education sciences

MDPI

Article

How Generative AI Is Reshaping Student Writing: A Data-Driven Perspective for Writing Instructors

Maryam Eslami [1,*], Penelope Collins [1] and Bradley Queen [2]

[1] School of Education, University of California, Irvine, CA 92697, USA; p.collins@uci.edu
[2] School of Humanities, University of California, Irvine, CA 92697, USA; bradley.queen@uci.edu
* Correspondence: eslamim@uci.edu

Abstract

Generative AI has rapidly entered college writing classrooms, raising practical questions about how student texts are changing and what that means for instruction. This study analyzes 255 final-draft analytical essays written in first-year writing classes across three instructional contexts—pre-Gen-AI (Winter/Spring 2022), AI-prohibited, and AI-permitted with specified uses (Winter/Spring 2024). We combined holistic quality ratings of essays with Coh-Metrix indices of writing volume, lexicality, referential cohesion, and syntax. Analytically, we estimated a regression of essay quality on class type and demographics, and MANCOVAs (with essay score and demographics as covariates) for the four linguistic constructs. Essay quality did not differ by AI policy. However, compared to 2022, essays of AI-permitted classes were organized into fewer but shorter paragraphs; displayed greater lexical diversity and used less frequent, less familiar vocabulary; showed lower local and global anaphor overlap (other cohesion indices were stable); and exhibited lower verb-phrase, passive, and negation densities but higher gerund density. We interpret these as selective redistributions of linguistic resources rather than uniform gains or losses. For instructors, the actionable implication is two-fold: leverage AI-era gains in lexical precision while explicitly teaching referential continuity and clause-level strategies that sustain argumentative coherence.

Keywords: generative AI; college writing; digital tools; language and literacy development; corpus linguistics; Coh-Metrix

1. Introduction

Since late 2022, Generative AI (Gen AI), specifically large language models (LLMs) like ChatGPT, have become widely available to students, changing writing practices in higher education (Albadarin et al., 2024). Early findings indicate both rapid uptake and uneven implementation across courses and institutions. Empirical work is starting to explore the opportunities and risks for learning and literacy development (Abdallah et al., 2025). Initial meta-analytic and systematic reviews report small to moderate improvements in learning outcomes when AI is thoughtfully integrated, while cautioning about the importance of fit-for-purpose and instructional design (Deng et al., 2025; Jin & Sercu, 2025; J. Wang & Fan, 2025).

However, critiques have noted that meta-analytic claims are often overstated and based on limited outcome measures (Bardach et al., 2025). Bardach et al. (2025) encourage researchers to focus on meaningful outcomes, such as course writing performance, and

Academic Editor: Zhongfeng Tian

Received: 8 November 2025
Revised: 15 December 2025
Accepted: 17 December 2025
Published: 19 December 2025

interpret effects in relation to outcome type (distal versus proximal) rather than assuming uniform benefits across measures.

Our study responds to that call by examining how AI affects students' language choices and writing quality in their course essays. More specifically, we adopt Corpus Linguistics as the primary analytic framework to investigate the effects of AI on student writing. Corpus Linguistics enables us to treat texts as evidence of patterned language choices across comparable contexts such as genre, task, and length. These patterns can be measured through frequency, association, and co-occurrence profiles (Biber et al., 1998; Gries, 2015; McEnery & Hardie, 2011; Sinclair, 1991). This corpus-based approach allows us to move beyond anecdotal observations and examine whether particular linguistic features occur systematically more or less frequently across cohorts, while maintaining comparability and representativeness for generalization (Biber, 1994, 1995).

Additionally, we use Systemic Functional Linguistics to interpret how these distributional patterns shape meaning in context. Systematic Functional Linguistics treats language as a social-semiotic system through which writers make lexico-grammatical choices to express ideational meaning and to create cohesive textual organization (M. Halliday, 1994; M. A. K. Halliday & Matthiessen, 2014). In the present study, we focus on two metafunctions, cohesion (M. A. K. Halliday & Hasan, 1976) and ideational meaning, encompassing lexical density, nominal groups, and grammatical metaphor (Eggins, 2004; M. A. K. Halliday & Matthiessen, 2014). Biber and Gray (2016) connect these insights to corpus-observable trends in academic writing.

While our study centers on linguistic and functional analyses of student writing, it builds on a growing body of research that has begun to examine how generative AI affects writing performance and quality. This emerging body of literature provides context for understanding the potential mechanisms through which AI may influence language use and writing outcomes.

Recent studies indicate that generative AI can influence students' writing, yet its impact on writing quality depends on how it is used (Yang et al., 2025). When AI provides formative feedback aligned with tasks, students often produce clearer prose, stronger organization, and fewer surface errors. Additionally, AI-driven feedback can increase revision productivity and essay length (Amyatun & Kholis, 2023; Deep & Chen, 2025; Yan et al., 2024; Yang et al., 2025). These benefits typically occur when students remain the primary authors and use AI output as suggestions rather than replacements (Bahari, 2025; Liu & Jin, 2025). However, results vary across genres and courses, and are limited in scope, constraining generalization (Bahari, 2025; Deep & Chen, 2025). Conversely, when students rely on AI for authorship, their writing tends to be more generic, less aligned with disciplinary conventions, and may limit the development of students' authorial decision-making skills (Gültekin Talayhan et al., 2023; Marzuki et al., 2023; Reinhart et al., 2025).

Before addressing the effects of Gen AI on writing quality, it is essential to clarify what "writing quality" entails. Research has found that writing volume, lexical choice, referential cohesion, and syntactic organization are important contributors to judgments of writing quality (Crossley et al., 2014; McNamara et al., 2010). However, these features contribute in varying ways. Crossley et al. (2014) identified different profiles of successful writing that differ in their use of linguistic resources. For example, the academic style tends to be longer, consisting of more sentences and paragraphs, and uses lower-frequency words and more passive constructions. In contrast, the accessible style is typically shorter, features simpler syntax and cohesive overlap, and uses simpler vocabulary. Moreover, longer essays generally receive higher holistic ratings due to greater elaboration, more integrated evidence, and clearer structure (Attali & Burstein, 2006; Crossley et al., 2014; Kobrin et al., 2007). In short, high-quality writing can be realized through different combinations of lexical resources, cohesion, and sentence/clause architecture.

Building on this understanding, the present study examines how generative AI use influences these underlying linguistic dimensions of writing quality. Specifically, we focus on lexicality, cohesion, and syntax to capture how students use linguistic resources to construct meaning, organize information, and create textual flow. These may be considered indicators of writing sophistication and development and provide quantifiable evidence of how AI integration may shape language use in authentic academic settings.

1.1. Linguistic Dimensions of Writing

In the present study, lexicality, cohesion, and syntax are treated as complementary yet distinct dimensions of writing that together contribute to the quality and linguistic control in academic discourse. These dimensions have been widely recognized in both corpus-based and systematic functional linguistic research as central to writing proficiency (Biber & Gray, 2016; Crossley et al., 2014; M. A. K. Halliday & Matthiessen, 2014).

