



The AI Detector Paradox: A Corpus-Based Analysis of Lexical Features and ZeroGPT Scores in Chinese L2 Undergraduate Writing

Huyue Liao^{1*}, Lubna Ali Mohammed²

^{1&2} Faculty of School of Education, Lincoln University College, Malaysia

DOI: doi.org/10.60072/ijeissah.2026.v4i04.004

*Corresponding author: Huyue Liao,
School of Education, Lincoln University
College, Malaysia

Corresponding author's e-mail:
285696442@qq.com

Received: 02 Aug 2026
Revised: 07 Oct 2026
Accepted: 23 Oct 2026

ABSTRACT

This study examines the "AI Detector Paradox": the association between L2 English lexical features and AI-detection scores. Applying standardized multiple regression to a baseline corpus of 200 verified human-authored undergraduate theses, we analyzed how specific lexical features are associated with ZeroGPT AI-likelihood scores. The results point to a counterintuitive "Penalty of Proficiency." Reliance on formulaic language is not associated with higher scores. Rather, higher lexical diversity (TTR) is significantly associated with higher AI scores ($\beta = 0.188$, $p < 0.01$), while dense academic vocabulary (AWL) paradoxically acts as a protective shield ($\beta = -0.284$, $p < 0.001$). These associations suggest that ZeroGPT scores may penalize signs of L2 lexical development, which could pressure L2 students to fall back on rigid, template-based writing to avoid false accusations.

Keywords: *Academic Integrity; AI Detection Bias; Corpus Analysis; Formulaic Language; L2 Learners*

Background

The use of large language models (LLMs) such as ChatGPT, Gemini, and Claude in academic writing has grown exponentially in recent years. In response to this irreversible trend, higher education institutions have widely adopted AI detectors, including GPTZero, Turnitin, Quillbot, and ZeroGPT, to defend academic integrity (Yan *et al.*, 2023). Yet a growing number of recent empirical studies have found that these algorithms depend too heavily on statistical probability indicators (namely, perplexity and burstiness). The outcome is inaccurate AIGC probability scores and, when advanced writing is assessed, even the devastating misclassification of human-written texts as AI-generated content (Elkhatat *et al.*, 2023; Walters, 2023). To avoid false accusations of plagiarism against students and the hegemony of any single AI detector, scholars have therefore turned to a makeshift solution: aggregating the results of several AI detectors to lower the risk of a student being misflagged (Hyatt *et al.*, 2025).

A broader spectrum of interest in this field focuses on the specific groups disproportionately affected by this algorithmic incompetence. A pioneering study conducted by Liang *et al.* (2023) demonstrated that currently available AI detectors pose a systemic bias against texts written by Non-Native English Speakers (NNES), while Alexander *et al.* (2023) resonated with this point and went further into the failing of AI benchmarks in assessing ESL or EFL writing in interdisciplinary research. As supporting evidence, Ibrahim (2023) investigated an ESL database, which reinforced that there exists strong inconsistency and vulnerability when AI classifiers scan ESL texts. Concerning L2 learners' language competence, they are more prone to choose words in a more conservative way and use highly structured syntax, increasing the likelihood of triggering the detector's "low information entropy" red lines. To some extent, this algorithmic flaw may contribute to the academic marginalization and stigmatization of L2 learners.

Objectives

Despite these empirical studies, current research leaves a critical empirical gap: major scholars still regard L2 writing features as an ambiguous monolithic entity, taking it superficially as a dichotomous question varying from AI tools. Rarely does any study adopt a corpus linguistics perspective to measure which specific linguistic features in L2 English writing are independently associated with detector scores through multivariate statistical modelling. To bridge the research gap, this study constructs a "Zero-AI L2 English human-written corpus", encompassing undergraduate theses, and applies standardized multiple regression. This approach is designed to estimate how each linguistic feature is associated with AIGC probability scores while holding the other features constant. To achieve the precise extraction of L2 linguistic features, this study draws upon the Academic Word List (AWL) and the University of Manchester's Academic Phrasebank to formulate the following three directional hypotheses for empirical testing:

- **H1 (Formulaic Language Reliance):** Formulaic language reliance, measured by the frequency of 4-word bundles, positively predicts the ZeroGPT AI score.
- **H2 (Lexical Formality):** Lexical formality, measured by academic vocabulary density (AWL%), positively predicts the ZeroGPT AI score.
- **H3 (Lexical Diversity):** Lexical diversity, as measured by the Type-Token Ratio (TTR%), is a positive predictor of the ZeroGPT AI score.

