1.1. Number of Different Words (NDW)	N of different words used in a language sample
3.1.2. Number of Different Words in the First 50 Words (NDW-50)	
3.1.3. Expected Random NDW for 50 Words (NDW-ER50)	
3.1.4. Expected Sequential NDW for 50 Words (NDW-ES50)	
3.2. TTR	
3.2.1. Type–Token Ratio (TTR)	
3.2.2. Mean Segmental Type–Token Ratio (50-word segments) (MSTTR-50)	N of word types /N words in a text
3.2.3. Corrected Type–Token Ratio (CTTR)	
3.2.4. Root Type–Token Ratio (RTTR)	
3.2.5. Logarithmic Type–Token Ratio (LogTTR)	
3.2.6. Uber Index (Uber)	
3.3. Verb diversity	
3.3.1. Verb Variation 1 (VV1)	
3.3.2. Squared Verb Variation 1 (SVV1)	
3.3.3. Corrected Verb Variation 1 (CVV1)	
3.4. Lexical word diversity	
3.4.1. Lexical Variation (LV)	variation of specific classes of words
3.4.2. Verb Variation 2 (VV2)	
3.4.3. Noun Variation (NV)	
3.4.4. Adjective Variation (AdjV)	
3.4.5. Adverb Variation (AdvV)	
3.4.6. Modifier Variation (ModV)	

Measures of Syntactic Complexity

To analyze syntactic complexity, the study adopted Lu's (2010) framework, which categorizes various indices of syntactic structures. The reason for choosing this classification was that, unlike the simpler measures of syntactic complexity, which focus on narrow aspects of syntax, this multidimensional classification allows for a more sophisticated analysis of writing quality (Bulté & Housen, 2018; Lu, 2010). Additionally, studies indicated that Lu's classification is a reliable and valid measure to assess syntactic complexity across different genres(Housen et al., 2019; Lahuerta Martínez, 2018). The indices of Lu's (2010) syntactic complexity are presented in Table 3.

Table 3. Measures of Syntactic Complexity (Lu, 2010)

Measure	Definition
Length of Production Unit	
Mean Length of Clause (MLC)	N of words/ N of clauses
Mean Length of Sentence (MLS)	N of words/ N of sentences
Mean Length of T-unit (MLT)	N of words/ N of T-units
Overall Sentence Complexity	
Clauses per Sentence (C/S)	N of clauses/ N of sentences
Subordination	
Clauses per T-unit (C/T)	N of clauses/ N of T-units
Complex T-units per T-unit (CT/T)	N of complex T-units/ N of T-units
Dependent Clauses per Clause (DC/C)	N of dependent clauses/ N of clauses
Dependent Clauses per T-unit (DC/T)	N of dependent clauses/ N of T-units
Coordination	
Coordinate Phrases per Clause (CP/C)	N of coordinate phrases/ N of clauses
Coordinate Phrases per T-unit (CP/T)	N of coordinate phrases/ N of T-units
T-units per Sentence (T/S)	N of T-units/ N of sentences
Phrasal Sophistication	
Complex Nominals per Clause (CN/C)	N of complex nominals/ N of clauses
Complex Nominals per T-unit (CN/T)	N of complex nominals/ N of T-units
Verb Phrases per T-unit (VP/T)	N of verb phrases/N of T-units

Procedure and Data Analysis

The Statistical Package for the Social Sciences (SPSS), version 27, was used to analyze the data. Before analyzing the corpus, the human-written and Grok-generated essays were assessed by an IELTS examiner to ensure that they were at the same level of proficiency. The examiner was not informed which essays were human-written and which were Grok-generated. After a one-month interval, the examiner reassessed the essays, and intra-rater reliability was calculated. The result showed an acceptable reliability value of 0.82. To analyze the lexico-syntactic complexity, the corpus was assessed using individual indices. Shapiro-Wilks Test of normality was used to ensure that the data were normally distributed. To answer the research questions of the study, a set of independent samples t-tests were used to compare the scores. As multiple independent samples t-tests were performed to analyze the data, Bonferroni Adjusted Alpha was used to control Type I errors.

4. RESULTS

In this section, the results of the analysis for each research question are presented. Prior to analyzing the data for each research question, Shapiro-Wilk goodness-of-fit test was used to check the assumption of normal distribution. Tables 4 and 5 show the results.