1.1.1. Lexicality

Lexicality refers to the range and density of words used to convey meaning. It captures features such as lexical diversity (the variety of word types), lexical sophistication (the use of low-frequency or abstract words), and lexical density (the proportion of content words relative to total words). Together, these features contribute to precision, representation of key concepts, and disciplinary alignment (Biber et al., 1998; Gries, 2015; McEnery & Hardie, 2011). Higher levels of lexical sophistical and diversity are often associated with more advanced academic writing, as they indicate greater precision, conceptual depth, and control over register (Crossley et al., 2011, 2014; McNamara et al., 2010).

Emerging research suggests that AI may influence the lexical features of writing in diverse ways. While AI-generated texts tend to contain longer or less frequent words, student-authored texts often exhibit greater lexical variety (Georgiou, 2024; Z. Wang, 2024). When students use AI primarily for formative feedback—rather than using AI-generated text verbatim—they tend to refine their language, replacing vague or wordy phrases with more precise, discipline-appropriate wording (Amyatun & Kholis, 2023; Z. Wang, 2024; Zhao, 2025). Thus, the nature and extent of the effects on lexical choices depend on how AI is integrated into the writing process, and may vary across genres and user roles (Georgiou, 2024; McNamara et al., 2010; Z. Wang, 2024).

1.1.2. Cohesion

Cohesion refers to the linguistic features that link clauses and sentences together, creating continuity and meaning within a text (M. A. K. Halliday & Hasan, 1976). In computational linguistics, cohesion is often examined through recurrence signals like noun and pronoun overlap or argument overlap, which help readers track entities across sentences (Graesser et al., 2004; McNamara et al., 2010). Within Systematic Functional Linguistics, these cohesive devices are interpreted through reference, lexical relations, and conjunction, which together structure the text into a coherent and logical whole (M. A. K. Halliday & Hasan, 1976; M. A. K. Halliday & Matthiessen, 2014). Maintaining referents and building causal connections are distinct yet interrelated dimensions of cohesion (M. A. K. Halliday & Hasan, 1976; McNamara et al., 2010; Thompson, 2013). In analytical writing, cohesive referential chains support argument tracking and conceptual development, whereas indiscriminate repetition can lead to redundancy. Thus, strong academic writing balances the repetition of key terms with structural cues for continuity (Crossley & McNamara, 2012; Crossley et al., 2014; McNamara et al., 2010).

Building on these theoretical and linguistic perspectives, recent research has begun to explore how such cohesive mechanisms appear in AI-generated and human-authored texts. AI-generated narratives may show high referential cohesion but reduced deep or causal cohesion (Zhou et al., 2023). A comparison of student-authored and ChatGPT -generated

discursive essays, revealed that student writing typically shows higher lexical density and greater referential cohesion (Nkhobo & Chaka, 2023). However, cohesion effects appear to be task-sensitive and may depend on whether AI serves as an assistant or the primary author (McNamara et al., 2010; Seo, 2024; Zhou et al., 2023). When AI is used to support revision rather than generate texts, it has been found to enhance the narrative and deep cohesion of EFL postgraduate students (Liu & Jin, 2025). Together, these findings suggest that the dimensions of cohesion may operate somewhat independently, shifting according to the role AI plays in text production.

1.1.3. Syntax

Syntax refers to the structural organization of language, encompassing the grammatical patterns through which words combine to form phrases and clauses to convey meaning and rhetorical goals (M. A. K. Halliday & Matthiessen, 2014; Thompson, 2013). In academic writing syntactic complexity manifests through features such as clausal embedding and phrasal or nominal packaging. Together, these features reflect how writers manage information density and abstraction (Biber, 1994; Biber et al., 1998; Biber & Gray, 2016). Importantly, syntactic complexity is not a linear indicator of quality; effective texts can achieve similar rhetorical goals with different structures depending on genre, stance and evidence integration (Biber & Gray, 2016; Crossley et al., 2014; Ortega, 2015). Within Systematic Functional Linguistics, these choices relate to transitivity (how processes, participants, and circumstances are represented) and to the organization of the text as a message (Eggins, 2004; M. A. K. Halliday & Matthiessen, 2014; Thompson, 2013).

Building on these theoretical and linguistic perspectives, recent research has explored how syntactic complexity should be interpreted in context. Effective academic argumentation can involve a range of clause- and phrase-level choices depending on stance, evidence integration, and disciplinary voice (Bahari, 2025; Crossley et al., 2014). For example, when ChatGPT was used to assist EFL students in revising narratives, their writing showed improved fluency and overall performance despite decreased global complexity metrics. However, clause-level analyses revealed richer embedded structures, such as an increase in clausal complements per clause (Seo, 2024). Other studies report selective changes in sentence-level features, such as greater use of passive constructions, non-finite forms (-ing/infinitives), and denser nominal phrases, suggesting a redistribution of linguistic resources rather than a uniform increase or decrease in syntactic "complexity" (Bahari, 2025; Fredrick & Craven, 2025; Seo, 2024). Collectively, these studies suggest that syntactic metrics should be viewed as a family of rhetorical choices, rather than a fixed measure of textual sophistication, that varies dynamically with the role of AI in the writing process (Bahari, 2025; Crossley et al., 2014).

1.2. Generative AI and Student Writing

A growing corpus-analytic literature examines how generative AI shapes student writing. Studies comparing AI-generated and student-authored texts show that AI often produces denser, nominalized prose characterized by frequent noun phrases and participle modifiers (Reinhart et al., 2025). The informationally compressed style generated by AI can impede readability. These results suggest that greater complexity does not necessarily translate to clear or higher writing quality. In contrast, when students use AI for revision or feedback rather than full text generation, their writing tends to show stronger cohesion, clearer organization, and greater lexical sophistication (Amyatun & Kholis, 2023; Yang et al., 2025). These findings suggest that the linguistic effects of AI depend not only on model output but also on students' roles and engagement during the writing process (Bahari, 2025; Seo, 2024).

1.3. The Current Study

Currently, much of the literature draws from small or homogeneous samples, emphasizes EFL contexts, or relies on teacher perceptions rather than direct linguistic analyses (Fredrick & Craven, 2025; Georgiou, 2024; Nkhobo & Chaka, 2023; Seo, 2024; Z. Wang, 2024; Zhou et al., 2023). Consequently, we know little about how instruction policies concerning AI use influences writing in first-year composition courses, or how the linguistic dimensions of lexicality, cohesion, and syntax shift under AI-permitted conditions.

To address these gaps, our study focuses on (a) authentic first-year analytical essays written across distinct instructional contexts (pre-AI, AI-prohibited, AI-permitted); (b) a coordinated examination of language-use families linked to reader experience (volume, lexicality, referential cohesion, syntax); and (c) connections to holistic writing quality as judged in courses. Guided by these foci, we formulate two exploratory research questions:

RQ1. Does classroom policy on AI usage influence the overall quality of student writing?

