International Journal of TESOL Studies (2026)
Vol. 8 (3) 158-182 https://doi.org/10.58304/ijts.260510

Article

Flexibility versus Formulaicity: Comparing Phrasal Verb Use between Human-written and AI-generated Academic Essays

Siyang Zhou
The Hong Kong University of Science and Technology, China

Chen Chen*
Xi'an Jiaotong–Liverpool University, China

Qingyang Wu
University of Oxford, UK

Received: 2 February 2026/Received in revised form: 20 April 2026/Accepted: 27 April 2026/
Available online: 13 May 2026

Abstract
With the rise of Generative AI technology, numerous studies have compared linguistic features of human-produced and AI-generated writings. However, little attention has been paid to the use of phrasal verbs (PVs), a difficult type of two-part verbs, in academic writing. This study chose the 1.7-million-word arts and humanities essays from the British Academic Written English corpus and generated an equivalent AI-written corpus using ChatGPT3.5. Extracting PVs with an innovative dependency-based method, the authors compared the frequency, diversity, and disciplinary distribution of all the identified PVs, and examined the polysemy, semantic transparency, and collocations of the two PVs with the highest frequency in both corpora. Findings show that human writers used PVs approximately five times more than ChatGPT3.5 and demonstrated twice the PV diversity of ChatGPT3.5. There were also significant sub-disciplinary differences in both corpora, with essays in archaeology and linguistics using significantly fewer PVs than classics and comparative American studies. Regarding PV-specific comparison, humans used PVs with more meanings, more varied transparency, and broader collocations than ChatGPT3.5. Overall, this study revealed that human writing is more flexible, personal, and spontaneous, while AI writing is more formulaic, rigid, and predictable, which provides important implications for education, applied linguistics, and computer science.

Keywords
Artificial intelligence, ChatGPT, corpus analysis, writing, multi-word expressions, phrasal verbs

*Corresponding author. Email: Chen.Chen@xjtlu.edu.cn

1 Introduction

A recent breakthrough in the development of Artificial Intelligence (AI) technology is Generative AI (GenAI). Notable GenAI models include the Generative Pre-trained Transformer (GPT) series, Google Gemini, Claude, and DeepSeek. Among them, ChatGPT, developed by OpenAI, is one of the most influential models built upon the GPT architecture (Wu et al., 2023). As of 2026, ChatGPT boasts 900 million weekly active users, doubling from 400 million in February 2025 (Singh, 2026). ChatGPT can understand prompts and generate text in natural language, completing various complex linguistic tasks, such as answering questions, providing suggestions, and composing texts (M. Zhang & Crosthwaite, 2025). ChatGPT can even develop reading materials with different difficulties for students (Nguyen et al., 2026) and generate example sentences for vocabulary learning of elementary-level learners (Nakahara, 2025). As a result, ChatGPT has been widely applied across various fields, including education, training, content creation, and research assistance.

The proliferation of ChatGPT in higher education raises a potential concern about academic integrity, as some students or academics may use ChatGPT to replace human writing without disclosure. Since some journal reviewers and novice English teachers have difficulty recognizing AI-written works (Casal & Kessler, 2023; De Wilde, 2024), this may lead to ethical issues and harm educational equality. While students can consult and leverage ChatGPT in their writing process, it is important for both students and teachers to understand the differences between human-written and AI-generated content, to produce high-quality writing that effectively achieves its intended purpose. Therefore, it is necessary to compare the differences between GenAI writing and student writing systematically.

Researchers have compared GenAI writing and student writing in terms of vocabulary and multi-word expressions (Du, 2025; Jiang & Hyland, 2025a, 2025c; Reviriego et al., 2024; M. Zhang & Crosthwaite, 2025), syntax (Fredrick & Craven, 2025; W. Liu & Liu, 2025), and rhetorical styles (Jiang & Hyland, 2025b, 2025d; Reinhart et al., 2025). However, to the best of our knowledge, no study has analyzed PV in these comparisons. PVs are widely used by English L1 speakers for their flexibility, idiomaticity, adaptability, and efficiency (Siyanova & Schmitt, 2007; Thao & Bon, 2023). According to W. Li et al. (2003), PVs account for one third of the English verb vocabulary. Despite its common presence in informal discourse, researchers also noted the important role of PVs in some formal contexts such as academic writing, because PVs are perceived as "the most natural-sounding way of expressing a particular idea" (Fletcher, 2005, p. 212).

PVs have traditionally been a challenge to Natural Language Processing because they share the grammatical complication of lexicon and syntax (W. Li et al., 2003). With the development of GenAI, it remains unclear how Large Language Models (LLMs) handle PVs. By investigating the PV use in student essays and AI-generated ones, we argue that PVs may serve as a robust marker to distinguish human writing and machine output. The findings are expected to offer pedagogical implications for teachers of English for Academic Purposes and practical implications for computer scientists researching AI-writing detection. Understanding PV usage differences not only enables teachers to guide L2 learners in co-creating contextually-appropriate writing with AI across different registers, but also contributes to more effective identification of potentially AI-written text.

This study focuses on ChatGPT3.5 as an earlier-generation LLM, treating it as a representative case of initial AI writing strategies rather than a proxy for all contemporary AI systems. We will first review the related literature about AI-generated writing and PV research, before highlighting the gaps in the literature. We will then elaborate on our methodology and present PV findings from both the macro and micro levels, followed by an in-depth discussion. Finally, the implications and limitations will be addressed in the conclusion.

2 Literature Review

2.1 GenAI and writing assistance

According to a systematic review by B. Li et al. (2025), among all the research about AI in language education, writing was the most prominent topic, taking up 42.4 % of the total empirical studies in this field in 2023-2024. Research shows that the use of GenAI in academic writing is prevalent, but it often goes unnoticed. A study conducted at a British university showed that 94% of texts produced by ChatGPT were not identified by human markers in undergraduate assessments, and there was an 84.3% chance that AI-generated essays were graded higher than student-written essays (Scarfe et al., 2024). In the same vein, in the German context, AI-generated essays consistently demonstrated high quality and were also rated higher than human-written essays (Herbold et al., 2023). This phenomenon gives rise to concerns such as plagiarism and students' over-reliance on AI. Therefore, distinguishing GenAI writing from human writing is a pressing issue that should be tackled. Existing AI writing detectors have shown varied results, with common false positive and false negative results (Mizumoto et al., 2024). This justifies further investigation into effective parameters to differentiate GenAI writing from human writing beyond the existing ones.

2.2 Comparison of GenAI writing and student writing

Researchers have investigated the differences between AI-generated writing and human-produced writing from different aspects. Regarding syntactic differences, the L1 background of human writers seemed to play a decisive role. In a study of 73 essays, ChatGPT-4o mini demonstrated similar syntactic complexity as English L1 writers from America, but exhibited lower syntactic diversity (W. Liu & Liu, 2025). ChatGPT4.0 also used simpler syntactic constructions than British university students, who used a more balanced syntactic distribution in a study of 145 essays (Jiang & Hyland, 2025c). When compared with English L2 learners, studies yielded opposite results. ChatGPT3.5 demonstrated higher syntactic complexity than Japanese university students in 125 essays (Mizumoto et al., 2024). ChatGPT3 and ChatGPT4 also displayed higher syntactic complexity than German high school students (Herbold et al., 2023). Likewise, ChatGPT4's writing outperformed L2 students from the United Arab Emirates (UAE) in measures of syntactic complexity in 50 essays (Fredrick & Craven, 2025). Therefore, students' L1 background may have a direct influence on the comparison result.

Regarding lexical comparison, findings are also inconclusive. While it is widely agreed that AI-generated texts employ objective and formal vocabulary and more nominalization (Mizumoto et al., 2024), findings are inconsistent regarding lexical diversity. In single-word vocabulary, ChatGPT3.5 showed lower lexical diversity than human authors in 120 essays, while ChatGPT4.0 had a similar lexical diversity as humans, and sometimes even higher (Reviriego et al., 2024). A limitation of Reviriego et al.'s (2024) study is that they used several human corpora but did not give clear information on the L1 backgrounds of all the human datasets. In another study using two versions of ChatGPT, German high school students outperformed ChatGPT3 in lexical diversity of English, but were outshone by ChatGPT4 (Herbold et al., 2023). English L2 learners from the UAE demonstrated lower lexical diversity than that of ChatGPT4 (Fredrick & Craven, 2025). As for multi-word expressions, in a comparison of 145 essays, ChatGPT4.0 used fewer 3-word bundles than British university students, but they were more rigid and formulaic and showed greater lexical variation (Jiang & Hyland, 2025a). In a study of 30 essays, ChatGPT3.5 used more diverse bigrams with higher frequency than L2 learners studying in Australia, as L2 learners tended to rely on familiar and simpler word combinations (M. Zhang & Crosthwaite, 2025). In a small-scale analysis of 10 IELTS essays written by Chinese students, students exhibited greater diversity in verb collocations than ChatGPT4.0, which was asked to write like learners with IELTS scores

of Band 5.5-6.0 (Du, 2025). Two caveats of this study are the absence of inferential statistical tests and the unreasonable prompt restriction given by the researcher, which should be taken into consideration when interpreting the findings.