Table 4. Tests of Normality for Lexical Complexity Indices

	Shapiro-Wilk		
	Statistic	**df**	**Sig.**
LD	.983	70	.450
LS1	.976	70	.191
LS2	.971	70	.104
VS1	.976	70	.200
VS2	.980	70	.322
CVS1	.984	70	.519
NDW	.978	70	.248
NDW50	.960	70	.124
NDWER50	.989	70	.822
NDWES50	.936	70	.081
TTR	.969	70	.080
MSTTR	.987	70	.710
CTTR	.962	70	.133
RTTR	.988	70	.721
LogTTR	.970	70	.091
Uber	.984	70	.527
VV1	.985	70	.569
SVV1	.975	70	.174
CVV1	.980	70	.326
LV	.943	70	.053
VV2	.935	70	.211
NV	.978	70	.264
AdjV	.948	70	.406
AdvV	.960	70	.325
ModV	.976	70	.189

The results of Shapiro-Wilk goodness-of-fit test for lexical complexity indices in Table 4 revealed that the data were normally distributed as the observed significant values were greater than the critical value ($p > .05$), suggesting that the assumption of normality was met.

Table 5. Tests of Normality for Syntactic Complexity Indices

	Shapiro-Wilk		
	Statistic	**df**	**Sig.**
MLS	.976	70	.189
MLT	.984	70	.490
MLC	.986	70	.611
C/S	.971	70	.107
C/T	.971	70	.107
CT/T	.972	70	.120
DC/C	.993	70	.972
DC/T	.973	70	.137
T/S	.917	70	.061
CP/T	.969	70	.083
CP/C	.979	70	.273
CN/T	.967	70	.061
CN/C	.979	70	.271
VP/T	.985	70	.542

According to Table 5, the results of Shapiro-Wilk goodness-of-fit test demonstrated that the assumption of normality was met as the significant value for each index is greater than the critical value ($p > .05$).

RQ1: Human vs. Grok Essays: Lexical Complexity Compared

The first research question sought to investigate the differences between human-authored and Grok 4-generated argumentative essays in terms of lexical complexity. The results of independent samples t-tests are presented in Table 6.

Table 6. Results of Independent Samples T-Test for Lexical Complexity

	Grok 4		Human		t	sig.
	M	SD	M	SD		
LD	0.653	0.035	0.677	0.036	-2.821	.006
LS1	0.441	0.039	0.440	0.044	.151	.880
LS2	0.368	0.079	0.343	0.077	1.329	.188
VS1	0.322	0.040	0.327	0.034	-.535	.594
VS2	3.185	0.750	3.648	0.659	-2.737	.008
CVS1	1.526	0.247	1.428	0.250	1.645	.105
NDW	157.829	7.205	158.143	6.293	-.194	.846
NDW50	43.171	1.654	42.686	2.026	1.099	.276
NDWER50	42.706	0.971	41.797	0.925	4.007	.000*
NDWES50	44.277	0.784	42.343	1.277	7.638	.000*
TTR	0.611	0.036	0.535	0.026	10.099	.000*
MSTTR	0.899	0.047	0.894	0.037	.533	.596
CTTR	6.768	0.518	6.773	0.398	-.048	.962
RTTR	9.699	0.524	9.740	0.544	-.324	.747
LogTTR	0.923	0.037	0.911	0.037	1.306	.196
Uber	28.240	1.713	28.251	1.483	-.029	.977
VV1	0.798	0.054	0.791	0.047	.572	.569
SVV1	27.323	2.764	26.452	2.669	1.341	.184
CVV1	3.507	0.234	3.516	0.238	-.151	.881
LV	0.662	0.067	0.703	0.023	-3.465	.001*
VV2	0.203	0.024	0.237	0.049	-3.621	.001*
NV	0.587	0.054	0.693	0.039	-9.404	.000*
AdjV	0.156	0.015	0.181	0.026	-5.046	.000*
AdvV	0.062	0.066	0.055	0.052	.487	.628
ModV	0.194	0.036	0.220	0.020	-3.832	.000*
df = 68						

*** Significant at Bonferroni adjusted alpha = .002**

According to the results in Table 6, Grok 4 showed a significantly higher mean for NDWER50 *(M* = 42.70, *SD* = 0.97) than human authors *(M* = 41.79, *SD* = .92), *t*(68) = 4.00, *p* < .002. The results for NDWES50 also indicated a significantly higher mean for Grok 4 *(M* = 44.27, *SD* = 0.78) than human authors *(M* = 42.34, *SD* = 1.27), *t*(68) = 7.63, *p* < .002. Additionally, the TTR was significantly higher for Grok 4 (*M* = .61, *SD* = 0.03) than for human authors *(M* = 0.53, *SD* = .02), *t*(68) = 10.09, *p* < .002.