RQ2. How does classroom policy on AI usage influence the linguistic features of student writing? These features include writing volume, word usage, coherence, and syntax.

By integrating corpus-linguistics and systematic functional perspectives, the current study investigates how different instructional contexts (classes that permit the use of AI, classes that prohibit its use, and those that predate the public launch of AI) shapes the linguistic and rhetorical dimensions of undergraduate student writing in first-year composition classes. By focusing on lexicality, cohesion, and syntax, the study examines how classroom policies regarding AI use correspond with variation in students' linguistic and rhetorical choices. Including pre-AI writing serves as a baseline for understanding linguistic tendencies prior to the introduction of generative AI, while the AI-prohibited and AI-permitted conditions enable us to assess the influence of policy interventions with contemporary cohorts. This comparative approach enables us to examine whether access and regulation of AI alter the distribution of linguistic resources that underlie textual sophistication, coherence, and authorial control. Through this design, we seek to provide insight into how instructors' responses to generative AI may shape writing development.

2. Materials and Methods

2.1. Research Site

This work compares the quality and linguistic features of analytical essays written in composition courses that allow or implement the use of Gen AI, essays written in courses that have banned the use of Gen AI, and essays written by students in the winter and spring of 2022, before Gen AI was released to the public. The research was conducted at a highly selective university in a western state with a demographically, economically, and geographically diverse student body.

2.2. Data Sources

The corpus was compiled from final drafts of 2000-word Advocacy Project essays written by students in a lower-division writing course. All essays were written in English, including those authored by students who reported a non-English or multilingual first language. These argumentative essays covered topics chosen by the students, who received formative feedback on earlier drafts from their instructors and peers. The corpus included papers written under three conditions: Specified Use (generative AI was allowed for specific purposes), Prohibited Use (generative AI was not allowed) and Pre-Generative AI (before ChatGPT's release in 2022). Essays from the Specified Use and Prohibited Use conditions were written in the winter and spring terms of 2024, while those in Pre-Generative AI condition were written in the winter and spring terms of 2022.

The final sample included 255 essays: 85 from each condition. A power analysis using G*Power (Version 3.1.9.7) for multivariate analysis with a sample of 255, three groups, $\alpha = 0.05$, and power = 0.90 indicates that the smallest detectable effect size was $f^2(V) = 0.0502$, corresponding to a Pillai's trace of 0.048. This suggests that there was sufficient power to detect small to moderate effects in the planned MANCOVA tests (Cohen, 1988, p. 368). We also collected institutional demographic data, with Chi-squared tests showing no significant differences across conditions (see Table 1 for a summary of the demographics).

Table 1. Demographic Composition of Each Instructional Group.

Variable	Category	Prohibited (n = 85)	Specified Use (n = 85)	Year 2022 (n = 85)
Gender	Female	45	42	48
	Male	40	42	37
STEM	STEM	44	55	52
	Non-STEM	34	24	23
First Language	English only	28	28	28
	English/Non-English	35	35	35
	Non-English	22	22	22
First Generation	Yes	34	23	29
	No	51	59	56
Low Income	Yes	28	18	30
	No	57	67	55
URM	Yes	25	17	30
	No	60	67	52

All essays were final drafts of the same first-year writing course's program-wide Advocacy Project assignment, which maintains shared genre expectations, length targets (~2000 words), and a common rubric across sections. Essays were sampled from winter and spring quarters in both 2022 (pre-AI) and 2024 (Prohibited and Specified Use sections), which helps align timing within the academic year across cohorts.

Instructors adopted AI policies at the section level, and students enrolled through regular registration rather than being assigned, so the design is quasi-experimental rather than randomized. Students did not know about the course policy before enrollment. Demographic profiles together with essay score, were included as covariates in all MANCOVAs to partially account for pre-existing differences.

2.3. Measures

Writing quality. The quality of each essay was assessed using the university's Analytical Writing Placement Examination (AWPE) rubric. Although the AWPE is typically used as an on-demand writing assessment for incoming students, its overall holistic scoring system can be applied to undergraduates' argumentative writing on a six-point scale. Higher-scoring essays (scores of 5 or 6) exhibit clear arguments, appropriate and well-developed support, fluency of sentence construction, and a strong command of written English conventions. In contrast, lower-scoring essays (scores of 1 or 2) often fail to engage with the prompt, show illogical or poorly developed arguments, and contain frequent errors in syntax and grammar (see Appendix A for full rubric).

We used ChatGPT-4.0 to score essays using the following prompt: "Pretend you are a first-year college writing instructor. Score the attached essay based on the following holistic rubric and only write the score in the output, without any feedback or explanation." This prompt was followed by the full Analytical Writing Placement Examination (AWPE) rubric. Essays were submitted as written by the students, without any modifications to

paragraphing or formatting. The temperature was set to 0.1 to minimize randomness and enhance determinism in model outputs. Essays were submitted in a batch to eliminate variability and reduce human intervention, consistent with best practices for LLM reliability (Chang et al., 2024).

ChatGPT's scoring was compared with human ratings from a former Composition lecturer on a random subset of 64 essays (25%), yielding an 80% agreement and a Cohen's Kappa of 0.637, indicating substantial agreement (Landis & Koch, 1977). The internal consistency was high (Cronbach's $\alpha = 0.885$), suggesting that AI scoring be a valid proxy for holistic evaluation in large-scale writing analysis (Yan et al., 2024). Additionally, automated scoring mitigates common sources of human bias, including fatigue, implicit assumptions about student identity, and enhances reproducibility and transparency.

Language Use. We derived the linguistic characteristics of each text using Coh-Metrix 3.0, a tool that measures various aspects of language and discourse (Graesser et al., 2004; McNamara et al., 2014). Specifically, we focused on scores for writing volume, word usage (lexicality), degree of overlap across sentences and text (cohesion), and grammatical use (syntax). All the aforementioned indices were computed automatically by Coh-Metrix rather than rated by human coders or ChatGPT, so traditional interrater reliability does not apply to these measures.

Writing Volume. We used Coh-Metrix to tally fundamental textual characteristics such as essay length in terms of the *word count* (DESWC), *sentence count* (DESSC), and *paragraph count* (DESPC). We also tallied the mean *sentence length* in words (DESSL) and the mean *paragraph length* in sentences (DESPL).

Word Usage reflected the range and characteristics of the words that students used in their writing. Lexical diversity reflected the variety of words used in student essays and was operationalized using the Coh-Metrix indices LDTTRc, or the *type-token ratio* for content words, and LDMTLD, the *measure of textual lexical diversity*, which is less sensitive to essay length. Higher scores for both *type-token ratios* and the *measure of textual lexical diversity* reflect greater variety in word usage.