In short, the mixed findings on the writing performance of humans and AI may be attributed to four reasons: the target linguistic feature, the L1 background of learners, the ChatGPT version, and the prompt used by the researchers. Newer version of ChatGPT tends to outperform L2 writers in single-word vocabulary with effective prompts, whereas English L1 writers seem to have some advantage in multi-word expression performance over ChatGPT. The multi-word expressions of GenAI deserve further investigation due to the multifaceted nature of this concept. The current study aims to compare a unique type of multi-word expression— PVs—in AI writing and student writing, to reveal more nuanced differences in the multi-word expressions between human works and machine output.

In addition to the gaps mentioned above, the studies comparing AI-generated and human-produced writing typically employ relatively small corpora, which may limit the generalisability of their findings. Research on university students usually compiles small-sized learner corpora of L1 or L2 speakers of English, as a proxy for human writing (Du, 2025; Fredrick & Craven, 2025; Mizumoto et al., 2024; M. Zhang & Crosthwaite, 2025). The size of the student corpora ranged from 145 essays (Jiang & Hyland, 2025a) to seven essays (Nkhobo & Chaka, 2023). Only a few studies employing non-student human corpora had larger corpora size. For example, Reinhart et al. (2025) compiled AI corpora that parallel a 4-million-word human corpus from the Corpus of Contemporary American English (COCA), while Tudino and Qin's (2024) human corpus contains 0.47 million words drawn from The Elsevier OA CC-BY Corpus of published papers. However, Reinhart et al. (2025) used texts across spoken and written genres, while the data from Tudino and Qin (2024) only represented advanced academic writers. Research on English L1 student writing with larger sample sizes is still needed to enhance the generalisability of the findings. Therefore, the current study will employ a 1.7-million-word L1-student corpus and an equivalent-sized GenAI corpus, totalling 3.4 million words, directly addressing this limitation.

2.3 PV research in learner corpus

In the current study, PVs, a subtype of multi-word expressions, are defined as "two-part verbs consisting of a lexical verb followed by a contiguous (adjacent) or noncontiguous adverbial particle" (Gardner & Davies, 2007, p. 341). We adopt this definition because it is clear-cut without ambiguity, and it has been widely adopted in the PV lists analyzed by several key studies in the field (Garnier & Schmitt, 2015; D. Liu, 2011; D. Liu & Myers, 2018).

PVs are worth researching because of their ubiquity in day-to-day English (Gardner & Davies, 2007) and their emerging importance in academic writing (Alangari et al., 2020). Gardner and Davies (2007) estimated that learners, on average, encounter one PV in every 150 English words they are exposed to and two PVs per average page of written text. Although English L2 learners rarely notice PVs because of their multi-word composition, PVs are an integral part of idiomatic and natural language expression (Siyanova & Schmitt, 2007). Therefore, some intervention studies were conducted to improve students' PV knowledge (Teng, 2020). Traditionally, PVs are usually avoided in academic writing because of their colloquial tone (D. Liu, 2011). However, Leech et al. (2009) have observed a colloquialization trend in academic writing since the late 20th century, which reflects an informalization of writing styles. The frequent appearance of PVs in published journal articles lent support to this trend (Alangari et al., 2020). Therefore, analyzing the PV usage in human and AI writings can help L2 learners better use GenAI writing assistance and develop stronger genre awareness to know when and how to use PVs appropriately. This study will take a step further from Alangari et al. (2020), who only researched PV used in applied linguistics journal articles, by comparing the frequencies of PVs used in different sub-disciplines in the arts and humanities, to reveal potential disciplinary differences of academic PV use.

To date, corpus analysis of PVs has been carried out on English L2 learners of various L1 backgrounds (Badem & Şimşek, 2021; Ryoo, 2013; Wei, 2021) and some English L1 speakers (M. Chen, 2013). Regarding quantitative analysis of PVs, PV frequency is the most common inquiry. Researchers not only count how many PVs appear in total in the given corpus, but also calculate which PVs appear most often in the given corpus (Badem & Şimşek, 2021), ranging from top 10 PVs (Wei, 2021) to top 25 PVs (Waibel, 2007). Besides, researchers often compare the PV rankings or PV frequencies against those in the English L1 corpora, to identify potential underuse or overuse patterns of PVs in academic writing by L2 learners (H. H.-J. Chen et al., 2015; Tran & Tran, 2019). PV diversity could be another direction of PV quantitative analysis. To the best of our knowledge, no study has examined PV diversity. Similar to lexical diversity measures, PV diversity can be operationalized by Type/Token Ratio (TTR) (or its variants to reduce the influence of text length), where *type* refers to the number of different PVs used and *token* means the total number of PVs, including repeated ones. A higher value of TTR suggests higher PV diversity. The current study will analyze PV frequency, rank the top 20 PVs, and compare the PV diversity in human and GenAI corpora, which can provide a quantitative overview of their mastery of this linguistic structure.

When it comes to qualitative analysis, case study is a less common approach for PV analysis. For example, Kiativutikul and Phoocharoensil (2014) analyzed 500 concordances for three PVs ("*carry out", "find out", "point out"*) respectively in COCA, to compare their genre distributions (e.g. academic, spoken registers), collocations, and grammatical patterns. This study reported different usage patterns of these PVs and revealed that dictionaries could not capture all the real-life uses of PVs. This justifies the need for conducting case analyses of high-frequency PVs in this study.

As qualitative analysis can compensate for the limitations of quantitative findings, the present study selects the most frequent PVs from both corpora for detailed case analyses of their polysemy, transparency, and collocational patterns. Polysemy refers to the multiple meanings of the same PV. According to Liu and Myers (2018), among the most frequent 150 PVs in English (D. Liu, 2011), 70.66% PVs had significant differences in their meaning distribution in the spoken subcorpus and the academic subcorpus in COCA, and 2-3 meaning senses were identified per PV in the subcorpus. Semantic transparency means the extent to which the PV's meaning can be deduced literally. Transparency affects L2 learners' PV acquisition and leads to PV avoidance (Liao & Fukuya, 2004), but how GenAI tools deal with such nuances has not been explored. As for collocations, analyzing the collocations of PVs can reveal the formulaicity and predictability of GenAI language, with more diverse collocations suggesting lower predictability. The only study that applied concordance analysis of formulaic language in GenAI corpora is probably the study by Tudino and Qin (2024). They compared the two most frequent 4-grams in two ChatGPT corpora (i.e., *a nuanced understanding of; the finding of this*) and found that ChatGPT produced similar collocations across different essays, regardless of whether it was prompted to use precise academic language or not. It implied ChatGPT's tendency to select the most strongly-associated tokens based on its algorithm, demonstrating high formulaicity and low creativity. Our concordance analysis can shed new light on the collocation pattern of PVs in ChatGPT writing.

2.4 Gaps in the literature and research questions

The current study aims to make four contributions to the field. First, this study pioneers a comparison of PVs in English-L1 student writing with AI-generated writing, as PVs might be a salient linguistic feature to distinguish AI-generated and human texts. Past L2-focused research on PVs often draws on findings from L1 speakers' data (Garnier & Schmitt, 2016; Sonbul et al., 2020). Therefore, although our study uses data from English L1 speakers, the findings still hold value for L2 learners of English. The current study employs mixed methods to provide a comprehensive understanding of the phenomenon. Second, the study focuses on the particular discipline of the arts and humanities, and it also aims to uncover potential sub-disciplinary stylistic differences in PV use within this field. We chose to focus on this discipline because it has not been thoroughly investigated before, and it may exhibit more idiosyncratic

linguistic features among different writers, unlike scientific writing, where creativity in language use is relatively limited. Third, the study employs a large corpus comprising 3.4 million words, including 1.7 million tokens from 679 essays in each of the parallel corpora. Previous studies comparing learners and GenAI writing typically had a small corpus size of 20 to 145 essays. A larger sample size can enhance the generalisability and reliability of the findings on potential differences in PV usage patterns between humans and AI, potentially explaining some contradictions in previous studies mentioned above. Lastly, methodologically, this study employs a novel approach to identifying PVs in corpora by relying on dependency parsing, which is found to be more efficient than traditional corpus-based methods (details are provided in Section 3.3).

Based on the review above, two research questions (RQs) are proposed:

RQ1: How do PVs in human writing differ from GenAI writing based on ChatGPT3.5 in terms of frequency, diversity, and distribution across sub-disciplines in the arts and humanities?

RQ2a: To what extent do the top 20 PVs used by humans and ChatGPT3.5 resemble the actual PV usage pattern in the academic register of COCA?

RQ2b: How do humans and ChatGPT3.5 differ in their use of the top PV in terms of polysemy, meaning transparency, and collocational patterns?