Literature Review

Most AI detectors work by asking how predictable a text is. DetectGPT flags writing whose word choices sit close to what a language model would itself expect (Mitchell *et al.*, 2023). The idea is neat, but it falters

in practice. When Weber-Wulff *et al.* (2023) tested a range of tools, they found them neither accurate nor reliable, and Walters (2023) saw performance vary widely across 16 detectors. Elkhatat *et al.* (2023) reported similar inconsistency, with some human texts labelled as machine-written. The problem is sharper for second-language writers. Liang *et al.* (2023) showed that detectors misjudge essays by non-native English speakers far more often than those by native speakers, a pattern echoed in ESL and EFL settings (Alexander *et al.*, 2023; Ibrahim, 2023). Pratama (2025) describes this as a trade-off between accuracy and fairness, while Hyatt *et al.* (2025) suggest pooling several detectors to reduce false positives. What is still unclear is why L2 texts get flagged. Research on learner writing shows that lexical diversity and academic vocabulary are key markers of L2 development (Maamuuja, 2021), yet few studies have tested which of these features actually pushes detection scores up. That is the question this study takes on.

Methods

Corpus Construction

The corpus for this research consists of 200 undergraduate theses by English majors (5000 words to 8000 words), randomly sampled from the Chinese National Paper Open Access Database. All manuscripts were originally submitted between 2018 and 2020, the critical period before off-the-shelf Generative AI models (e.g., ChatGPT, Gemini) became publicly available. Using these historical samples establishes a “Zero-AI baseline” with a very low likelihood of AI involvement, so that any AIGC probability score assigned by the detector used here (ZeroGPT) can be treated as a false positive. To ensure analytical validity, irrelevant components (e.g., cover page, Chinese abstract, table of content, tables and figures, reference, and appendices) were systematically removed from the original texts. The files were then converted to plain text (.txt) format so that features could be extracted precisely in AntConc 4.3.

Feature Extraction

This study employs AntConc 4.3 to analyze multi-level linguistic features. Three linguistic factors, each entered as a single observed variable, were selected to measure the L2 learners’ formulaic language dependency and lexical profiles: 1. Formulaic Language: to extract 4-word lexical bundles with high frequency and quantify the frequency as normalized frequency per 10,000 words (**Raw bundle count/Total token * 10,000**); 2. Lexical Formality (Academic Word List, AWL): to calculate the density of academic vocabulary appeared in every sample (**AWL token/Total token * 100%**); 3. Lexical Diversity (Type-Token Ratio, TTR%): to measure lexical abundance and vocabulary variation per text (**Type token/Total token * 100%**). Because raw TTR falls as text length rises, and thesis length ranged from 5,000 to 8,000 words, TTR is length-sensitive, and its results should be interpreted with caution.

Statistical Analysis

ZeroGPT’s overall weighted AIGC rate, the percentage of each thesis that ZeroGPT labels as AI-generated (0–100%), served as the dependent variable; one score was obtained per thesis. To test the hypothesized paths, all extracted linguistic independent variables were standardized as Z-scores, and the paths were fitted with multiple linear ordinary least squares (OLS) regression; no latent constructs or PLS estimation were involved, so the model is an observed-variable path analysis. This approach controls for the other linguistic variables in the model and so estimates the independent association of each linguistic feature with the ZeroGPT score, expressed as the standardized path coefficient (β).

Results and Discussion

Descriptive statistics

This study first carried out an overall descriptive analysis of the 200 Zero-AI baseline samples written by Chinese L2 English learners (see Table 1). The mean TTR is 23.69% (SD=2.51%). Set against the typical 30%-40% TTR threshold observed in proficient L1 academic writing (Maamujav, 2021), however, this figure reflects the relatively restricted and conservative lexical diversity typical of undergraduate non-native writers. The density of academic words (AWL%) averages 4.69% (SD=2.00%), which points to a strong adherence to academic formality. The normalized frequency of 4-word bundles averaged 62.58 per 10,000 words (SD=21.50). Notably, although the corpus meets absolute Zero-AI criteria, it produced an average AIGC misclassification rate of 7.16% (SD=5.19%), and the highest false-positive spike reached 27.66%—substantially above the 20% threshold used by some higher education institutions. These descriptive findings therefore show that ZeroGPT assigns non-zero AI scores to verified human L2 texts; without a comparable L1 corpus, however, they do not by themselves establish a bias specific to L2 writers.