We used a range of indices to assess the difficulty and types of words students used. We analyzed two word-frequency indices: the *frequency of content words* (WRDFRQc) and the *frequency of all words* (WRDFRQa), where lower scores reflected greater use of rare words, suggesting more advanced vocabulary. *Age of acquisition* (WRDAOAc) measures when words are typically learned, with higher scores indicating more advanced words learned at older ages. *Familiarity* (WRDFAMc) reflects how commonly known a word is, where lower scores suggest rarer, more challenging words. *Concreteness* (WRDCNCc) gauges how tangible a word is, with lower scores associated with more abstract usage. *Imageability* (WRDIMGc) measures how well words evoke mental images, with lower scores reflecting less imageable, more abstract words. Finally, *meaningfulness* (WRDMEAc) reflects the semantic richness of words, with lower scores indicating fewer associations with other words.

Cohesion refers to the lexical and grammatical linking within sentences and texts that enhance meaning. Using Coh-Metrix 3.0 (Graesser et al., 2004; McNamara et al., 2014), we identified eight cohesion indices. We used two indices of *noun overlap*, the repetition of nouns across adjacent sentences (*local noun overlap*: CRFNO1) and across the passage (*global noun overlap*: CRFNOa), which help the reader identify key subjects and objects. *Argument overlap* involves the repetition of noun + role (subject/object) across sentences and was indexed at the local level (CRFAO1) and globally across the essay (CRFAOa). *Stem overlap* supports continuity of the essay topic and is operationalized as the repetition of word stems across adjacent sentences (CRFSO1) and globally (CRFSOa). Finally, *anaphor overlap*,

the consistent use of pronouns with their referents, was indexed locally (CRFANP1) and globally (CRFANPa). For all measures, higher scores reflected greater cohesion in essays.

We examined the syntax complexity of students' essays through the density, or frequency, of different syntactic structures measured in phrases per thousand words. These indices include *noun phrase density* (DRNP) with higher scores reflecting denser nominal packaging of information, *verb phrase density* (DRVP), reflecting the expression of actions, and *adverbial phrase density* (DRAP), which reflects the way writers modify verbs and adjectives. *Prepositional phrase density* (DRPP) reflects how writers modify, attribute, and link evidence. *Passive voice density* (DRPVAL)measures the use of passive constructions, common in the academic register (Schleppegrell, 2001). *Negation density* (DRNEG) tracks negative constructions, signaling stance or qualification. *Gerund density* (DRGERUND) accounts for -ing forms as nouns and relates to nominalization. Finally, *infinitive density* (DRINF) relates to purpose or goal framing in academic writing.

2.4. Procedures

Student essays were cleaned and prepared for Coh-Metrix analysis, following Bruss et al. (2004) guidelines. Nonessential content, such as headers, footers, personal identifiers, citations, and appendices, was removed to focus analysis on the core essay (Bruss et al., 2004). We also corrected spelling, grammatical errors, and formatting inconsistencies to ensure a cleaner dataset for analysis. The essays were saved in plain text format (.txt), with consistent use of line breaks, punctuation, and paragraphing, facilitating accurate parsing and analysis of their linguistic and coherence properties.

2.5. Analytic Plan

We addressed the first research question regarding how classroom AI policy affects essay quality through multiple regression analysis. The dependent variable was students' holistic scores. The predictor variables were class type (with Specified Use as the reference level) and the control variables, gender, first language (with English only as the reference level), major (STEM vs. non-STEM), and status as first-generation, low-income, or underrepresented minority.

To examine the effect of AI policy on the linguistic characteristics of students' essays, we calculated four sets of multivariate analyses of covariance (MANCOVA). For each MANCOVA, the independent variable was class type (Prohibited Use vs. Specified Use vs. Pre-GenAI), and the covariates were essay scores, gender, first language, first-generation status, low-income background, under-represented minority status, and major (STEM vs. non-STEM). The first MANCOVA examined group differences in writing fluency and included five dependent variables (*word count*, *sentence count*, *paragraph count*, *sentence length*, and *paragraph length*). The second MANCOVA addressed word usage and included nine dependent variables (*type-token ratios*, *measure of textual lexical diversity*, *frequency of content words*, *frequency of all words*, *age of acquisition*, *familiarity*, *concreteness*, *imageability*, and *meaningfulness*). The MANCOVA examining referential cohesion had eight dependent variables (*local and global noun overlap*, *local and global argument overlap*, *local and global stem overlap*, *and local and global anaphor overlap*). The fourth MANCOVA addressed syntactic usage and had eight dependent variables (*noun phrase density*, *verb phrase density*, *adverbial phrase density*, *prepositional phrase density*, *passive voice density*, *negation density*, *gerund density*, and *infinitive density*). Each MANCOVA was followed by a series of Bonferroni-adjusted univariate analyses of variance (ANCOVAs) to determine the specific variables that differed. Significant differences were followed with post hoc comparisons using Tukey's HSD or Games-Howell tests to identify specific group differences.

3. Results

Does classroom policy about the use of AI influence the overall quality of student writing?

Students' overall performance on the final drafts of their papers was quite high, with mean scores of 5.27 (s.d. = 0.61) in Prohibited Use classrooms, means of 5.38 (s.d. = 0.53) in Specified Use classes, and mean scores of 5.27 (s.d. = 0.54) in Pre-GenAI classes. Despite having sufficient power to detect moderate effects, the regression analysis shown in Table 2 revealed that neither class type nor student demographic characteristics predicted the quality ratings of the essays.

Table 2. Regressions predicting quality ratings as a function of class type and student demographics.

	Model 1	**Model 2**
Class Type		
Prohibited Use	−0.16	−0.12
Pre-Generative AI	−0.22	−0.18
Demographic Controls		
Male		−0.05
First Generation		−0.07
Low Income		−0.00
URM		−0.07
First Language: English & Non-English		−0.14
Non-English		−0.12
STEM Major		−0.07
Constant	5.37 ***	5.45 ***
R^2	0.01	0.03

Please note: All coefficients are standardized beta weights. The reference level for class type is Specified Use. The reference level for First Language is English Only. *** $p < 0.001$.

How does classroom policy on AI use influence the linguistic features of student writing?

Although classroom AI policy had little effect on students' final draft quality scores, we examined how AI use might influence students' language across four domains: writing fluency, word usage, coherence, and syntax. We first considered the effects on writing fluency, as shown in Table 3.

Table 3. Descriptive Statistics of Writing Fluency as a Function of Instructional Group.

Essay Characteristic	**Prohibited Use (n = 85)**	**Specified Use (n = 85)**	**Pre-GenAI (2022) (n = 85)**
Paragraph Count	11.06 [ab] (4.01)	11.26 [b] (3.93)	9.85 [a] (3.53)
Sentence Count	80.75 (20.90)	83.99 (21.86)	81.61 (22.50)
Word Count	2039.28 (481.55)	2162.49 (482.80)	2145.61 (519.65)
Paragraph Length in Sentences	7.91 [b] (2.57)	8.02 [b] (2.52)	8.92 [a] (2.77)
Sentence Length in Words	25.99 (4.26)	26.46 (3.85)	27.07 (5.03)

Please note: Standard deviations are presented in parentheses. Scores with different superscripts vary significantly.