3 Methodology

3.1 Data collection

To compare academic English written by UK university students with parallel texts generated by ChatGPT, we utilized the British Academic Written English (BAWE) corpus (Heuboeck et al., 2010) and created a corresponding ChatGPT-generated corpus (hereafter referred to as BAWE_GPT). The BAWE corpus is publicly available[1]. The texts in the corpus were produced by undergraduate university students studying in the UK. The corpus was collected during 2004 – 2007. The BAWE corpus spans four disciplinary categories: arts and humanities, life sciences, physical sciences, and social sciences. For this study, we selected the arts and humanities component, which constitutes approximately 25% of the total corpus and includes disciplines such as archaeology, classics, comparative American studies, English, history, linguistics, and eight others. We chose this discipline because of its argumentative and analytical nature, compared with others. Genres requiring academic sources (e.g., literature reviews, N=7) or personal experiences (e.g., methodology recounts, N=17; empathy writing, which involves reflective personal insights, N=1) were excluded, as ChatGPT's outputs in these areas may introduce hallucinations or lack authenticity. The BAWE and BAWE_GPT corpora comprise 679 essays each (after exclusions), totalling approximately 1.7 million words per corpus.

3.2 Prompt engineering process for generating academic essays

This study employs an iterative prompt-engineering process to address challenges in generating academic essays using LLMs, including originality, structural issues, and output length limitations. This approach ensures that the generated essays align with academic writing standards. The process was developed through a sequence of refinements using ChatGPT3.5, progressively improving prompt design and output quality.

Step 1: Initial direct input and basic prompting

The process began with exemplar-based prompts: "Please write an essay with a similar topic, genre, writing style, structure, and word limit as the above example." However, the outputs often exhibited

excessive similarity to the source texts. To enhance originality, the prompt was revised to focus on the essay's title and word count (Y. Liu et al., 2023), which yielded more creative content but often produced disorganized formats that occasionally deviated from academic writing norms, such as lacking clear structure or formal register.

Step 2: Incorporating academic writing requirements and persona

Subsequent prompts incorporated academic criteria and author roles to improve coherence and register appropriateness. Instructions such as "Please act as a university student to write an academic essay about <title>, including theories, examples, and citations" yielded more sophisticated essays but also hallucinated citations. Comparative testing revealed that prompting with terms such as "critical thinking" or "theories" yielded minimal improvements, suggesting that such terms have a limited impact on output quality. To maintain authenticity while minimizing fabrication, we simplified prompts and adopted a persona-based approach (White et al., 2023), which improved tone but often resulted in shorter-than-required texts.

Step 3: Outline-based structuring and part-wise generation

To overcome output length limitations, we introduced a two-stage process: (1) generating an outline with word allocation per section, and (2) producing each section sequentially based on the outline: "I'd like you to act as a university student who will submit an academic essay. Please write a <word count> essay entitled <title>. Before that, could you provide a brief outline and indicate how many words are allocated to each section?" This was followed by targeted requests: "Based on your outline which includes <number of parts> parts, could you write the first/second/third... part with <word count> words?" This approach ensured balanced structure, coherence, and controlled output length.

Overall, the iterative refinement of prompts—from direct imitation to structured generation—enabled improved academic essay quality while mitigating common LLM limitations such as content similarity, disorganization, and length inconsistency.

3.3 Data Analysis

To identify PVs, we followed a multi-step process. First, we used CoreNLP (Version 4.5.4) (Manning et al., 2014) to perform part-of-speech (POS) tagging and dependency parsing on both corpora. The taggers were reported to have accuracies of 97.3% (Stanford NLP Group, 2020) and 92.2% (D. Chen & Manning, 2014), respectively. To ensure precision, four research assistants, all university students majoring in English Studies and trained in syntax courses with components on dependency parsing, manually verified the POS tags, syntactic labels, and dependency distances. Prior to individual review, the assistants collaboratively annotated approximately 10% of the texts in pairs, followed by discussions to refine annotation criteria and resolve discrepancies. This process yielded an inter-rater reliability (Cohen's kappa) of above 90%. The remaining 90% of the texts were then independently reviewed by the assistants.

In dependency-parsed texts, PVs were identified by locating particles tagged as "compound:prt" that link to a verb, with the linkage indicated by the dependency distance (i.e., the position of the verb relative to the particle). This method proved more efficient than traditional approaches that rely solely on POS tags for PV identification. To avoid excluding PVs of which the particle may be incorrectly tagged, we also included two other cases: "advmod" + RP (particles sometimes tagged as adverbs), and "obl" + "case" (prepositional verbs like "look into"). The code resulted in 4192 PVs from BAWE and 763 from BAWE_GPT. After careful manual screening and reviewing by the two authors, we identified 3166 PVs from BAWE and 539 PVs from BAWE_GPT.

To answer RQ1, we checked the assumptions of parametric testing and performed independent-samples t-test in SPSS to compare the frequency and diversity of PVs across the two parallel corpora. For frequency, we calculated the raw and normalized PV frequency[2], to account for the different lengths of essays. For diversity, Corrected Type/Token Ratio (CTTR) (Carroll, 1964) was chosen over TTR because it minimizes the impact of essay length on PV diversity. CTTR is calculated by dividing the number of word types by the square root of twice the total number of tokens. Two-way between-groups ANOVAs were further performed in SPSS to examine the effects of discipline and authorship, as well as their interaction, on both the PV frequency and diversity. Residuals of the ANOVA were checked on R.

To answer RQ2a, we compared the frequency patterns of the top 20 PVs in the two corpora and contrasted these patterns with those of the top PVs in COCA (Liu, 2011), to evaluate their representativeness to L1 speakers' megacorpus. It should be noted that COCA includes more informal subcorpora than formal academic subcorpora. Therefore, the PV ranking patterns may not be free of a spoken language bias. For RQ2b, we analyzed all the concordance lines of the most frequent PV in BAWE ("*point out*") and BAWE_GPT ("*open up*"), respectively, in terms of their polysemy, transparency, and collocation patterns. For polysemy, we manually coded the meaning of each concordance with reference to the PV definitions provided in the online Cambridge Dictionary. For transparency, we evaluated the transparency of each PV and assigned each meaning to one of the three levels: opaque (level 3), semi-transparent (level 2), and transparent (level 1). Garnier and Schmitt (2016) coded PV transparency dichotomously into two levels, while Zhang and Wen (2019) used a 5-point scale to judge its transparency. We decided to use a three-point scale because it can capture the nuance of semantic transparency without over-complicating the differences. For collocations, after screening out the PVs used intransitively, we listed all the collocates for the target PVs and counted their raw frequency. Flexibility is operationalized by the diversity of collocates and the meanings for each PV. The more unusual the collocates are, and the more meanings of a PV are used, the more flexible we perceive its use.

4 Findings

4.1 RQ1: Overall comparison between PVs used by humans and ChatGPT

In this section, we will compare the PV frequency and diversity, with descriptive statistics presented before inferential statistics results. In terms of frequency, within the parallel corpora consisting of 679 essays each, human writers used 567 different PVs 3166 times in 625 essays, while ChatGPT used only 109 different PVs 539 times in 302 essays. In each human-written essay, PVs appear on average 3.88 times, whereas in each ChatGPT-authored essay, PVs appear on average 0.79 times (Table 1). In other words, human writers had a much higher frequency in using PVs in any of these calculated measures. The standard deviation (SD) of PV appearance in human-written essays was also higher than that of ChatGPT-generated essays, suggesting greater individual variation in human essays and more lexical uniformity in AI-written essays.

Table 1

Frequency of PVs in the Two Corpora

	BAWE	BAWE_GPT
	Mean (SD)	Mean (SD)
Raw PV frequency per essay	3.88 (3.43)	0.79 (1.34)
Normalized PV frequency per essay	1.94 (1.51)	0.32 (0.50)

In inferential statistics, a Shapiro-Wilk test was performed, to check the normal distribution of the raw and normalized frequencies of PVs in human essays and AI-written essays. Although both p values < 0.05, parametric testing is robust against violations (Glass et al., 1972). Therefore, independent-samples t-tests were performed to determine if there were significant differences in the PV frequency between human-written and AI-generated essays. Results in Table 2 show that the difference was significant with a large effect size when using either raw or normalized frequency data (similarity due to the similar number of tokens in the parallel corpora).

Table 2

Independent-samples t-test result on PV frequency in BAWE and BAWE_GPT

Variable	t	df	p	Cohen's d
Raw PV frequency	-21.90	1356	< 0.001	-1.435
Normalized PV frequency	-26.44	1356	< 0.001	-1.189

It should be noted that in human-written essays, 286 PVs appeared only once, taking up about half of the total PV types. They were used with both literal and figurative meanings, some of which are typically associated with informal contexts. For example, "bubble over" was used once, meaning to be very excited and enthusiastic (Cambridge Dictionary, 2025a). "Churn out" was used once to express producing large amounts of something quickly (Cambridge Dictionary, 2025b). This indicates that human writers occasionally employ colloquial PVs in academic writings, reflecting a flexible and emotive lexical choice and a dynamic and fluid language style. In contrast, in ChatGPT essays, only 54 PVs appeared once. Although they also accounted for about half of the total PV types, they tended to be less colloquial and emotional.