Conversely, human-authored essays displayed greater LV (*M* = 0.70, *SD* = 0.02) than Grok-4 essays (*M* = 0.66, *SD* = 0.06), *t*(68) = -3.46, *p* < .002. Regarding VV2, human-authored essays indicated a significantly higher mean (*M* = 0.23, *SD* = 0.04) than Grok-4 texts (*M* = 0.20, *SD* = 0.02), *t*(68) = -3.62, *p* < .00. In addition, the results showed significant differences in NV (human: *M* = 0.69, *SD* = 0.03 vs. Grok 4: *M* = 0.58, *SD* = 0.05; *t*(68) = -9.40, *p* < .002), AdjV

(human: *M* = 0.18, *SD* = 0.02 vs. Grok 4: *M* = 0.15, *SD* = 0.01; *t*(68) = -5.04, *p* < .002), and ModV (human: *M* = 0.22, *SD* = 0.02 vs. Grok 4: *M* = 0.19, *SD* = 0.03; *t*(68) = -3.83, *p* < .002), in favor of human authors. No other indices were significant after using the Bonferroni adjusted alpha level (*p* > .002).

RQ2: Human vs. Grok Essays: Syntactic Complexity Comparison

The second research question of the study examined the differences between Grok 4-generated and human-authored argumentative essays in terms of syntactic complexity and the results are presented in Table 7.

Table 7. Between-Group Comparison of Syntactic Complexity

	Grok 4		Human		t	Sig.
	M	SD	M	SD		
MLC	10.241	0.937	10.179	0.913	.280	.780
MLS	18.901	1.485	18.979	1.271	-.234	.816
MLT	18.217	1.487	17.511	1.509	1.971	.053
C/S	1.931	0.096	1.811	0.178	3.486	.001*
C/T	1.841	0.133	1.849	0.122	-.272	.787
CT/T	0.618	0.044	0.576	0.037	4.336	.000*
DC/C	0.427	0.032	0.436	0.038	-1.080	.284
DC/T	0.477	0.084	0.461	0.068	.875	.385
CP/C	0.498	0.072	0.433	0.066	3.912	.000*
CP/T	0.665	0.153	0.538	0.160	3.397	.001*
T/S	1.084	0.103	0.993	0.030	5.050	.000*
CN/C	1.761	0.134	1.705	0.191	1.417	.162
CN/T	2.750	0.263	2.824	0.219	-1.280	.205
VP/T	2.523	0.341	2.602	0.342	-.958	.342

df = 68

*** Significant at Bonferroni adjusted alpha = .003**

As the results in Table 7 show, Grok 4-generated and human-authored argumentative essays do not differ significantly (*p* > .003) on length-based syntactic measures, including MLT, MLC, and MLS. Regarding the overall sentence complexity, measured by C/S, the Grok 4-generated essays indicated a significantly higher mean (*M* = 1.93, *SD* = 0.09) than the human-authored essays (*M* = 1.81, *SD* = 0.17), *t*(68) = 3.48, *p* < .003. Considering subordination, only the complex T-unit ratio significantly differentiated the two corpora, while other indices were not statistically significant. Accordingly, Grok 4 essays had a higher CT/T mean (*M* = 0.61, *SD* = 0.04) compared to human texts (*M* = 0.57, *SD* = 0.03), *t*(68) = 4.33, *p* < .003. Grok 4-generated essays also included greater coordination. For CP/C, Grok 4 essays indicated a higher mean (*M* = 0.49, *SD* = 0.07) compared to human texts (*M* = 0.43, *SD* = 0.06), *t*(68) = 3.91, *p* < .003. Also, for CP/T, Grok 4 demonstrated a significantly higher mean (*M* = 0.66, *SD* = 0.15) than human-authored essays (*M* = 0.53, *SD* = 0.16), *t*(68) = 3.39, *p* < .003. The other differentiating measure was T/S, with Grok 4 essays exhibiting a significantly higher mean (*M* = 1.08, *SD* = 0.10) compared to human essays (*M* =

0.99, *SD* = 0.03), *t*(68) = 5.05, *p* < .003. With respect to the measures of phrasal sophistication (CN/C, CN/T, and VP/T), no significant differences were observed between the two corpora.