As illustrated in Table 3, essay characteristics are broadly consistent across instructional conditions. When controlling for essay scores, gender, STEM major, income, and first-generation status, the MANCOVA revealed that the three instructional conditions showed comparable writing fluency, $F(10,434) = 1.52$, $p = 0.131$, yet gender was a significant covariate

$F(5,216) = 3.19$, $p = 0.008$. Subsequent Bonferroni-adjusted ANCOVAs revealed significant differences in the number of paragraphs used, $F(2,220) = 3.83$, $p = 0.052$, and the number of sentences contained in the paragraphs, $F(2,220) = 4.63$, $p = 0.01$. Tukey's post hoc tests revealed that students who wrote before the introduction of AI wrote longer paragraphs than students in the Specified and Prohibited conditions. Further, final essays in the Specified condition were organized using more paragraphs than those written before the introduction of generative AI. No other effects were significant.

Table 4 summarizes students' word usage as a function of instructional type. We found that essay scores, $F(9,212) = 3.55$, $p < 0.001$, and gender, $F(9,212) = 2.36$, $p = 0.015$, were significant covariates contributing to word usage. Word usage also varied as a function of the instructional condition, $F(18,426) = 2.28$, $p = 0.002$. Subsequent Bonferroni-adjusted ANCOVAs revealed that essay quality was a significant covariate for content word frequency, $F(1,220) = 14.21$, $p < 0.001$, overall word frequency, $F(1,220) = 13.22$, $p < 0.001$, word familiarity, $F(1,220) = 12.09$, $p < 0.001$, and word concreteness, $F(1,220) = 7.96$, $p = 0.005$, indicating that essays with higher scores used rarer words that were less familiar and more abstract. Gender was also a significant covariate for mean textual lexical diversity, $F(1,220) = 6.72$, $p = 0.010$, word concreteness, $F(1,220) = 8.27$, $p = 0.004$, and imageability, $F(1,220) = 7.97$, $p = 0.005$, suggesting that female students' word usage was more diverse and abstract. We found significant instructional group differences in mean textual lexical diversity scores, $F(2,220) = 3.55$, $p = 0.030$, content word frequency, $F(2,220) = 9.47$, $p < 0.001$, overall word frequency, $F(2,220) = 12.54$, $p < 0.001$, and word familiarity, $F(2,220) = 7.62$, $p < 0.001$. Tukey's post hoc tests revealed that students' essays from the Specified condition included more diverse vocabulary, consisting of less frequent, more diverse words, than those written before the introduction of generative AI. Further, essays written in the specified condition contained overall less frequent vocabulary than that of their contemporaries in the Prohibited use condition.

Table 4. Descriptive Statistics of Word Usage as a Function of Instructional Group.

Index	Prohibited Use Mean (SD)	Specified Use Mean (SD)	Pre-Generative AI Mean (SD)
Type-Token Ratio	0.53 (0.05)	0.52 (0.05)	0.52 (0.05)
Mean Textual Lexical Diversity	97.8 ^{ab} (17.2)	102.1 ^{a} (16.4)	95.6 ^{b} (16.7)
Content Word Frequency	2.07 ^{ab} (0.12)	2.02 ^{a} (0.14)	2.11 ^{b} (0.12)
Overall Word Frequency	2.86 ^{a} (0.10)	2.83 ^{b} (0.10)	2.90 ^{c} (0.08)
Age of Acquisition	391.3 (19.8)	394.5 (21.2)	393.0 (17.7)
Familiarity	563.1 ^{a} (7.0)	560.0 ^{b} (8.3)	564.9 ^{a} (7.5)
Concreteness	366.6 (15.4)	368.2 (15.9)	367.4 (16.3)
Imageability	398.4 (13.5)	399.2 (14.6)	398.8 (14.4)
Meaningfulness	424.9 (11.2)	422.6 (12.0)	424.3 (12.0)

Please note: Standard deviations are presented in parentheses. Scores with different superscripts vary significantly.

We next considered the cohesion of students' final drafts, as shown in Table 5. We found that underrepresented minority status, $F(8,213) = 2.29$, $p = 0.023$, was a significant covariate with coherence. Coherence also varied as a function of the instructional condition, $F(16,428) = 1.72$, $p = 0.040$. Subsequent Bonferroni-adjusted ANCOVAs revealed that underrepresented minority status was a significant covariate for local anaphor overlap, $F(1,220) = 7.45$, $p = 0.007$, and global anaphor overlap, $F(1,220) = 7.30$, $p = 0.0017$. We found significant instructional group differences that were limited to anaphor overlap, both locally, $F(2,220) = 4.52$, $p = 0.015$, and globally, $F(2,220) = 5.933$, $p = 0.005$. Tukey's post hoc tests revealed that students' essays from the Specified condition showed less

anaphor overlap than essays written before the introduction of generative AI. No other effects were significant.

Table 5. Descriptive Statistics of Cohesion as a Function of Instructional Group.

Index	Prohibited Use Mean (SD)	Specified Use Mean (SD)	Pre-Generative AI Mean (SD)
Local Noun Overlap	0.58 (0.13)	0.59 (0.13)	0.58 (0.13)
Global Noun Overlap	0.44 (0.12)	0.44 (0.12)	0.44 (0.13)
Local Argument Overlap	0.67 (0.11)	0.67 (0.12)	0.69 (0.12)
Global Argument Overlap	0.54 (0.12)	0.54 (0.12)	0.55 (0.13)
Local Stem Overlap	0.70 (0.11)	0.72 (0.11)	0.70 (0.13)
Global Stem Overlap	0.58 (0.12)	0.58 (0.12)	0.57 (0.13)
Local Anaphor Overlap	0.27 ^{ab} (0.09)	0.25 ^{a} (0.09)	0.29 ^{b} (0.12)
Global Anaphor Overlap	0.05 ^{ab} (0.02)	0.04 ^{a} (0.02)	0.06 ^{b} (0.04)

Please note: Standard deviations are presented in parentheses. Scores with different superscripts vary significantly.

Finally, we examined the density of the syntactic structures used in students' final drafts, as shown in Table 6. Although none of the covariates were significant, we found that the density of the different syntactic structures varied significantly as a function of the instructional condition, $F(16,436) = 2.21$, $p = 0.005$. Subsequent Bonferroni-adjusted ANCOVAs revealed significant instructional group differences in verb phrase density, $F(2,224) = 3.41$, $p = 0.035$, passive voice density, $F(2,224) = 6.58$, $p = 0.002$, negation density, $F(2,224) = 3.15$, $p = 0.045$, and gerund density, $F(2,224) = 6.46$, $p = 0.002$. Tukey's post hoc tests revealed that student essays written before the introduction of generative AI showed greater density in the use of verb phrases, the passive voice, and the use of negation than those written in the Specified Use condition. Furthermore, prior to the introduction of AI, students were less likely to incorporate gerunds into their academic writing. However, there were no significant differences in the syntactic patterns incorporated by the contemporaneous students in the Specified and Prohibited writing conditions. No other effects were significant.

