

Lexical Richness and Academic Writing Performance in a Thai University Academic English Test

Zixuan Hu

Chawin Srisawat

*Kornwipa Poonpon**

Khon Kaen University, Thailand

*Corresponding author's email: korpul@kku.ac.th

Received August 2, 2026; revised August 29, 2026; accepted August 30, 2026

Abstract

Although lexical richness is an important feature of L2 writing quality, research on its predictive ability in university entrance tests, particularly in the Thai EFL context, is limited. Following Read's (2000) framework, this study analyzes the relationship between lexical richness and academic writing performance in an academic English test at a Thai university. The study aims to (1) compare lexical richness characteristics across proficiency levels; (2) examine their relationship with writing scores; and (3) determine their predictive power independent of text length. A corpus of 103 argumentative essays was analyzed using TAALED and TAALES. The essays were classified into two groups based on university admissions criteria. The Meets Standard Group outperformed the Below Standard Group on text length ($d = 1.54$), lexical diversity as measured by HD-D ($d = 0.76$) and MATTR ($d = 0.75$), and bigram frequency ($d = 0.47$), but not on the Maas index, lexical density, word frequency (KF and BNC written), word range, and AWL proportion. Text length correlated most strongly with scores, followed by HD-D, MATTR, and bigram frequency, while other indicators did not reach significance. Regression analysis showed that, when text length was controlled, the model explained 64.8% of the variance, with text length as the dominant predictor ($\beta = .691$). Lexical diversity (HD-D) retained a small but significant independent effect ($\beta = .174$, $p = .010$), whereas lexical sophistication (bigram frequency, BNC written frequency, AWL proportion) and lexical density did not. These findings provide empirical evidence from the Thai EFL context for vocabulary instruction and rubric design.

Keywords: lexical richness, lexical diversity, Thai EFL, TAALED, TAALES

Academic writing demands precise and concise expression, grounded in vocabulary knowledge (Coxhead, 2012; Hyland, 2003). Lexical richness is a key measure of L2 vocabulary in writing, encompassing lexical variation, sophistication, and density, and can predict writing quality (Engber, 1995; Laufer & Nation, 1995; Read, 2000).

Findings on the relationship between lexical richness and writing performance have been mixed across indices, tasks, and learner populations. Although multiple studies report a positive correlation between lexical diversity and writing performance, a few report weaker or insignificant results (Engber, 1995; Zhang, 2022; Zhang & Ren, 2018). These differences may stem from methodological choices, particularly the measures used. One possible explanation is the use of the length-sensitive TTR, which typically decreases as text length increases (McCarthy & Jarvis, 2007; Torruella & Capsada, 2013). Furthermore, many studies did not control text length as a potential confounding variable; associations attributed to lexical richness may partly reflect differences in text productivity. These methodological issues indicate a need for further studies that use length-robust indices and explicitly address the role of text length in the relationship between lexical richness and writing quality.

A limited number of studies have explored lexical richness in academic English writing by Thai EFL learners. Research has identified limited knowledge of English grammar and vocabulary as a major source of writing errors among Thai EFL students (Sermsook et al., 2017), suggesting that vocabulary instruction may deserve greater attention in the Thai EFL context. Moreover, findings from other Asian EFL contexts may not be directly generalizable to Thai learners.

Against this backdrop, the Khon Kaen University Academic English Language Test (KKU-AELT) serves as the official English entrance exam for master's and doctoral programs. In the context of the KKU-AELT (including this study), "academic writing" refers primarily to the test's institutional purpose as a graduate admissions measure rather than the disciplinary or linguistic register of the essays themselves, as the writing task prompt is a general-topic argumentative task. The current version of the KKU-AELT writing scoring criteria includes lexical richness indicators. Previous research found that, while academic vocabulary was positively correlated with higher scores, lexical density and diversity (as measured by TTR) were negatively correlated with scores (Srisawat & Poonpon, 2023). This counterintuitive finding may reflect TTR's sensitivity to text length rather than a genuine relationship between lexical diversity and writing performance. Moreover, previous research examined lexical sophistication only through the proportion of academic words, while other prominent measures of lexical sophistication, such as word range and n-gram frequency, should also be considered (Kyle & Crossley, 2015).

These research gaps point to three needs in studies on lexical richness. First, there is a need for validated, text-length-robust alternatives to TTR. Second, in addition to academic vocabulary, the measurement of lexical sophistication should consider a broader range of indicators, such as n-gram frequency and word range. Third, it is worth determining whether the observed relationships are confounded by text length or reflect genuine effects on writing performance. The aims of this study are therefore threefold: (1) to compare the characteristics of lexical richness in the academic writing of KCU-AELT test takers across different proficiency levels; (2) to examine the relationship between lexical richness indicators and academic writing scores; and (3) to determine the extent to which lexical richness indicators predict academic writing proficiency independent of text length.