Table 1. Descriptive statistics of linguistic features and AI detection bias (N = 200)

Variables	M	SD	Min	Max
Dependent Variable				
AI Detection Bias (ZeroGPT %)	7.16	5.19	0.00	27.66
Independent Variables				
Formulaic Language (4-word bundles)	62.58	21.50	20.38	129.13
Lexical Formality (AWL %)	4.69	2.00	0.06	11.67
Lexical Diversity (TTR %)	23.69	2.51	16.77	32.90

Source: Author construct based on the corpus analysis

Correlation & Path Analysis

Table 2 below reports the Pearson correlation matrix for the selected variables. All predictor intercorrelations are below the 0.70 threshold in absolute value (the highest is $r = 0.54$, between formulaic language and AWL), which rules out severe multicollinearity. Crucially, the zero-order correlations, positive for TTR ($r = 0.20$, $p < 0.01$) and negative for AWL ($r = -0.30$, $p < 0.01$) and Formulaic Language ($r = -0.21$, $p < 0.01$), anticipate the directional β coefficients later estimated in the regression analysis.

Table 2. Pearson correlation matrix of the studied variables

Variables	1	2	3	4
1. AI Detection Bias	1.000			
2. Formulaic Language	-0.21 **	1.000		
3. Lexical Formality (AWL)	-0.30 **	0.54 **	1.000	
4. Lexical Diversity (TTR)	0.20 **	-0.24 **	-0.04	1.000

Note: $p < 0.01$ **. N = 200.

Given the acceptable intercorrelations shown in Table 2, the standardized multiple regression was then estimated. The overall model was statistically significant ($F(3, 196) = 9.338$, $p < 0.001$), although it explained a modest share of the variance in ZeroGPT scores ($R^2 \approx 0.13$).

The path testing results point to an unexpected pattern (see figure 1). First, the data did not support Hypothesis 1 (H1), which proposed that reliance on formulaic language increases AI detection rates ($\beta = -0.010$, $t = -0.122$, $p = 0.903$). **After controlling for other lexical variables, mere reliance on structural templates is not associated with higher ZeroGPT scores.** The model did, however, suggest a "Penalty of Proficiency" pattern for the remaining features:

1. Lexical diversity (TTR%) is significantly and positively associated with the ZeroGPT score ($\beta = 0.188$, $t = 2.729$, $p = 0.007$), supporting H3. **Empirically, this indicates that the more unique words and varied expressions writers use, the greater their risk of being flagged as AI-generated by ZeroGPT.**
2. Academic vocabulary density (AWL%), by contrast, is significantly and negatively associated with the ZeroGPT score ($\beta = -0.284$, $t = -3.575$, $p < 0.001$), which runs counter to the direction hypothesized in H2. **The dense accumulation of formal academic vocabulary paradoxically appears to act as a digital shield, being linked with a lower AI detection probability.**

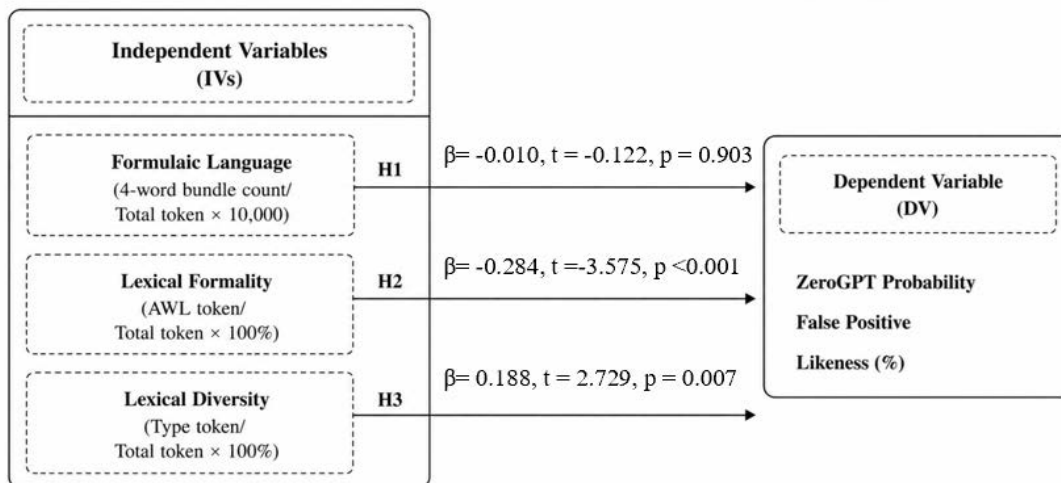


Figure 1. Standardized Regression (Path) Results

Limitations. This study used a single detector (ZeroGPT) and a single L2 corpus without an L1 comparison group, so the results describe associations with ZeroGPT scores rather than a bias specific to L2 writers. Raw TTR is sensitive to text length, so length-robust indices such as MTLTD are recommended for future work. Detector outputs may also change across software versions; future studies should report the scanning date, detector version, bundle frequency threshold, and the list of sampled theses to support replication.