Table 6. Descriptive Statistics of Syntactic Density as a Function of Instructional Group.

Index	Prohibited Mean (SD)	Specified Use Mean (SD)	Twenty-Two Mean (SD)
Noun Phrase Density	359.1 (15.9)	359.7 (18.9)	355.5 (18.3)
Verb Phrase Density	212.3 ^{ab} (18.9)	207.0 ^{a} (24.6)	215.8 ^{b} (26.7)
Adverbial Phrase Density	26.9 (6.1)	27.7 (7.4)	28.6 (7.0)
Prepositional Phrase Density	120.1 (12.2)	122.8 (12.8)	122.7 (13.7)
Passive Voice Density	8.8 ^{a} (3.4)	8.9 ^{a} (3.6)	10.5 ^{b} (3.9)
Negation Density	7.6 ^{ab} (3.5)	6.8 ^{a} (3.2)	7.8 ^{b} (3.2)
Gerund Density	27.8 ^{ab} (8.5)	29.4 ^{a} (8.4)	25.7 ^{b} (6.1)
Infinitive Density	23.7 (4.7)	23.2 (6.0)	24.4 (6.0)

Please note: Standard deviations are presented in parentheses. Scores with different superscripts vary significantly.

4. Discussion

This study examined how first-year undergraduate writing varies across three instructional contexts: pre-AI, AI-prohibited, and AI-permitted classrooms. It focuses on key linguistic dimensions: lexical choice, referential cohesion, and syntactic patterning, within a SFL framework. Rather than treating observed changes as cosmetic shifts, we interpret them as shifts in how students deploy language resources for meaning-making. By using

a corpus of comparable essays under varying classroom AI policy contexts, we identify cohort-level patterns with direct relevance for teaching and assessment.

Our findings underscore the utility of combining SFL with corpus linguistics to examine generative AI's influence on undergraduate writing. SFL posits that language choices reflect specific social purposes and contexts (M. Halliday, 1994), while corpus methods systematically quantify these choices. This integrated approach explains the linguistic shifts observed in AI-permitted classrooms: greater lexical sophistication aligns with the ideational function (representing complex, nuanced ideas), while shifts in referential cohesion and syntax reflect textual functions (structuring meaning for readers). In SFL terms, the ideational domain concerns how writers select and organize content, while the textual domain focuses on how writers guide readers through coherent argumentation. Our results suggest that students in AI-permitted courses redistribute effort across these domains, by expanding their lexical range yet showing thinner referential chains and shifting some syntactic shifts. These are not improvements or detractions, but a different distribution of linguistic resources calls for instructional recalibration.

Our regression analyses, controlling for essay scores and student covariates, show no meaningful differences in overall writing quality across AI policy types (Specified, Prohibited, pre-Gen-AI/2022). Class-type coefficients were small and statistically nonsignificant, indicating that students across conditions produced comparably rated final drafts.

However, notable differences emerged in how students structured essays of similar lengths. Compared to the 2022 cohort, students in the AI-permitted (Specified Use) classrooms wrote more, but shorter, paragraphs. This indicates a shift in rhetorical structure: while pre-AI essays concentrated into longer units, allowing for deeper ideation and conceptual density, the shorter paragraphs in Specified Use essays may be more readable at the expense of depth with ideas listed rather than developed (M. A. K. Halliday & Matthiessen, 2014).

4.1. Lexical Differences

AI-permitted essays displayed greater lexical variety, characterized by less frequent, more content-bearing words, hallmarks of academic argumentation (Crossley et al., 2012). This stylistic shift suggests that Ai may support more formal, information-rich prose, distancing students from conversational registers (Crossley et al., 2011; Crossley & McNamara, 2012). In contrast, students in the Year 2022 group relied more on frequent, familiar, and less diverse vocabulary while the Prohibited group showed intermediate lexical characteristics. This pattern aligns with corpus comparisons reporting that AI-aided or AI-generated texts tend to use longer words, more complex words, proxies for lexical sophistication (Georgiou, 2024; Kendro et al., 2025; Z. Wang, 2024).

Viewed through SFL, these lexical enhancements strengthen the ideational function, enabling students to express nuanced meaning with greater specificity (Eggins, 2004; M. A. K. Halliday & Matthiessen, 2014). Our corpus shows that AI-permitted essays tend to use more sophisticated and varied vocabulary, whereas 2022 essays rely on more frequent, familiar wording. Given that final drafts across cohorts were rated similarly, it appears that students can achieve comparable ratings writing quality through different lexical routes—a finding consistent with profile-based accounts of writing quality (Crossley et al., 2014)

Pedagogically, this suggests a broad implication rather than a prescriptive recommendation: AI-permissive classrooms may encourage vocabulary expansion, supporting precision and denser argumentation. The instructional challenge is to help students deploy this expanded vocabulary—anchored in evidence, genre norms, and audience expectations—so that lexical sophistication enhances argumentation rather than merely decorating prose.

4.2. Referential Cohesion Differences

Analyses of referential cohesion, specifically anaphor overlap, revealed consistent, though modest group differences. AI-permitted essays displayed *lower* anaphor overlap than those from the pre-AI cohort, suggesting weaker referential continuity. This implies that as vocabulary increases, writers may create fewer ties back to previously introduced concepts, potentially compromising textual cohesion.

Interestingly, this contrasts with studies of L2 narrative writing, where ChatGPT narratives showed stronger referential cohesion but weaker deep cohesion (Zhou et al., 2023). These discrepancies emphasize that cohesion operates in genre- and context-sensitive layers. What strengthens surface continuity in one genre or authorship role (e.g., narratives) may not strengthen the referential chaining students need to carry arguments in analytical writing.

Instructionally, this highlights the need to reinforce practices that maintain coherence, such as strategically reintroducing key terms and maintaining consistent references, without constraining students' range of expression.

4.3. Syntactic Density Differences

Most syntactic features were stable across conditions, but four indices showed significant variation: verb phrase density, passive voice density, gerund usage, and negation. The 2022 cohort exhibited higher verb phrase density, more negation, and a greater prevalence of passive constructions, possibly reflecting pre-AI instructional norms (Biber & Gray, 2016). In contrast, gerund usage was higher in the AI-permitted (Specified Use) group, suggesting stylistic shifts in syntactic packaging influenced by AI exposure.