The significance of this study is threefold. First, it contributes to the limited empirical research on vocabulary richness in Thai EFL and provides evidence on the relationship between vocabulary use and writing performance on a locally developed university entrance exam. Second, methodologically, it demonstrates the importance of comparing regression models that include and exclude text length as a control variable, providing a model for future research. Third, the findings provide practical implications for the KCU-AELT writing rubric and vocabulary instruction in Thai higher education.

Literature Review

Theoretical Framework: Read's (2000) Multidimensional Model of Lexical Richness

For measuring lexical richness, Read's (2000) four-dimensional model is widely used in L2 writing research and comprises variation, sophistication, density, and lexical error. This study selected the first three dimensions and excluded lexical errors because of documented challenges in identifying them reliably, including low inter-rater reliability and the difficulty of distinguishing them from syntactic or spelling errors (Hasko, 2022).

Lexical Richness in L2 Writing Research

Previous studies have broadly supported a positive relationship between lexical richness and writing quality. Engber (1995) found that lexical diversity and formal accuracy markedly predicted essay ratings, while Laufer and Nation (1995) indicated that more developed lexical knowledge was associated with higher writing scores. Lexical density is also related to writing quality, though its effect can vary by L1 background (Jarvis, 2002). In East Asian EFL contexts, Zhang and Ren (2018) found

that high-scoring TEM 4 essays showed higher lexical diversity and sophistication with fewer errors.

However, the relationship is not consistently positive in all contexts and for all measures. For example, Zhang (2022) found that lexical density and diversity did not significantly predict L2 writing quality, suggesting that findings can vary by task, methodology, and learner proficiency. In a study of the KKU-AELT, Srisawat and Poonpon (2023) found positive correlations for academic vocabulary but negative correlations for TTR-based diversity and density, which they ascribed to TTR's length sensitivity. That variability shows the importance of methodological choices, especially index selection.

Methodological Advances in Measuring Lexical Richness

Lexical variation measurement has been a core methodological issue in lexical richness studies. Traditionally, lexical diversity has been calculated using type-token ratio (TTR), the ratio of distinct words to total words. However, the use of TTR has been criticized for its sensitivity to text length: as a text becomes longer, its TTR decreases, potentially distorting the results (McCarthy & Jarvis, 2007; Torruella & Capsada, 2013), which complicates comparisons across texts. More recent methods have sought to address this limitation through alternative estimates of lexical richness. McCarthy and Jarvis (2007, 2010), for instance, proposed HD-D as a measure of lexical diversity that estimates the expected number of distinct types in random samples of 42 words using a hypergeometric distribution, whereas MATTR, developed by Covington and McFall (2010), calculates the mean TTR by measuring overlapping segments. Others, such as Zenker and Kyle (2021), examined the Maas index, which modifies TTR to reduce its length dependence. Their analyses found that MATTR, MTLD, and HD-D estimate richness more effectively than conventional TTR. However, stability can vary: MATTR performs consistently across text lengths, whereas MTLD and HD-D require a minimum length for stable estimates (McCarthy & Jarvis, 2010; Torruella & Capsada, 2013; Zenker & Kyle, 2021).

Kyle and Crossley (2015) extended this methodological work to lexical sophistication. Their framework combines word frequency, word range, academic language, and n-gram frequency in TAALES to measure the relative sophistication of a writing sample; this approach has since been validated by studies of vocabulary use and writing ability (Kyle & Crossley, 2015; Kyle et al., 2018). These new methods are crucial to the current study, enabling vocabulary use to be analyzed with a broader set of lexical indices while addressing differences in text length.

Research Gaps and the Present Study

Despite valuable previous contributions, research gaps remain. First, few studies have examined lexical richness in academic writing in the Thai context. Although Srisawat and Poonpon (2023) examined lexical richness in the KKU-AELT corpus, TTR was their only measure of lexical diversity, and the proportion of academic words was their only measure of lexical sophistication, excluding other important factors such as n-gram frequency and word range (Kyle & Crossley, 2015). Second, to our knowledge, no previous Thai EFL studies have examined the role of text length when assessing the relationship between lexical richness and writing scores. As shown by Zenker and Kyle (2021) and others, text length substantially affects lexical diversity estimation; unless effectively controlled, it may yield misleading conclusions. To address these gaps, the present study employed a corpus-based design with multiple length-robust lexical indices and regression models controlling for text length, as detailed in the Methods section.