5. DISCUSSION AND CONCLUSION

RQ1: Human vs. Grok Essays: Lexical Complexity Compared

The current study indicated a slightly higher lexical density in human argumentative essays than in Grok 4-generated essays, although this difference was not statistically significant under the Bonferroni-adjusted alpha. Despite the non-significant result, the higher mean for human essays contrasts with the view that AI-generated texts are typically lexically denser and more representative of academic registers (Georgiou, 2025). According to Biber and Conrad (2019), academic discourse is characterized by dense vocabulary and conventionalized lexical bundles.

Regarding lexical sophistication, no significant differences were observed between human-authored and Grok 4-generated argumentative essays. This finding contrasts with previous studies reporting greater lexical sophistication in AI-generated writing (Georgiou, 2025; Zindela, 2023). One explanation for this discrepancy lies in differences in research design. The present study analyzed essays written on a single topic by highly proficient learners (IELTS band 8), whereas Zindela (2023) examined first-year university students writing on three different topics. In addition, lexical sophistication was operationalized differently across the two studies. Since lexical sophistication is a well-established indicator of L2 writing proficiency (Kyle & Crossley, 2015), the absence of significant differences suggests that both the human and Grok-generated essays reflected comparable proficiency levels.

Concerning lexical variation, Grok 4 argumentative essays demonstrated significantly higher TTR, NDW-ER50, and NDW-ES50 values than human essays. This pattern aligns with previous IELTS-based and corpus-driven studies reporting equal or greater lexical diversity in AI-generated argumentative writing (Fredrick & Craven, 2025; Kujur, 2025; Wu, 2025). The ability of AI systems to minimize lexical repetition under standardized conditions may account for these findings. Moreover, the significantly higher NDW-ER50 and NDW-ES50 values suggest that Grok 4 maintains greater lexical variation not only in raw TTR but also across token-controlled indices, extending previous evidence on lexical diversity in AI-generated academic texts.

The findings, however, contrast with Amirjalili et al. (2024), who reported higher TTR in a human-written literary essay than in a GPT-4-generated essay. This discrepancy is likely attributable to differences in genre and AI model. Whereas their study examined a literary essay, the present study focused on argumentative writing, a genre that naturally promotes lexical variation through stance-taking and evaluative language. The higher TTR observed for Grok 4 therefore suggests that lexical variation is neither uniform across genres nor consistent across AI models.

The current study further showed that human-authored argumentative essays exhibited greater lexical word, verb, adjective, and modifier variation than Grok 4-generated essays. These findings are partially consistent with Georgiou (2025) and Varga and Baksa (2025), who likewise reported greater verb and modifier variation in human writing. They are further supported by large-scale investigations demonstrating that human-written texts consistently display richer part-of-speech variation, particularly in verbs, adjectives, and modifiers (Akinwande et al., 2024; Reviriego et al., 2024).

RQ2: Human vs. Grok Essays: Syntactic Complexity Comparison

Regarding the indices measuring the length of the production unit, results for the MLT, MLC, and MLS across the two corpora showed no statistically significant difference. This pattern is consistent with recent studies reporting comparable performance between human writers and advanced AI models (Kujur, 2025; Liu & Liu, 2025; Wind, 2025). For example, Wind (2025) found no significant differences between ChatGPT-4.0 and human writing in MLC and MLS, while Liu and Liu (2025) reported convergence in MLS and MLT across particular rhetorical stages. However, these findings differ from those of Fredrick and Craven (2025) and Zanotto and Aroyehun (2024), who found that human-authored texts were longer than AI-generated texts. Such discrepancies may reflect differences in genre, corpus design, and the syntactic measures employed.

Regarding overall sentence complexity, Grok 4 essays exhibited significantly higher clauses per sentence (C/S) than human-authored essays. This finding aligns with Zanotto and Aroyehun (2024) who argued that AI-generated texts achieve greater syntactic density because they are not constrained by the cognitive demands of human writing. However, it contrasts with Wu (2025) and Liu and Liu (2025) who found that human essays demonstrated greater logical complexity through more sophisticated clause organization and higher C/S values.