In terms of PV diversity, human-written essays exhibited a high CTTR of 1.17 (SD = 0.57), whereas ChatGPT-written essays showed a lower CTTR of 0.34 (SD = 0.40). This implied that human writers had three times the PV diversity of ChatGPT. Despite the significant results in the Shapiro-Wilk test ($p < 0.05$), an independent-samples t-test was run to compare the PV diversity between the two corpora, given the reason explained above. The difference was also significant with a large effect size: $t(1356) = 30.64$, $p < 0.001$, Cohen's $d = 1.663$.

Regarding the distribution of PV frequency across sub-disciplines, a two-way between-groups ANOVA was conducted to examine the effects of sub-discipline and authorship type (human versus AI) on PV frequency in essays. Assumption checks indicated a violation of the homogeneity of variances (Levene's test, $p < 0.001$) and a non-normal distribution of the dependent variable (see above). Nevertheless, since our dataset has a large sample size (N=1358), parametric tests are still considered appropriate, as they can mitigate the impact of these violations (Lumley et al., 2002).

Results showed that PV frequency varies significantly between different sub-disciplines [F (7, 1342) = 6.140, $p < 0.001$, $\eta^2 = 0.031$], different authorship [F (1, 1342) = 492.190, $p < 0.001$, $\eta^2 = 0.268$], and the interaction between the sub-disciplines and authorship [F (7, 1342) = 3.454, $p = 0.001$, $\eta^2 = 0.018$]. The large effect size in the authorship difference echoed the significant result in the independent-samples t-test. For disciplinary differences, essays in archaeology and linguistics used significantly fewer PVs than classics and comparative American studies in both human writing and AI writing. The significant interaction indicated that the differences in PV frequency between AI and human writers varied across disciplines (Figure 1). In some sub-disciplines, these differences are greater than in others. For example, the mean PV frequency difference between humans and AI in classics was up to 4.79, compared with only 2.35 in archaeology. The mean PV frequencies by both author types in each sub-discipline are listed in Table 3. Lastly, inspection of the Q–Q plot (Figure 2) indicated that residuals were approximately normally distributed, as the points were generally aligned along a straight line, so violations of the normality assumption are unlikely to introduce substantial bias (Shatz, 2023).

Figure 1

The Interaction between Authorship and Sub-disciplines in PV Frequency

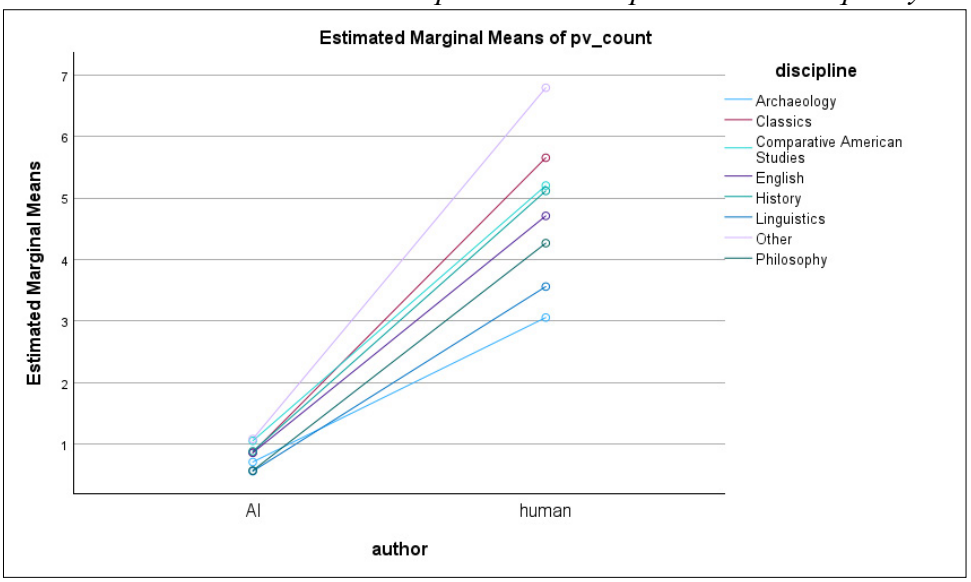


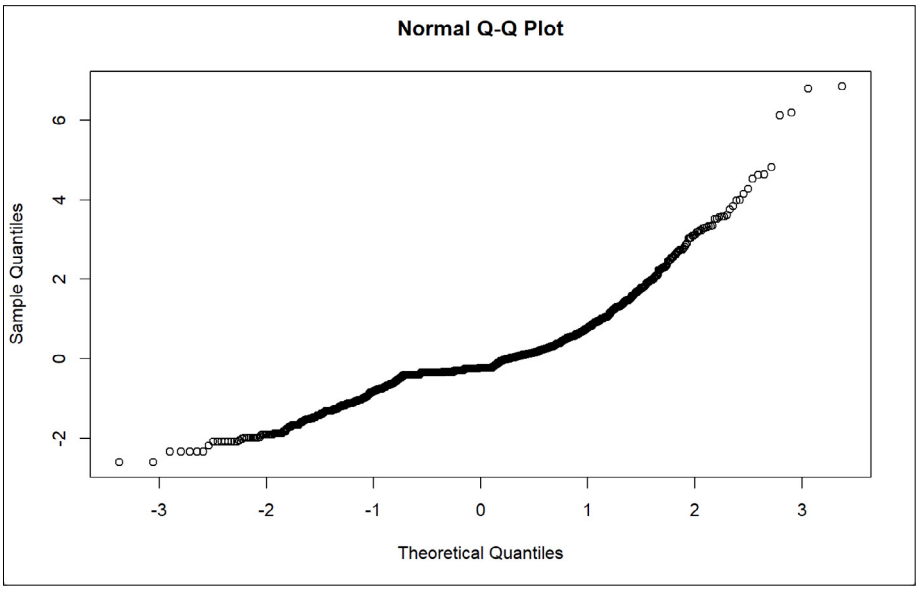
Table 3

Pairwise Comparison of PV Frequency in the Two-way ANOVA

Sub-discipline	N	AI		Human		Cohen's *d*
		Mean	**SD**	**Mean**	**SD**	
Archaeology	69	.71	0.88	3.06	2.80	1.13
Classics	81	.86	1.28	5.65	4.86	1.35
Comparative American Studies	74	1.05	1.74	5.20	4.17	1.30
English	100	.86	1.62	4.71	5.43	0.96
History	95	.88	1.18	5.12	4.55	1.27
Linguistics	107	.56	1.01	3.56	2.79	1.43
Other	48	1.08	2.31	6.79	5.09	1.45
Philosophy	105	.57	0.89	4.27	4.22	1.21

Figure 2

The Q-Q Plot of the Residuals of the ANOVA



Similarly, PV diversity (measured by CTTR) also differs significantly between sub-disciplines [F (7, 1342) = 7.046, $p < 0.001$, $\eta^2 = 0.035$], different authorship [F (1, 1342) = 970.510, $p < 0.001$, $\eta^2 = 0.420$], and the interaction between the two [F (7, 1342) = 4.503, $p < 0.001$, $\eta^2 = 0.023$]. The effect size of the differences between AI and humans is also larger than the difference between sub-disciplines, reaffirming the stark contrast. For disciplinary differences, linguistics and archaeology essays have significantly lower PV diversity than essays in comparative American studies (Figure 3). The interaction effects suggest that the differences in PV diversity between AI and humans were larger in some sub-disciplines than in others (see Table 4). For example, the difference of PV diversity in ChatGPT-authored and human-written archaeology essays was 0.59, while the difference in history essays was 0.89. This result reveals that within the arts and humanities, different sub-disciplines exhibit nuanced stylistic variations that require different levels of formality, with linguistics and archaeology being more formal and objective than other subjects like history, classics, and comparative American studies. Again, a residual check was performed to confirm the robustness of the ANOVA against the significant result in the normality check (Figure 4).