Lower verb density and fewer passives and negations suggest a reduction in clause-level operations, whereas increased gerund density points indicate a rise in -ing forms functioning as nouns, a form of nominalization. In SFL terms, this signals a shift in clause grammar toward denser ideational packaging within nominal structures, known as grammatical metaphor. Rather than distributing meaning across more finite clauses, students appear to condense content within compact noun-like forms, an established way of producing concise, stance-managed academic prose (M. A. K. Halliday & Matthiessen, 2014; Schleppegrell, 2001).

However, noun phrase density remained consistent across groups, suggesting that this trend in nominalizations is selective, specific to gerunds, rather than a wholesale increase in nominal complexity. This supports the broader interpretation that syntactic changes represent a redistribution of linguistic resources rather than a generalized simplification.

While these patterns suggest that students in AI-permissive environments use these constructions more frequently, we cannot definitively attribute this to AI exposure, as it is plausible that the trend reflects evolving genre expectations or broader stylistic shifts in academic writing.

Across all the syntactic indices, the Prohibited group consistently exhibited intermediate values, showing comparable performance to both the 2022 and the Specified Use groups. Given that these students were subjected to AI bans in their classes, this intermediate position may reflect informal or extracurricular AI use, or the influence of uncontrolled factors. As such, the instructional impact of AI class policy remains somewhat ambiguous.

The instructional implication is one of alignment: instructors should make explicit which syntactic options best support genre conventions, evidence integration, and disciplinary voice, enabling students to make informed linguistic decisions based on rhetorical goals.

These findings highlight an important tension: AI-supported environments can foster lexical and stylistic innovation but may inadvertently undermine textual cohesion and disrupt established academic norms. Instructors must explicitly address linguistic struc-

tures, carefully balancing the benefits of technological support with the reinforcement of foundational academic writing skills. Moreover, the absence of demographic moderation effects suggests that these linguistic shifts are widely distributed across student populations and carry broad pedagogical implications.

5. Limitations and Future Directions

Several limitations should temper our conclusions. Our analysis focused on the final products of writing rather than real-time tool use and was limited to a single institution and genre (analytical/argumentative essays of comparable length). We compared cohorts rather than assessing the causal effects of a specific intervention; instructors selected AI policies at the section level and students self-enrolled, so the design is quasi-experimental. Essays came from winter and spring terms in both 2022 and 2024, and across most indices the Prohibited group tends to fall between 2022 and Specified Use and often does not differ significantly from either. These patterns argue against a simple uniform cohort shift from 2022 to 2024, but unmeasured instructor- or cohort-level differences and finer-grained pedagogical variation may still contribute to the observed effects. Moreover, because we analyzed final drafts, revised and instructor-reviewed, the essays reflect high quality, which may narrow observable differences and contribute to ceiling effects in the quality ratings.

Future work should pair corpus analyses with process data (e.g., revision logs or screen capture), extend to additional genres and institutions, and test whether targeted instructional supports can preserve referential cohesion and syntactic control alongside AI-driven gains in lexical richness.

Ultimately, the central issue is not whether AI improves or degrades student writing. Instead, it is how classroom conditions that permit AI use appear to redistribute students' linguistic efforts towards more varied and content-rich vocabulary, but with a corresponding need to reinforce coherence and calibrate syntax for rhetorical clarity. For educators, the task is to embrace AI's lexical benefits while redoubling efforts to teach the structural and syntactic practices that turn sophisticated wording into strong, persuasive academic prose.

Author Contributions: Conceptualization, M.E. and P.C.; methodology, M.E. and P.C.; software, M.E.; formal analysis, M.E. and P.C.; investigation, M.E.; resources, B.Q.; data curation, M.E. and B.Q.; writing—original draft preparation, M.E.; writing—review and editing, P.C. and B.Q.; visualization, M.E. and P.C.; project administration, M.E. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board of University of California Irvine (protocol code HS# 2016-2749 and date of last approval was 13 October 2022).

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The raw data supporting the conclusions of this article can be made available by the authors upon request.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A

The **2024—AWPE Scoring Guide**

Holistic reading requires scorers to assign essays to categories based on the dominant characteristics of the essay. The categories below are used to identify scores of papers in six distinct groups.

A 6 paper commands attention by engaging the material in an insightful and mature manner. The response is clear, logical, and convincing. Further, the response is fully developed, relying on well-chosen examples and persuasive reasoning. The 6 paper demonstrates a strong control of language, a fluid use of sophisticated sentences and a notable understanding of the conventions of written English.

A 5 paper is clearly thoughtful and competent. The response consistently relies on effective examples that are carefully developed. All parts of a 5 paper will be on topic, though the overall response will not be as accomplished or elaborated upon as a 6 paper. A 5 paper often has a less fluent and complex style than a 6 paper, but it does show a clear control over word choice and sentence variety while coherently observing the conventions of written English.

A 4 paper is satisfactory, though uneven in its delivery. It demonstrates an understanding of the TOPIC/prompt and presentation of a thesis, though that thesis may not be logically positioned or clearly stated. Still, the thesis does have support in the form of examples, though the examples may not be fully developed, and some examples may not directly support or relate to the position and topic. The reasoning may be marginal in some parts of the response, and the response in general will be less developed and accomplished than a 5 paper. A 4 paper demonstrates adequate accuracy in sentence control and sentence construction. The 4 paper recognizes and adheres to the conventions of written English.

A 3 paper demonstrates that the student will benefit from additional instruction in reading, logic, and sentence construction. The response may indicate an unclear understanding of the text or topic, and the examples may benefit from additional development and analysis. A 3 paper may be characterized by any of the following: little sentence variety, word choice that could be more appropriate and concise, occasional misunderstandings of major concepts in grammar and usage, or frequent minor misunderstandings at the sentence level.

A 2 paper indicates a student is likely to receive significant benefits from instruction in reading, essay organization, and sentence generation. Typically, the response is basic in nature and often presents unclear logic or coherence throughout, and because of these qualities, the essay often diverges from the text or topic. Its prose is usually characterized by at least one of the following: simplistic or imprecise word choice; use of a single type of sentence construction; many repeated misunderstandings of grammar and syntax.

A 1 paper shows that a student would benefit from both intensive and extensive instruction in all aspects of reading and writing. The response may not engage the topic's demands, or there may not be a recognized response to the basic aspects of the prompt. The response may be very brief in nature. It often displays a pervasive struggle with word choice and coherence at the sentence level.