With respect to subordination, Grok 4-generated essays showed a significantly higher complex T-unit ratio (CT/T), whereas C/T, DC/C, and DC/T did not differ significantly between the two corpora. These findings partially contradict Wu (2025), who reported higher DC/C and DC/T values in student essays, and Liu and Liu (2025), who found human writers consistently outperforming AI across multiple subordination indices. In contrast, the present findings suggest that Grok 4 differs primarily in the organization of subordination rather than its overall density. They are also partially consistent with Wind (2025), who reported no significant differences in DC/C between ChatGPT-4.0 and human argumentative essays.

The study further demonstrated significantly higher CP/C, CP/T, and T/S values in Grok 4 essays, indicating that their syntactic complexity is achieved primarily through coordinative expansion rather than deeper subordination. This finding is consistent with Wind (2025), Wu (2025), and Liu and Liu (2025), all of whom reported that AI-generated essays relied more heavily on coordination and surface-level structural expansion. Liu and Liu (2025) further argued that this pattern reflects the uniform and formulaic organization of AI-generated texts.

Analysis of phrasal sophistication showed no significant differences between the two corpora in CN/C, CN/T, or VP/T. These findings are partially consistent with Kujur (2025), who characterized AI-generated writing as linguistically rich, yet rhetorically formulaic. Although Kujur (2025) operationalized sophistication through lexical density rather than complex nominal indices, both studies suggest that the linguistic strengths of AI do not necessarily extend to deeper phrasal sophistication. In contrast, Nkhobo and Chaka (2023), Liu and Liu (2025), and Wind (2025) reported superior AI performance on some phrasal complexity measures, including noun-phrase modification and complex nominals, while Liu and Liu (2025) found human writers to exhibit greater verb-phrase density.

Conclusive Implications and Limitations of the Study

Using Grok 4 for generating writings and adopting Lu's (2010) model of syntactic and lexical complexity, the present study extends previous comparative studies by examining disparities between AI-generated and human-authored essays under controlled conditions of topic, genre, length, and proficiency. While prior studies reported significant lexical diversity and varying syntactic complexity in AI-generated writing, the current study revealed that Grok 4 neither simply

replicated nor entirely diverged from those patterns. Grok 4 demonstrated greater lexical diversity, whereas human writers showed greater lexical variation across part-of-speech-based indices. No significant differences emerged in lexical sophistication, density, or length-based measures. Grok 4 essays exhibited higher sentence complexity and greater reliance on coordination and complex T-units, supporting the view that AI writing differs systematically among models.

The present study also has practical implications for academic writing, particularly for argumentative essays. Instructor can use the Grok-generated essays as excellent contrastive models to elucidate the features of complexity and to facilitate their teaching. For example, Grok 4's inclination towards coordination provides a productive foundation for the classroom analysis. Student can differentiate between the coordinative expansion, subordination, and phrasal elaboration and then revise AI essays by restructuring information into more rhetorically effective and purposeful forms (e.g., embedding causal relationships, refining noun phrase modifications, or diversifying clause-combining strategies). The human superiority in part-of-speech-based variation highlights the importance of pedagogical emphasis on how lexical selections across grammatical categories enhance clarity and cohesiveness in argumentation. In practice, instructors may encourage students to identify AI passages that exhibit substantial lexical diversity and minimal variation in nouns, adjectives, or modifiers, and rewrite the passages to improve rhetorical accuracy through deliberate selection of nominal, adjectival, and modifier choices. Finally, as the corpora were homogenized for overall proficiency, the study advocates for considering the AI output not solely as a model text, but as a resource for guided revision.

This study has some limitations. First, the corpus was limited to IELTS Writing Task 2 argumentative essays written by Iranian EFL learners at a high proficiency level, limiting the generalizability of the findings to other genres, proficiency levels, or learner populations. Second, the analysis relied entirely on automated lexical and syntactic complexity indices, which do not capture discourse organization or rhetorical effectiveness. Future research should include multiple genres, proficiency levels, additional AI models, and multidimensional linguistic indices combined with qualitative analyses.