Figure 3

The Interaction between Authorship and Sub-disciplines in PV Diversity

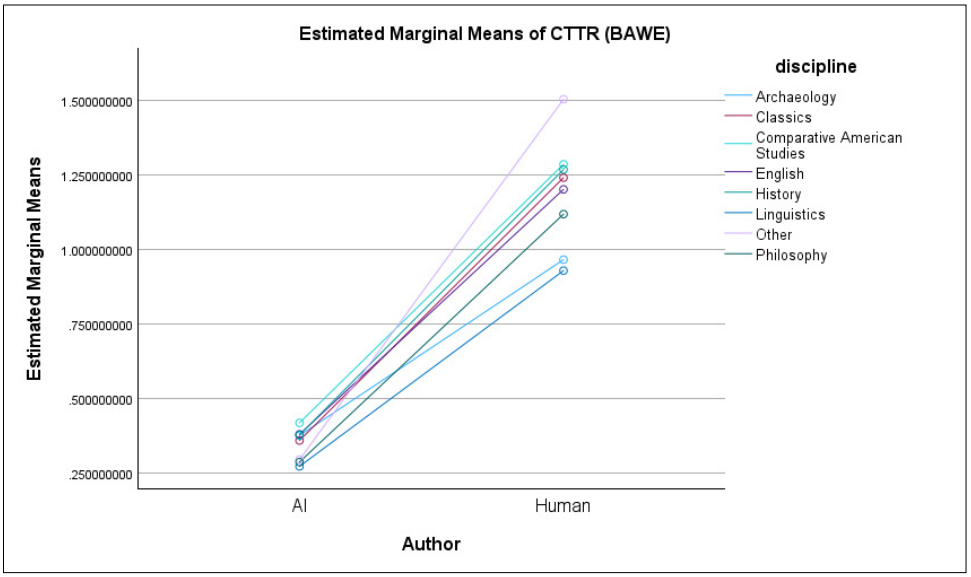
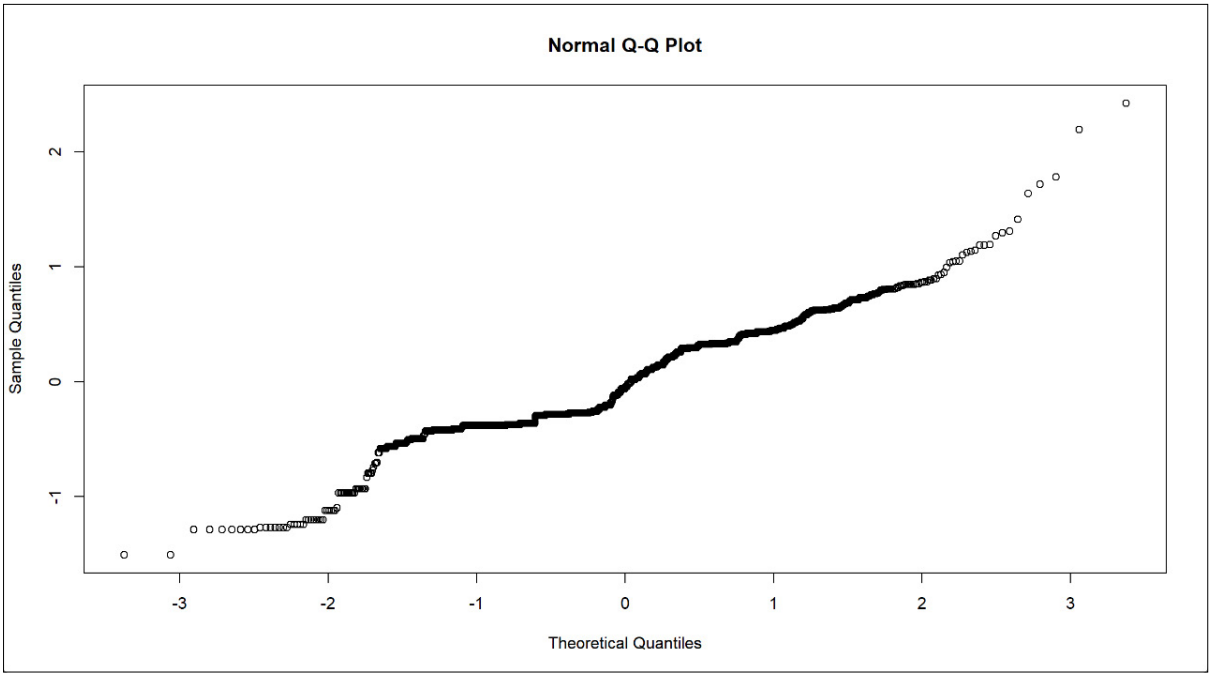


Table 4

Pairwise Comparison PV Diversity in the Two-way ANOVA

Sub-discipline	N	AI		Human		Cohen's *d*
		Mean	**SD**	**Mean**	**SD**	
Archaeology	69	0.38	0.41	0.97	0.55	1.21
Classics	81	0.36	0.41	1.24	0.58	1.75
Comparative American Studies	74	0.42	0.42	1.29	0.59	1.69
English	100	0.38	0.42	1.20	0.62	1.55
History	95	0.38	0.40	1.27	0.57	1.81
Linguistics	107	0.27	0.38	0.93	0.44	1.59
Other	48	0.30	0.40	1.50	0.61	2.34
Philosophy	105	0.29	0.38	1.12	0.48	1.92

Figure 4
The Q-Q Plot of the Residuals of the ANOVA



4.2 RQ2: Comparison of the most frequent PVs by human and ChatGPT

For RQ2a, we first compare the high-frequency PVs in BAWE and BAWE_GPT against the PV frequency in the academic register of COCA. Then for RQ2b, we perform a case analysis on the most frequent PVs used by humans and by ChatGPT in terms of their polysemy, meaning transparency, and collocational patterns.

Among the top 20 most frequent PVs that appear in BAWE and BAWE_GPT, eight overlapped across the corpora: "point out", "carry out", "set out", "pick up", "take on", "break down", "bring about", and "open up" (Table 5 & Table 6). However, their specific rankings differed considerably, with the human PV use aligning more closely with the actual frequency of PVs in the academic register in COCA. For example, "point out" and "carry out", which ranked top two in BAWE, were also the two most frequent PVs in the academic register in COCA. All the top 20 PVs in BAWE were among the top 50 most common PVs in the academic register in COCA, suggesting that PV use in BAWE was highly representative of the human academic writing style. However, in BAWE_GPT, the PV rankings differed markedly from those in the academic register in COCA, and the numbers did not show an ascending trend. Seven PVs did not appear in the top 150 most frequent PV list, suggesting a deviation from the typical expression patterns of human writers.

According to Table 5 and Table 6, the most common PV in BAWE and BAWE_GPT are "point out" and "open up" respectively. To analyze these two PVs in more depth, we extracted all the concordances containing these two PVs in both corpora, and examined their meanings, transparency, and collocations. After detailed comparisons, we found that human writers used PVs in a greater diversity of meanings. For both PVs, human writers used two different meanings of the same PV, while ChatGPT only used the most common meaning of the PV (See Tables 7 and 8). As for transparency, human writers can adeptly use the transparent, semi-transparent, and opaque meanings of PVs, whereas ChatGPT demonstrates its command only in the opaque meanings, which may be attributed to the PVs' high frequencies.

Table 5
The Top 20 Most Frequent PVs in BAWE

Rank	PV	Frequency	Liu (2011) academic ranking
1	point out	315	1
2	carry out	166	2
3	make up	97	5
4	set up	84	6
5	set out	75	18
6	pick up	60	10
7	go on	57	3
8	end up	53	14
8	take on	53	4
10	sum up	50	29
10	break down	50	17
12	find out	48	13
13	build up	45	26
14	give up	41	9
15	take up	40	12
16	take over	33	20
17	bring up	32	39
18	move on	31	49
19	bring about	30	8
19	open up	30	19

Table 6
The Top 20 Most Frequent PVs in BAWE_GPT

Rank	PV	Frequency	Liu (2011) academic ranking
1	open up	59	19
2	bring about	52	8
3	stand out	43	31
4	take on	36	4
5	carry out	25	2
6	put forth	21	n/a
7	break down	21	17
8	pass down	20	n/a
9	grow up	17	11
10	bring forth	14	n/a
11	point out	12	1
12	seek out	11	n/a
13	break away	9	n/a
14	slow down	8	51
15	fill in	8	83
16	set out	8	18
17	pick up	7	10
18	leave behind	6	n/a
19	carve out	5	n/a
20	cut off	5	42

Note: n/a means not in the high-frequency PV list in Liu (2011)

The most salient distinction between human writers and ChatGPT lies in the collocation patterns. In the concordances of "point out", human writers used 41 different collocations in 315 concordances, whereas ChatGPT used only three in 12 concordances (Table 7). If we normalize the count, in 100 concordances, humans would use 13 different collocates and ChatGPT would use 25. The relatively homogeneous collocates of human beings for "point out" may be due to the habitual use of "that" (151 times, see Appendix) for this high-frequency PV by students. In the case of "open up", although ChatGPT used it more often than human writers, human writers produced 24 distinct collocations in 30 concordances, while ChatGPT employed only 10 collocations across 59 concordances (Table 8). If we normalize the count, for every 100 concordances, humans would use 80 different collocates and ChatGPT would use 17 different ones. "Open up" is not a PV that students used regularly, so the collocates are more diverse and unpredictable. From these two cases, it could be inferred that, in general, human writers use more varied, flexible, and creative PV collocations, while ChatGPT uses more predictable, strongly-associated, and formulaic PV collocations. The detailed collocation lists can be found in Appendix.

Table 7
Comparison of Point Out in BAWE and BAWE_GPT

	Frequency	Meaning	Transparency	Number of Collocations
BAWE	314	Tell somebody something	Opaque	40
	1	Direct someone's gaze or attention towards, especially by extending one's finger	Transparent	1
BAWE_GPT	12	Tell somebody something	Opaque	3

Table 8
Comparison of Open Up in BAWE and BAWE_GPT

	Frequency	Meaning	Transparency	Number of Collocations
BAWE	28	Become present, available, or accessible	Opaque	25
	2	Become more communicative	Semi-transparent	1
BAWE_GPT	59	Become present, available, or accessible	Opaque	10

Overall, our findings reveal significant differences in the PV use between humans and AI. Humans displayed a more informal writing style with a personal touch, using many high-frequency PVs as well as creative and uncommon PVs. Human writers were also aware of the multiple meanings of PVs and used PVs with a wide range of collocations. ChatGPT had a more formal writing style, which contained fewer PVs. Moreover, its PV use was highly predictable, rigid, and homogenous in terms of collocations, with an impersonal style.