Conclusion

In conclusion, this research shows that ZeroGPT assigns AI-likelihood scores to verified human-written L2 theses and that these scores are associated with specific lexical features. Besides finding no support for the assumption that formulaic 4-word bundles raise AI probability, the study points to a paradoxical "Penalty of Proficiency." While a dense accumulation of academic vocabulary acts as a protective shield, students' genuine developmental progress in lexical diversity (TTR) is associated with a higher risk of being misclassified as AI. These patterns suggest that ZeroGPT, and possibly similar detectors, should not be relied on alone to identify academic misconduct, as their scores may penalize students' language growth.

Higher education institutions must therefore critically assess the reliability and authority of these digital tools to prevent a pedagogical regression into rigid, formulaic writing practices.

Declarations

Ethics Approval and Consent to Participate: This study analysed publicly accessible undergraduate theses from the Chinese National Paper Open Access Database; no human participants were recruited, so ethics approval and informed consent were not required. The theses were analysed in anonymised form, no personal information was extracted or reported, and the texts were submitted to ZeroGPT solely to obtain detection scores.

AI Assistance: During the drafting and revision process, the authors employed a writing-assistance tool to enhance the manuscript's language and readability.

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

Funding: The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Authors' contributions: Both authors contributed to the study conception, research design, data analysis, manuscript drafting, and revision, and approved the final manuscript.

Acknowledgement: The authors express sincere appreciation to the scholars and researchers whose work provided the foundation for this study.

References

- Alexander, K., Savvidou, C., & Alexander, C. (2023). Who wrote this essay? Detecting AI-generated writing in second language education in higher education. *Teaching English with Technology*, 23(2), 25–43. <https://doi.org/10.56297/BUKA4060/XHLD5365>
- Elkhatat, A. M., Elsaid, K., & Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *International Journal for Educational Integrity*, 19(1), 17. <https://doi.org/10.1007/s40979-023-00140-5>
- Hyatt, J.-P. K., Bienenstock, E. J., Firetto, C. M., Woods, E. R., & Comus, R. C. (2025). Using aggregated AI detector outcomes to eliminate false positives in STEM-student writing. *Advances in Physiology Education*, 49(3), 486–495. <https://doi.org/10.1152/advan.00235.2024>
- Ibrahim, K. (2023). Using AI-based detectors to control AI-assisted plagiarism in ESL writing: “The Terminator Versus the Machines”. *Language Testing in Asia*, 13(1), 46. <https://doi.org/10.1186/s40468-023-00260-2>
- Liang, W., Yuksekgonul, M., Mao, Y., Wu, E., & Zou, J. (2023). GPT detectors are biased against non-native English writers. *arXiv preprint arXiv:2304.02819*.
- Maamuujav, U. (2021). Examining lexical features and academic vocabulary use in adolescent L2 students' text-based analytical essays. *Assessing Writing*, 49, 100540. <https://doi.org/10.1016/j.asw.2021.100540>

- Mitchell, E., Lee, Y., Khazatsky, A., Manning, C. D., & Finn, C. (2023). DetectGPT: Zero-shot machine-generated text detection using probability curvature. In *International Conference on Machine Learning* (pp. 24950–24962). PMLR. <https://proceedings.mlr.press/v202/mitchell23a.html>
- Pratama, A. R. (2025). The accuracy-bias trade-offs in AI text detection tools and their impact on fairness in scholarly publication. *PeerJ Computer Science*, 11, e2953. <https://doi.org/10.7717/peerj-cs.2953>
- Walters, W. H. (2023). The effectiveness of software designed to detect AI-generated writing: A comparison of 16 AI text detectors. *Open Information Science*, 7(1), 20220158. <https://doi.org/10.1515/opis-2022-0158>
- Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S., Foltýnek, T., Guerrero-Dib, J., Popoola, O., Šigut, P., & Waddington, L. (2023). Testing of detection tools for AI-generated text. *International Journal for Educational Integrity*, 19(1), 26. <https://doi.org/10.1007/s40979-023-00146-z>
- Yan, D., Fauss, M., Hao, J., & Cui, W. (2023). Detection of AI-generated essays in writing assessments. *Psychological Test and Assessment Modeling*, 65(1), 125–144.