References

Ahmad, S., Khan, W. M., Nadeem, A., & Kashif, M. (2025). The role of AI in supporting writing development while sustaining deep learning processes. *Journal of Arts and Linguistics Studies*, *3*(2), 2993–3003. https://doi.org/10.71281/jals.v3i2.357

Akinwande, M., Adeliyi, O., & Yussuph, T. (2024). Decoding AI and human authorship: Nuances revealed through NLP and statistical analysis. *International Journal on Cybernetics & Informatics (IJCI)*, *13*(4), 85–103. https://doi.org/10.5121/ijci.2024.130408

Al-Zubaidi, K. (2025). The role of generative AI in higher education: Institutional guidelines, generational gaps, and the Grok 4 challenge. *Arab World English Journal (AWEJ)*, *11*, 1–4. https://doi.org/10.24093/awej/call11.1A

AlAfnan, M. A., & MohdZuki, S. F. (2023). Do artificial intelligence chatbots have a writing style? An investigation into the stylistic features of ChatGPT-4. *Journal of Artificial Intelligence and Technology*, *3*(3), 85–94. https://doi.org/10.37965/jait.2023.0267

Alqurashi, N., & Hamed, D. M. (2026). Phraseological patterns within human–artificial intelligence togetherness (HAIT): A corpus-based comparison of ESL development and AI simulation. *Sage Open*, *16*(1), 21582440261419890. https://doi.org/10.1177/21582440261419890

Amirjalili, F., Neysani, M., & Nikbakht, A. (2024). Exploring the boundaries of authorship: A comparative analysis of AI-generated text and human academic writing in English literature. *Front. Educ*, *9*. https://doi.org/10.3389/feduc.2024.1347421

Bahari, A. (2025). Balancing syntactic complexity and clarity: The role of AI in enhancing academic writing proficiency. *Saudi Journal of Language Studies*, *5*(4), 271–290. https://doi.org/10.1108/sjls-10-2024-0062

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press. https://doi.org/10.1017/9781108686136

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development? *TESOL Quarterly*, *45*(1), 5–35. https://doi.org/10.5054/tq.2011.244483

Bozorgian, H., & Rahimi, H. (2025). Peer e-feedback and chatgpt-4o in EFL writing: A cognitive-interpersonal comparison based on EFL students. *Journal of Language and Education*, *11*(4). https://doi.org/10.17323/jle.2025.27195

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, I. Vedder, & F. Kuiken (Eds.), *Dimensions of L2 performance and proficiency – Investigating complexity, accuracy and fluency in SLA* (pp. 21–46). John Benjamins.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing*, *26*, 42–65. https://doi.org/10.1016/j.jslw.2014.09.005

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics*, *28*(1), 147–164. https://doi.org/10.1111/ijal.12196

Csomay, E., & Prades, A. (2018). Academic vocabulary in ESL student papers: A corpus-based study. *Journal of English for Academic Purposes*, *33*, 100–118. https://doi.org/10.1016/j.jeap.2018.02.003

Daller, H., van Hout, R., & Treffers-Daller, J. (2003). Lexical richness in the spontaneous speech of bilinguals. *Applied Linguistics*, *24*(2), 197–222. https://doi.org/10.1093/applin/24.2.197

Dergaa, I., Chamari, K., Zmijewski, P., & Ben Saad, H. (2023). From human writing to artificial intelligence generated text: Examining the prospects and potential threats of ChatGPT in academic writing. *Biol Sport*, *40*(2), 615–622. https://doi.org/10.5114/biolsport.2023.125623

Etaat, F. (2026). Exploring linguistic fingerprints in human and AI-generated texts: An NLP-based approach in second language writing. *Ampersand*, *16*, 100258. https://doi.org/10.1016/j.amper.2026.100258

Fedoriv, Y., Pirozhenko, I., & Shuhai, A. (2023). Linguistic analysis of human- and AI-created content in academic discourse. *Journal of Vasyl Stefanyk Precarpathian National University. Philology*(10), 47–67. https://doi.org/10.15330/jpnuphil.10.47-67

Fredrick, D. R., & Craven, L. (2025). Lexical diversity, syntactic complexity, and readability: A corpus-based analysis of ChatGPT and L2 student essays [Original Research]. *Frontiers in Education, Volume 10 - 2025*. https://doi.org/10.3389/feduc.2025.1616935

Georgiou, G. P. (2025). Differentiating between human-written and AI-generated texts using automatically extracted linguistic features. *Information*, *16*(11), 979. https://doi.org/10.3390/info16110979