5 Discussion

5.1 RQ1: The frequency of PVs

Our study found that human writers consistently used more PVs with greater diversity than ChatGPT in academic essays. The most striking finding is probably that human writers used PVs five times more

frequently than ChatGPT. A ratio of five is notably high, as previous investigations on lexical diversity (Fredrick & Craven, 2025; Herbold et al., 2023; Mizumoto et al., 2024) have reported maximum ratio differences between ChatGPT and humans of less than two (Table 9). Similar ratios were reported in other key comparison metrics such as syntactic complexity, modals, and nominalizations. In the very few cases with a ratio difference larger than two, the raw frequencies of the measures were typically minimal, such as the frequency of epistemic markers (Herbold et al., 2023) and engagement markers (Jiang & Hyland, 2025b), indicating that the high ratios may be insignificant due to the small sample sizes (e.g. 0.06 versus 0). This suggests that PV may be among the most sensitive linguistic markers for distinguishing human and GenAI writing.

Table 9
Lexical Diversity of ChatGPT and Human Writing

Study	Linguistic measures	AI	Students
Fredrick and Craven (2025)	TTR	0.69	0.61
	Measure of textual lexical diversity (MTLD)	118	66.56
	Vocabulary diversity (Voc-D)	105	69.73
Herbold et al. (2023)	MTLD	108.91 (GPT4), 75.68 (GPT3)	95.72
Mizumoto et al. (2024)	MTLD	108.44	69.38
The current study	Normalized PV frequency	0.32	1.94

The significant difference in PV use between humans and AI can be attributed to several reasons. Firstly, because GenAI models generate texts by predicting the next word in a sequence of text and do this repeatedly until the output is complete, they tend to produce structurally simpler constructions. Complex verb phrases or clauses, which typically involve multiple morphosyntactic choices concerning tense, aspect, and mood (Huddleston & Pullum, 2005), tend to be more difficult to process than noun-related chunks. Additionally, GenAI models face particular difficulty in producing PVs because a proportion of them are separable (e.g. *put* it *down* VS *put down* it*). Under a next-word prediction mechanism, separable PVs are disfavored: the verb and its particle can be separated by intervening words, creating discontinuous dependencies that increase uncertainty for the generator. Moreover, due to intensive training, AI may adhere to academic writing conventions more strictly than university students. It can therefore be inferred that algorithm limitations and better undeterstanding of academic writing expectations both led to the underused PVs in AI writing.

5.2 RQ1: Disciplinary differences of PV use

Regarding the disciplinary differences in AI usage, archaeology and linguistics essays used significantly fewer PVs and had lower PV diversity than classics and comparative American studies. Compared with science and engineering, the humanities employ more verbs related to argument and persuasion (Hyland, 2008). In the disciplines involving human behaviour and interpretation, a slightly more conversational tone is acceptable and sometimes even preferred for clarity (Zuvaytova, 2025). Classics, history, and comparative American studies involve analysis of texts, history, and cultures, which requires considerable personal interpretation, and PVs may appear more frequently in descriptive and narrative texts in these disciplines. Comparatively, archaeology and linguistics are more closely aligned with science subjects (Kertész, 2024; Smith et al., 2012), which often involve experiments and fieldwork and value objectivity. Moreover, linguistic students may be more aware of the academic writing conventions in English, thereby minimizing PV use to project a more professional tone.

5.3 RQ2: The collocational patterns of PVs

Analysis of the most frequent PVs in both corpora revealed that humans used more diverse meanings, varied semantic transparencies, and a wider range of collocations. It is striking that in the case of "open up", human writers used "possibility" only four times but produced 24 other distinct collocates, including concrete nouns such as "door" and "church" (see Appendix). In the ChatGPT corpus, the collocates "avenue" appeared 22 times and "possibility" 18 times, reflecting an abstract, sophisticated, and vague writing style.

This result suggests that English L1 speakers have a nuanced understanding of the different uses of PVs in a variety of contexts, probably stemming from extensive real-world linguistic exposure. As a result, the collocations in human writing are more specific and contextually appropriate, which makes human writing highly dynamic, context-sensitive, and unpredictable. However, ChatGPT generates texts by predicting the next token based on the preceding context. Because it tends to select words based on probabilistic likelihood, the resulting PVs tend to reflect their most conventional meanings and favour the most frequent, decontextualized collocates. This process is statistical in nature and lacks the internal richness and individual variation rooted in the lived experience of humans and individual cognition (Fredrick & Craven, 2025). The above contrast illustrates that AI writing often lacks subjective details, contextual specificity, and rhetorical spontaneity, and produces verbose, tangential, or ambiguous texts (Attanasio et al., 2024). Therefore, ChatGPT rarely uses creative or uncommon collocations, which enhances its formulaicity and reduces idiosyncrasy.

The PV use reflects that ChatGPT writing is rigid and predictable, and rigorously adheres to academic writing style. These findings lend support to a recent concern of language homogenization (Bauer, 2025) due to the rise of GenAI. Overusing GenAI may impair the unique linguistic features of individual writers, leading to homogenous and impersonal writing discourse. A recent study reveals that AI-generated texts comprise at least 30% of online content, potentially reaching up to 40% (Spennemann, 2025). As GenAI becomes more widespread, this stylistic convergence could significantly affect linguistic diversity in both academic and public writing.

5.4 Digging deeper: Using PVs in academic writing

Although PVs are more common in informal discourses than formal discourses, we support a "cautious but not absolute approach" to using PV in academic writing (Zuvaytova, 2025, p. 3). The frequent and diverse use of PVs in human essays suggested that human essays have a more personal, creative, informal, and spontaneous writing style. This observation echoes the findings from Jiang and Hyland (2025b), who note that human writing is typically more informal and interactive, characterized by a higher use of engagement markers than ChatGPT essays. This personal style can probably be attributed to the inherent subjectivity in authors' academic "voices". For example, the high-frequency PVs "point out" and "find out" can express stance and opinion, which are essential rhetorical resources in academic writing of arts and humanities. Also, PVs may be a conventional and idiomatic way to narrate certain empirical actions, such as "set up" and "carry out", which are frequently used in English academic writing (D. Liu, 2011). Considering the alignment of PV use in BAWE and COCA, we can infer that the PV usage pattern in BAWE is broadly representative of human academic writing norms, and some PVs can be well integrated into academic writing.

This study raises an important question: should we "humanize" AI writing, or should humans move closer to the highly standardized academic prose produced by AI? Putting this question into the context of our study, should we train AI to use more PVs in its academic writing to "personify" the model, or should humans reduce their PV use in academic writing to achieve a more impersonal style? The answer may not be straightforward, as language is constantly evolving and stylistic appropriateness is

always context-specific. Chen M. (2013) found that British university students used fewer PVs than their American counterparts, reflecting stronger genre awareness, yet their PV frequency still surpassed that of ChatGPT. If British writers already used PV cautiously in academic writing and their PV frequency still outnumbered the PVs used by ChatGPT, this suggests that it is AI, not humans, that deviates from the current academic writing norm. While L2 learners can still use AI-generated writing as good samples for academic writing for their clear structure and cohesive argumentation, human writers can take pride in their creativity and lexical diversity, maintaining their personal writing style and using PVs where contextually suitable.

6 Conclusion

Overall, the current study made novel contributions to the burgeoning field of AI and human writing comparisons by investigating PVs, a unique lexical feature. By analyzing PVs in the parallel corpora, which are much larger than those in previous studies, the results revealed significant systematic differences in the usage of PVs between human writers and ChatGPT3.5. Humans use significantly more PVs in their academic writing than ChatGPT3.5, with more diverse meanings, more varied degrees of semantic transparency, and a wider range of collocations. Their writing seems to be more creative, informal, contextually specific, and linguistically flexible. There were also significant differences in the appearance of PVs across different sub-disciplines in the arts and humanities, suggesting varying levels of subjectivity or formality in the writing style. This study also introduced methodological innovations by employing a new approach to identifying PVs in corpora. Through dependency parsing and the extraction of PVs based on both dependency and POS tags, the identification process proved to be more efficient than in previous studies that relied solely on POS tagging. We recommend this method to researchers conducting future studies on PVs.

This study offers important implications for practitioners and educators in teaching the use of GenAI tools for writing and developing methods to differentiate human and GenAI writing. Firstly, this study identifies a lexical feature that shows a substantial difference between AI-generated and human writing. Teachers should adopt a contextual approach and provide students with explicit instructions on the PVs that can be used in academic communication, instead of discouraging students from using any PVs, as PVs are an important linguistic feature that retains the style of human authors. Second, teachers can integrate LLM tools like ChatGPT into the classroom and organize comparative activities, so that students can develop strong genre awareness and contextual understanding to judge whether PV use is appropriate for specific writing tasks. When they seek ChatGPT proofreading or feedback on their writing, they can exercise their discretion to discard or retain PVs in their writing, depending on the formality of the context. Third, for software engineers or AI scientists, beyond traditional lexical and syntactic features, PV frequency and diversity may serve as salient indicators to distinguish human and GenAI writing. These findings can provide valuable implications for the development of AI-detection software, potentially enhancing the accuracy of AI-writing detection and safeguarding academic integrity.