References

Abdallah, N., Katmah, R., Khalaf, K., & Jelinek, H. F. (2025). Systematic review of ChatGPT in higher education: Navigating impact on learning, wellbeing, and collaboration. *Social Sciences & Humanities Open*, *12*, 101866. [CrossRef]

Albadarin, Y., Saqr, M., Pope, N., & Tukiainen, M. (2024). A systematic literature review of empirical research on ChatGPT in education. *Discover Education*, *3*(1), 60. [CrossRef]

Amyatun, R. L., & Kholis, A. (2023). Can artificial intelligence (AI) like QuillBot AI assist students' writing skills? Assisting learning to write texts using AI. *ELE Reviews: English Language Education Reviews*, *3*(2), 135–154. [CrossRef]

Attali, Y., & Burstein, J. (2006). Automated essay scoring with e-rater® V. 2. *The Journal of Technology, Learning and Assessment*, *4*(3). [CrossRef]

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. [CrossRef]

Bardach, L., Emslander, V., Kasneci, E., Eitel, A., Lindner, M., & Bailey, D. H. (2025). Research syntheses on AI in education offer limited educational insights. *Preprints*. [CrossRef]

Biber, D. (1994). *Using register-diversified corpora for general language studies, Using large corpora* (pp. 180–201). MIT Press.
Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press.
Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.
Biber, D., & Gray, B. (2016). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press.
Bruss, M., Albers, M. J., & McNamera, D. (2004). Changes in scientific articles over two hundred years: A coh-metrix analysis. In *Proceedings of the 22nd annual international conference on Design of communication: The engineering of quality documentation* (pp. 104–109). Association for Computing Machinery. [CrossRef]
Chang, K., Xu, S., Wang, C., Luo, Y., Liu, X., Xiao, T., & Zhu, J. (2024). Efficient prompting methods for large language models: A survey. *arXiv*, arXiv:2404.01077. [CrossRef]
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.
Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading*, *35*(2), 115–135. [CrossRef]
Crossley, S. A., Roscoe, R., & McNamara, D. S. (2014). What is successful writing? An investigation into the multiple ways writers can write successful essays. *Written Communication*, *31*(2), 184–214. [CrossRef]
Crossley, S. A., Salsbury, T., & McNamara, D. S. (2012). Predicting the proficiency level of language learners using lexical indices. *Language Testing*, *29*(2), 243–263. [CrossRef]
Crossley, S. A., Salsbury, T., McNamara, D. S., & Jarvis, S. (2011). Predicting lexical proficiency in language learner texts using computational indices. *Language Testing*, *28*(4), 561–580. [CrossRef]
Deep, P. D., & Chen, Y. (2025). The role of AI in academic writing: Impacts on writing skills, critical thinking, and integrity in higher education. *Societies*, *15*(9), 247. [CrossRef]
Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. *Computers & Education*, *227*, 105224.
Eggins, S. (2004). *Introduction to systemic functional linguistics*. A&c Black.
Fredrick, D., & Craven, L. (2025). Lexical diversity, syntactic complexity, and readability: A corpus-based analysis of ChatGPT and L2 student essays. *Frontiers in Education*, *10*, 1616935.
Georgiou, G. P. (2024). Differentiating between human-written and AI-generated texts using linguistic features automatically extracted from an online computational tool. *arXiv*, arXiv:2407.03646.
Graesser, A. C., McNamara, D. S., Louwerse, M. M., & Cai, Z. (2004). Coh-Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202. [CrossRef]
Gries, S. T. (2015). Quantitative designs and statistical techniques. In *The Cambridge handbook of English corpus linguistics* (pp. 50–71). Cambridge University Press.
Gültekin Talayhan, Ö., Babayiğit, M., & Öğrenci, Y. Z. Y. A. (2023). The influence of AI writing tools on the content and organization of students' writing: A focus on EFL instructors' perceptions. *CUDES Current Debates in Social Sciences*, *6*(2), 83–93.
Halliday, M. (1994). *Introduction to functional grammar* (2nd ed.). Edward Arnold.
Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Routledge.
Halliday, M. A. K., & Matthiessen, C. M. (2014). *Halliday's introduction to functional grammar*. Routledge.
Jin, Y., & Sercu, L. (2025). ChatGPT interventions in higher education: A systematic review of experimental studies. *Journal of Computer Assisted Learning*, *41*(4), e70072. [CrossRef]
Kendro, K., Maloney, J., & Jarvis, S. (2025). Do LLMs produce texts with "human-like" lexical diversity? *arXiv*, arXiv:2508.00086.
Kobrin, J. L., Deng, H., & Shaw, E. J. (2007). Does quantity equal quality? the relationship between length of response and scores on the SAT essay. *Journal of Applied Testing Technology*, *8*(1), 1–15. Available online: https://jattjournal.net/index.php/atp/article/view/48359 (accessed on 7 November 2025).
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174. [CrossRef]
Liu, H., & Jin, J. (2025). Exploring whether generative artificial intelligence can enhance Chinese EFL learners' academic writings. *International Journal of Information and Education Technology*, *15*(4), 752–759. [CrossRef]
Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education*, *10*(2), 2236469. [CrossRef]
McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
McNamara, D. S., Crossley, S. A., & McCarthy, P. M. (2010). Linguistic features of writing quality. *Written Communication*, *27*(1), 57–86. [CrossRef]
McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press.
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. (2015). Syntactic complexity in L2 writing: Progress and expansion. *Journal of Second Language Writing*, *29*, 82–94. [CrossRef]
Reinhart, A., Markey, B., Laudenbach, M., Pantusen, K., Yurko, R., Weinberg, G., & Brown, D. W. (2025). Do LLMs write like humans? Variation in grammatical and rhetorical styles. *Proceedings of the National Academy of Sciences*, *122*(8), e2422455122. [CrossRef]
Schleppegrell, M. J. (2001). Linguistic features of the language of schooling. *Linguistics and Education*, *12*(4), 431–459. [CrossRef]
Seo, J. Y. (2024). Exploring the educational potential of ChatGPT: AI-assisted narrative writing for EFL college students. *Language Teaching Research*, *43*, 1–21. [CrossRef]
Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.
Thompson, G. (2013). *Introducing functional grammar*. Routledge.
Wang, J., & Fan, W. (2025). The effect of ChatGPT on students' learning performance, learning perception, and higher-order thinking: Insights from a meta-analysis. *Humanities and Social Sciences Communications*, *12*(1), 1–21. [CrossRef]
Wang, Z. (2024). A corpus-linguistics-based comparison of AI-aided writing and students' writing. *Journal of Big Data and Computing*, *2*(4), 133. [CrossRef]
Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., & Gašević, D. (2024). Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, *55*(1), 90–112. [CrossRef]
Yang, K., Raković, M., Liang, Z., Yan, L., Zeng, Z., Fan, Y., Gašević, D., & Chen, G. (2025, March 3–7). *Modifying AI, enhancing essays: How active engagement with generative AI boosts writing quality*. 15th International Learning Analytics and Knowledge Conference (pp. 568–578), Dublin, Ireland.
Zhao, D. (2025). The impact of AI-enhanced natural language processing tools on writing proficiency: An analysis of language precision, content summarization, and creative writing facilitation. *Education and Information Technologies*, *30*(6), 8055–8086. [CrossRef]
Zhou, T., Cao, S., Zhou, S., Zhang, Y., & He, A. (2023). Chinese intermediate English learners outdid ChatGPT in deep cohesion: Evidence from English narrative writing. *System*, *118*, 103141. [CrossRef]