Godwin-Jones, R. (2022). Partnering with AI: Intelligent writing assistance and instructed language learning. *Language Learning & Technology*, *26*(2). https://doi.org/10125/44747

Gregori-Signes, C., & Clavel-Arroitia, B. (2015). Analysing lexical density and lexical diversity in university students' written discourse. *Procedia - Social and Behavioral Sciences*, *198*, 546–556. https://doi.org/10.1016/j.sbspro.2015.07.477

Hadji Jamel, A. J. (2025). AI-human writing divide: Pedagogical considerations. *International Journal of Linguistics, Literature and Translation*, *8*(12), 103–113. https://doi.org/10.32996/ijllt.2025.8.12.12

Halliday, M., & Matthiessen, C. (2014). *An introduction to functional grammar* (3rd ed.). Routledge.

Herbold, S., Hautli-Janisz, A., Heuer, U., Kikteva, Z., & Trautsch, A. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports*, *13*(1), 18617. https://doi.org/10.1038/s41598-023-45644-9

Housen, A., De Clercq, B., Kuiken, F., & Vedder, I. (2019). Multiple approaches to complexity in second language research. *Second Language Research*, *35*(1), 3–21. https://doi.org/10.1177/0267658318809765

Hunt, K. W. (1966). Recent measures in syntactic development. *Elementary English*, *43*(7), 732–739. https://doi.org/https://www.jstor.org/stable/41386067

Ismail, H. Y. S. (2023). Cohesion and coherence in essays generated by ChatGPT: A comparative analysis to university students' writing. *CDELT Occasional Papers in the Development of English Education*, *83*(1), 143–165. https://doi.org/10.21608/opde.2023.325331

Jaashan, H. M. S., & Bin-Hady, W. R. A. (2025). Stylometric analysis of AI-generated texts: A comparative study of ChatGPT and DeepSeek. *Cogent Arts & Humanities*, *12*(1), 2553162. https://doi.org/10.1080/23311983.2025.2553162

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, *53*, 61–79–61–79.

Kuhn, D., & Moore, W. (2015). Argumentation as core curriculum. *Learning: Research and Practice*, *1*(1), 66–78. https://doi.org/10.1080/23735082.2015.994254

Kujur, A. (2025). A comparative analysis of AI-generated and human-written text: Linguistic patterns, detection accuracy, and implications for modern communication. *SSRN*. https://doi.org/10.2139/ssrn.5833302

Kyle, K. (2016). *Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication* [Unpublished doctoral dissertation]. Georgia State University.

Kyle, K. (2020). Measuring lexical richness. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (1st ed., pp. 454–458). Routledge.

Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, *49*(4), 757–786. https://doi.org/10.1002/tesq.194

Lahuerta Martínez, A. C. (2018). Analysis of syntactic complexity in secondary education EFL writers at different proficiency levels. *Assessing Writing*, *35*, 1–11. https://doi.org/10.1016/j.asw.2017.11.002

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal*, *25*(2), 21–33. https://doi.org/10.1177/003368829402500202

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in l2 written production. *Applied Linguistics*, *16*(3), 307–322. https://doi.org/10.1093/applin/16.3.307

Liu, W., & Liu, X. (2025). A comparative analysis of syntactic complexity in argumentative essays from rhetorical perspective: ChatGPT vs. English native speakers. *PLoS One, 20*(8), e0329410. https://doi.org/10.1371/journal.pone.0329410

Lu, C., Bu, Y., Wang, J., Ding, Y., Torvik, V., Schnaars, M., & Zhang, C. (2019). Examining scientific writing styles from the perspective of linguistic complexity. *Journal of the Association for Information Science and Technology*, *70*(5), 462–475. https://doi.org/10.1002/asi.24126

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

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly*, *45*(1), 36–62. https://doi.org/10.5054/tq.2011.240859

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Mohammed, A. A. Q., Mudhsh, B. A., Bin-Hady, W. R. A., & Al-Tamimi, A. S. (2025). Deepseek and grok in the spotlight after chatgpt in english education: A review study. *Journal of English Studies in Arabia Felix*, *4*(1), 13–22. https://doi.org/10.56540/jesaf.v4i1.114

Nkhobo, T., & Chaka, C. (2023). Student-written versus ChatGPT-generated discursive essays: A comparative coh-metrix analysis of lexical diversity, syntactic complexity, and referential cohesion. *International Journal of Education and Development Using Information and Communication Technology*, *19*(3), 69–84.