Despite its contributions, this study is not without limitations. First, this study is limited by its focus on ChatGPT3.5, an earlier-generational LLM. While this choice allows for a controlled examination of foundational AI writing patterns, subsequent model updates may exhibit different lexical behaviors. The findings should therefore be interpreted as model-specific rather than representative of all contemporary AI systems. Second, the corpora only cover topics in the arts and humanities, which do not fully represent all major disciplines, such as science and engineering. Therefore, while the findings offer valuable insights in this specific discipline, they may not be fully generalizable to other disciplines. Lastly, the human corpus consists mainly of students from English L1 backgrounds and some students with English L2 backgrounds. Such heterogeneity may lead to a potential confounder of the L1

background in the results. However, given that the majority of human writers are English L1 speakers and all English L2 speakers at least meet the language requirements for enrolling in British universities, their English proficiency can reasonably reflect the academic norms of proficient writers.

Notes

1. Readers may refer to https://www.coventry.ac.uk/research/research-directories/current-projects/2015/british-academic-written-english-corpus-bawe/search-the-bawe-corpus/ for further details.
2. Normalised frequencies (per million tokens) were calculated by dividing the raw frequency of each PV by the total number of tokens in the corresponding corpus (BAWE or BAWE_GPT) and multiplying by 1,000,000.

Acknowledgements

We would like to express our sincere gratitude to Wenyan XU, Jiabei XIAO, and Qianruo WANG for developing the *BAWE_GPT* corpus and for their meticulous work in verifying the dependency annotations.

Declaration of interest statement

The authors certify that they have NO affiliations with or involvement in any organization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Declaration of generative AI in scientific writing

During the preparation of this work, the authors used Grammarly/ChatGPT to improve the readability of the writing. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

Funding source

The development of the BAWE_GPT corpus was supported by the Xi'an Jiaotong-Liverpool University Summer Undergraduate Research Fund (SURF-2023-0133)

Appendix

The collocates of "point out" and "open up" in human and ChatGPT corpora

Table 10
BAWE – *open up*

Collocation	Count
possibilities	4
areas	2
Amoy	1
areas of observation	1
avenue	1
Church	1
culinary sites	1
debate	1
door	1
feminine	1
field	1
former	1
governmental organs	1
hinterland	1
it	1
market	1
objects	1
possibility	1
question	1
questions and possible ideas	1
secrets	1
the West	1
themselves	1
way	1
web	1
Total collocations	26

Table 11
BAWE_GPT – *open up*

Collocation	Count
avenue	22
possibility	18
space	7
opportunity	4
access	2
discussion	2
horizon	1
market	1
territories	1
question	1
Total collocations	10

Table 12
BAWE – *point out*

Collocation	Count
that	151
difference	4
fact	4
example	2
functions and patterns of use	2
nature	2
problem	2
ways	2
absence	1
areas	1
argument	1
awareness	1
color of the skin	1
competence	1
conclusion	1
confusion	1
correlates	1
criticism	1
differences	1
evils	1
harm	1
idea	1
idealism	1
information	1
issue	1
limitation	1
links	1
mismatch	1
outline	1
parallelism	1
parallels	1
passage	1
places	1
reasons	1
regularities	1
semantic classes	1
something	1
space	1
state	1
use	1
rock formation	1
Total collocations	41

Table 13

BAWE_GPT – *point out*

Collocation	Count
that	10
failure	1
difficulty	1
Total collocations	3

References

Alangari, M., Jaworska, S., & Laws, J. (2020). Who's afraid of phrasal verbs? The use of phrasal verbs in expert academic writing in the discipline of linguistics. *Journal of English for Academic Purposes*, *43*, 100814. https://doi.org/10.1016/j.jeap.2019.100814

Attanasio, M., Mazza, M., Le Donne, I., Masedu, F., Greco, M. P., & Valenti, M. (2024). Does ChatGPT have a typical or atypical theory of mind? *Frontiers in Psychology*, *15*, 1488172. https://doi.org/10.3389/fpsyg.2024.1488172

Badem, N., & Şimşek, T. (2021). A comparative corpus-based study on the use of phrasal verbs by Turkish EFL learners and L1 English speakers. *Advances in Language and Literary Studies*, *12*(6), 55. https://doi.org/10.7575/aiac.alls.v.12n.6.p.55

Bauer, N. F. (2025). Does ChatGPT increase language homogenization? In C. Vaih-Baur, V. Mathauer, E.-I. von Gamm, & D. Pietzcker (Eds.), *KI in Medien, Kommunikation und Marketing: Wirtschaftliche, gesellschaftliche und rechtliche Perspektiven* (pp. 11–31). Springer Fachmedien. https://doi.org/10.1007/978-3-658-46344-1_2

Cambridge Dictionary. (2025a, October 15). *Bubble over*. https://dictionary.cambridge.org/dictionary/english/bubble-over

Cambridge Dictionary. (2025b, October 15). *Churn out*. https://dictionary.cambridge.org/dictionary/english/churn-out

Carroll, J. B. (1964). *Language and Thought*. Prentice-Hall.

Casal, J. E., & Kessler, M. (2023). Can linguists distinguish between ChatGPT/AI and human writing?: A study of research ethics and academic publishing. *Research Methods in Applied Linguistics*, *2*(3), 100068. https://doi.org/10.1016/j.rmal.2023.100068

Chen, D., & Manning, C. (2014). A fast and accurate dependency parser using neural networks. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 740–750. https://doi.org/10.3115/v1/D14-1082

Chen, H. H.-J., Yang, C. T.-Y., Wei, I. F.-F., & Jiang, A. (2015). A corpus study on phrasal verb use in the academic writing of published authors, native English-speaking students, and Taiwanese EFL learners. *English Teaching & Learning*, *39*(4), 63–91.

Chen, M. (2013). Overuse or underuse? A comparative study on phrasal-verb use between British and American novice writers. *International Journal of Corpus Linguistics*, *18*(3), 418–442.

De Wilde, V. (2024). Can novice teachers detect AI-generated texts in EFL writing? *ELT Journal*, *78*(4), 414–422. https://doi.org/10.1093/elt/ccae031

Du, M. (2025). A corpus-based analysis of verb collocations in human and AI-generated IELTS writing. *Journal of Educational Technology and Innovation*, *7*(2). https://doi.org/10.61414/mzmmrq74

Fletcher, B. (2005). Register and phrasal verbs. *MED Magazine*, *33*(4), 212–224.

Fredrick, D. R., & Craven, L. (2025). Lexical diversity, syntactic complexity, and readability: A corpus-based analysis of ChatGPT and L2 student essays. *Frontiers in Education*, *10*, 1616935. https://doi.org/10.3389/feduc.2025.1616935

Gardner, D., & Davies, M. (2007). Pointing out frequent phrasal verbs: A corpus-based analysis. *TESOL Quarterly*, *41*(2), 339–359. https://doi.org/10.1002/j.1545-7249.2007.tb00062.x

Garnier, M., & Schmitt, N. (2015). The PHaVE List: A pedagogical list of phrasal verbs and their most frequent meaning senses. *Language Teaching Research*, *19*(6), 645–666. https://doi.org/10.1177/1362168814559798

Garnier, M., & Schmitt, N. (2016). Picking up polysemous phrasal verbs: How many do learners know and what facilitates this knowledge? *System*, *59*, 29–44. https://doi.org/10.1016/j.system.2016.04.004

Glass, G. V., Peckham, P. D., & Sanders, J. R. (1972). Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of Educational Research*, *42*(3), 237–288. https://doi.org/10.3102/00346543042003237

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

Heuboeck, A., Holmes, J., & Nesi, H. (2010). *The BAWE Corpus Manual.* https://www.coventry.ac.uk/globalassets/media/global/08-new-research-section/current-projects/bawemanual-v3.pdf

Huddleston, R., & Pullum, G. (2005). The Cambridge Grammar of the English Language. *Zeitschrift für Anglistik und Amerikanistik*, *53*(2), 193–194. https://doi.org/10.1515/zaa-2005-0209

Hyland, K. (2008). Genre and academic writing in the disciplines. *Language Teaching*, *41*(4), 543–562. https://doi.org/10.1017/S0261444808005235

Jiang, F. (Kevin), & Hyland, K. (2025a). Does ChatGPT argue like students? Bundles in argumentative essays. *Applied Linguistics*, *46*(3), 375–391. https://doi.org/10.1093/applin/amae052

Jiang, F. (Kevin), & Hyland, K. (2025b). Does ChatGPT write like a student? Engagement markers in argumentative essays. *Written Communication*, *42*(3), 463–492. https://doi.org/10.1177/07410883251328311

Jiang, F. (Kevin), & Hyland, K. (2025c). Metadiscursive nouns in academic argument: ChatGPT vs student practices. *Journal of English for Academic Purposes*, *75*, 101514. https://doi.org/10.1016/j.jeap.2025.101514

Jiang, F. (Kevin), & Hyland, K. (2025d). Rhetorical distinctions: Comparing metadiscourse in essays by ChatGPT and students. *English for Specific Purposes*, *79*, 17–29. https://doi.org/10.1016/j.esp.2025.03.001

Kertész, A. (2024). Relativism versus absolutism in linguistics. *Foundations of Science*, *29*(4), 1089–1120. https://doi.org/10.1007/s10699-023-09909-w

Kiativutikul, C., & Phoocharoensil, S. (2014). A corpus-based study of phrasal verbs: CARRY OUT, FIND OUT, and POINT OUT. *International Journal of Research Studies in Language Learning*, *3*(7). https://doi.org/10.5861/ijrsll.2014.820

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge University Press.