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

Ortega, L. (2012). Interlanguage complexity: A construct in search of theoretical renewal. In B. Kortmann & B. Szmrecsanyi (Eds.), *Linguistic complexity: Second language acquisition, indigenization, contact* (pp. 127–155). DeGruyter.

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

Rehman, F., & Farooq, M. (2024). Instructor-led versus AI-generated feedback: Implications for academic writing development. *Journal of Language and Education*, *10*(3), 101–118.

Reviriego, P., Conde, J., Merino-Gómez, E., Martínez, G., & Hernández, J. A. (2024). Playing with words: Comparing the vocabulary and lexical diversity of ChatGPT and humans. *Machine Learning with Applications*, *18*, 100602. https://doi.org/10.1016/j.mlwa.2024.100602

Smythos. (2025). *What's new in Grok 4: Release facts, benchmarks, and value*. Retrieved December 10, 2025, from https://smythos.com/developers/ai-models/whats-new-in-grok-4-release-factsbenchmarks-and-value

Spring, R., & Johnson, M. (2022). The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and SpaCy tools. *System*, *106*, 102770. https://doi.org/10.1016/j.system.2022.102770

Staples, S., & Reppen, R. (2016). Understanding first-year L2 writing: A lexico-grammatical analysis across L1s, genres, and language ratings. *Journal of Second Language Writing*, *32*, 17–35. https://doi.org/10.1016/j.jslw.2016.02.002

Sulistyo, T., & Heriyawati, D. F. (2017). Reformulation, text modeling, and the development of EFL academic writing. *Journal on English as a Foreign Language*, *7*(1), 1–16. https://doi.org/10.23971/jefl.v7i1.457

Sun, S., & Li, Y. (2025). AI-generated, L2 learner, and native German writing: A comparative analysis of linguistic complexity. *Glottotheory*, *16*(2), 127–148. https://doi.org/10.1515/glot-2025-2011

Varga, E., & Baksa, A. (2025). Syntactic comparison of human and AI-written scientific texts. *Annales Mathematicae et Informaticae*, *61*, 248–260. https://doi.org/10.33039/ami.2025.10.013

Verma, A., & Arora, S. (2025). Authorship, originality, and linguistic merits in the era of artificial intelligence. *Journal of English Language Teaching*, *67*(3), 20 – 28. https://doi.org/10.66121/08m1rd81

Wang, C. (2023, September). *A syntactic complexity analysis of revised composition through artificial intelligence-based question-answering systems* 2nd international conference on artificial intelligence and computer information technology (AICIT), Yichang, China.

Wangsa, K., Karim, S., Gide, E., & Elkhodr, M. (2024). A systematic review and comprehensive analysis of pioneering AI chatbot models from education to healthcare: Chatgpt, Bard, Llama, Ernie And Grok. *Future Internet*, *16*(7), 219. https://doi.org/10.3390/fi16070219

Wind, A. M. (2025). Linguistic complexity and cohesive features of AI-generated and human-produced argumentative essays: A corpus-based analysis. *DEAL*, 155–175. https://doi.org/10.21862/ELTE.DEAL.2025.7

Wolfe-Quintero, K. E., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. University of Hawaii, Second Language Teaching and Curriculum Center.

Wu, J. (2025). Comparing linguistic features between human-written high-scoring IELTS essays and AI-generated ones. *Journal of Humanities and Social Sciences Studies*, *7*(8), 68–77. https://doi.org/10.32996/jhsss.2025.7.8.8

xAI. (2025). *Grok-4 and Grok-4 heavy release notes*. Retrieved December, 28 from https://x.ai/news/grok-4

Zanotto, S. E., & Aroyehun, S. (2024). Human variability vs. machine consistency: A linguistic analysis of texts generated by humans and large language models. arXiv:2412.03025. Retrieved December 01, 2024, from https://ui.adsabs.harvard.edu/abs/2024arXiv241203025Z

Zindela, N. (2023). Comparing measures of syntactic and lexical complexity in artificial intelligence and L2 human-generated argumentative essays. *International Journal of Education and Development Using Information and Communication Technology*, *19*(3), 50–68.