Li, B., Tan, Y. L., Wang, C., & Lowell, V. (2025). Two years of innovation: A systematic review of empirical generative AI research in language learning and teaching. *Computers and Education: Artificial Intelligence*, *9*, 100445. https://doi.org/10.1016/j.caeai.2025.100445

Li, W., Zhang, X., Niu, C., Jiang, Y., & Srihari, R. (2003). An expert lexicon approach to identifying English phrasal verbs. *Proceedings of the 41st Annual Meeting of the ACL*, 513–520.

Liao, Y., & Fukuya, Y. J. (2004). Avoidance of phrasal verbs: The case of Chinese learners of English. *Language Learning*, *54*(2), 193–226. https://doi.org/10.1111/j.1467-9922.2004.00254.x

Liu, D. (2011). The most frequently used English phrasal verbs in American and British English: A multicorpus examination. *TESOL Quarterly*, *45*(4), 661–688. https://doi.org/10.5054/tq.2011.247707

Liu, D., & Myers, D. (2018). The most-common phrasal verbs with their key meanings for spoken and academic written English: A corpus analysis. *Language Teaching Research*, *24*(3), 403–424. https://doi.org/10.1177/1362168818798384

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

Liu, Y., Zhang, Z., Zhang, W., Yue, S., Zhao, X., Cheng, X., Zhang, Y., & Hu, H. (2023). *ArguGPT: Evaluating, understanding and identifying argumentative essays generated by GPT models* (arXiv:2304.07666). arXiv. http://arxiv.org/abs/2304.07666

Lumley, T., Diehr, P., Emerson, S., & Chen, L. (2002). The importance of the normality assumption in large public health data sets. *Annual Review of Public Health*, *23*(Volume 23, 2002), 151–169. https://doi.org/10.1146/annurev.publhealth.23.100901.140546

Manning, C. D., Mihai Surdeanu, Bauer, J., Finkel, J., Bethard, S. J., & McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing toolkit. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 55–60.

Mizumoto, A., Yasuda, S., & Tamura, Y. (2024). Identifying ChatGPT-generated texts in EFL students' writing: Through comparative analysis of linguistic fingerprints. *Applied Corpus Linguistics*, *4*(3), 100106. https://doi.org/10.1016/j.acorp.2024.100106

Nakahara, R. (2025). Is data-driven learning with ai-generated example sentences effective for elementary-level learners? A case study of the synonyms "collect" and "gather." *Asian Journal of English Language Teaching*, *34*(1), 5–22. https://doi.org/10.65961/AJELT-2025-1-002

Nguyen, N. H., Huang, X., & Dang, T. N. Y. (2026). Lexical profile of ChatGPT-generated reading materials targeting EFL learners across the CEFR levels. *International Journal of TESOL Studies*, *8*(3), 113–133. https://doi.org/10.58304/ijts.260211

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.

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. https://doi.org/10.1073/pnas.2422455122

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

Ryoo, M.-L. (2013). A corpus-based study of the use of phrasal verbs in Korean EFL students' writing. *The Journal of Asia TEFL*, *10*(2), 63–89.

Scarfe, P., Watcham, K., Clarke, A., & Roesch, E. (2024). A real-world test of artificial intelligence infiltration of a university examinations system: A "Turing Test" case study. *PLOS ONE*, *19*(6), e0305354. https://doi.org/10.1371/journal.pone.0305354

Shatz, I. (2023). Assumption-checking rather than (just) testing: The importance of visualization and effect size in statistical diagnostics. *Behavior Research Methods*, *56*(2), 826–845. https://doi.org/10.3758/s13428-023-02072-x

Singh, S. (2026, March 25). ChatGPT Statistics (2026) – Active Users & Growth Data. *DemandSage*. https://www.demandsage.com/chatgpt-statistics/

Siyanova, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. One-word verbs. *IRAL - International Review of Applied Linguistics in Language Teaching*, *45*(2), 119–139. https://doi.org/10.1515/IRAL.2007.005

Smith, M. E., Feinman, G. M., Drennan, R. D., Earle, T., & Morris, I. (2012). Archaeology as a social science. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(20), 7617–7621. https://doi.org/10.1073/pnas.1201714109

Sonbul, S., Salam El-Dakhs, D. A., & Al-Otaibi, H. (2020). Productive versus receptive L2 knowledge of polysemous phrasal verbs: A comparison of determining factors. *System*, *95*, 102361. https://doi.org/10.1016/j.system.2020.102361

Spennemann, D. H. (2025). *Delving into: The quantification of Ai-generated content on the internet (synthetic data)* (arXiv:2504.08755). arXiv. https://doi.org/10.48550/arXiv.2504.08755

Stanford NLP Group. (2020). *POS Tagger FAQ*. CoreNLP. https://stanfordnlp.github.io/CoreNLP/tools_pos_tagger_faq.html

Teng, F. (2020). The effectiveness of group, pair and individual output tasks on learning phrasal verbs. *The Language Learning Journal*, *48*(2), 187–200. https://doi.org/10.1080/09571736.2017.1373841

Thao, T. Q., & Bon, P. V. (2023). English majors' difficulties in using phrasal verbs in academic writing. *VNU Journal of Science: Education Research*, *39*(4), 78–86. https://doi.org/10.25073/2588-1159/vnuer.4738

Tran, P. N. T., & Tran, Q. T. (2019). The use of phrasal verbs in English language research proposals by Vietnamese M.A. students. *VNU Journal of Foreign Studies*, *35*(4), 114–129. https://doi.org/10.25073/2525-2445/vnufs.4399

Tudino, G., & Qin, Y. (2024). A corpus-driven comparative analysis of AI in academic discourse: Investigating ChatGPT-generated academic texts in social sciences. *Lingua*, *312*, 103838. https://doi.org/10.1016/j.lingua.2024.103838

Waibel, B. (2007). *Phrasal verbs in learner English: A corpus-based study of German and Italian students* [Doctoral dissertation].

Wei, Y. (2021). Use of English phrasal verbs of Chinese students across proficiency levels: A corpus-based analysis. *International Journal of TESOL Studies*, *3*(4), 25–41. https://doi.org/10.46451/ijts.2021.12.03

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J., & Schmidt, D. C. (2023). *A prompt pattern catalog to enhance prompt engineering with ChatGPT* (arXiv:2302.11382). arXiv. https://doi.org/10.48550/arXiv.2302.11382

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.-L., & Tang, Y. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, *10*(5), 1122–1136. https://doi.org/10.1109/JAS.2023.123618

Zhang, M., & Crosthwaite, P. (2025). More human than human? Differences in lexis and collocation within academic essays produced by ChatGPT-3.5 and human L2 writers. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2024-0196

Zhang, X., & Wen, J. (2019). Exploring multiple constraints on second language development of English polysemous phrasal verbs. *Applied Psycholinguistics*, *40*(05), 1073–1101. https://doi.org/10.1017/S0142716419000146

Zuvaytova, S. (2025). Using phrasal verbs in academic writing. *Mental Enlightenment Scientific-Methodological Journal*, *6*(04), 1–6. https://doi.org/10.37547/mesmj-V6-I4-01

Siyang Zhou holds a PhD in Applied Linguistics from Oxford University and is a lecturer at the Center for Language Education, Hong Kong University of Science and Technology. She is mainly interested in formulaic language and issues related to study abroad. Her most recent publications on these topics have appeared in *Applied Linguistics Review*, *the International Journal of Applied Linguistics and Language Teaching (IRAL)* and *Australian Review of Applied Linguistics*.

Chen Chen is an Assistant Professor in the Department of Applied Linguistics at Xi'an Jiaotong-Liverpool University. Her research interests include corpus linguistics, L2 vocabulary, English Medium Instruction, and English for Academic Purposes. Her work has appeared in journals such as *Reading in a Foreign Language*, *Applied Corpus Linguistics,* and *International Journal of Applied Linguistics*.

Qingyang Wu is a graduate student in Applied Linguistics and Second Language Acquisition at the University of Oxford. Her research interests include language variation, language attitudes, perceptual dialectology, and L2 learning motivation. Her current research examines how learners' first-language (L1) accents influence language attitudes and motivation in second-language (L2) learning contexts. She is particularly interested in how dialects shape learners' experiences, attitudes, and engagement in English as a second language.