Methods

Research Design

The study used a quantitative, corpus-based research design to investigate the role of lexical richness in writing performance in the KKU-AELT context.

Data Source

The data came from the KKU-AELT writing test developed by the Language Testing Center at Khon Kaen University, Thailand. Test takers wrote an argumentative essay in response to the following statement: “In today’s work environment, multitasking is a more valuable skill than being an expert in one area. Do you agree or disagree with this statement? What is your opinion? Why do you think so?” All test takers responded to the same prompt within one hour.

Each essay was scored independently by at least two trained raters using the KKU-AELT analytic writing rubric (Srisawat & Poonpon, 2023; Thongyoi & Poonpon, 2019). All raters were experienced English language instructors with backgrounds in TESOL or applied linguistics and had been trained to use the rubric before scoring. The rubric assesses writing performance across three dimensions: Topic Development, Organization, and Language Use. For Language Use, the descriptors explicitly reward “syntactic variety, appropriate word choice, and idiomaticity” at the highest levels, while lower bands are characterized by “inappropriate choice of

words or word forms” and “an accumulation of errors” (Thongyoi & Poonpon, 2019). If the two scores differed by more than 20 points, a third rater scored the essay, and the two closest scores were used. Inter-rater reliability, calculated as the Pearson correlation between the two raters’ total scores, was $r = .86$ (ICC = .84), indicating acceptable consistency (Koo & Li, 2016). The final score for each essay was the average of the two raters’ scores, recorded on a 0–100 scale, which corresponds to the seven-band KKU-AELT assessment scale (Bands 1–7).

According to the official English proficiency requirements for admission to the master’s program at Khon Kaen University (Announcement No. 76/2023), participants were divided into two groups: the Meets Standard Group, who met the requirement (Bands 4–7; $n = 39$), and the Below Standard Group, who did not meet the requirement (Bands 1–3; $n = 64$).

All essays were manually transcribed from handwritten scripts into plain text, with spelling and grammatical errors preserved to reflect the test-takers’ actual vocabulary production. A total of 117 essays were initially collected. Among these, 14 essays (12.0%) contained fewer than 50 tokens and were excluded to ensure the stability of lexical measures. The 50-token threshold was adopted because TAALED presumes a minimum text length of 50 words for stable index calculation (see Kyle et al., 2021; McCarthy & Jarvis, 2010). Although Zenker and Kyle (2021) recommend higher minimums for certain indices such as MTLT, the present study did not employ MTLT; the indices used (HD-D, MATTR, and the Maas index) are commonly used with 50 tokens and above. All excluded essays were from the Below Standard group (Bands 1–2), indicating that the exclusion disproportionately removed lower-scoring scripts; this limitation is addressed in the Limitations section. After data cleaning, 103 valid essays remained for analysis.

Although the two groups differed in size ($n = 64$ vs. $n = 39$), this imbalance represented the actual distribution of test-taker performance in the cohort and was addressed by using Welch’s t test for unequal variances where necessary (Field, 2018). To evaluate whether the sample size was sufficient to detect the observed effects, a post hoc power analysis was conducted using G*Power 3.1 (Faul et al., 2009). The results presented in the Results section confirmed that the sample was adequately powered to detect the primary effects.

Measurement of Lexical Richness

Lexical richness is a multidimensional construct comprising lexical variation, lexical sophistication, and lexical density (Read, 2000). This study applied two natural language processing tools—TAALED (Kyle et al., 2021) and TAALES (Kyle &

Crossley, 2015; Kyle et al., 2018)—to measure lexical richness; both tools have been validated in L2 writing research (Kyle et al., 2021; Kyle & Crossley, 2015; Kyle et al., 2018). The specific metrics for each dimension are presented here.

Lexical variation (sometimes referred to as lexical diversity) refers to the range of different words used in a text. In this study, it was measured using three length-robust indicators of lexical variation from the TAALED tool (Kyle et al., 2021) that are less affected by text length: MATTR (mattr50_aw), HD-D (hdd42_aw), and the Maas index (maas_ttr_aw). Traditional TTR and its variants were excluded because previous research has amply demonstrated their susceptibility to text length effects (Torruella & Capsada, 2013; Zenker & Kyle, 2021). In addition, the MTLT and MTLT-W metrics require approximately 100 tokens to produce stable estimates (Zenker & Kyle, 2021); as some texts in the corpus did not meet this threshold, these metrics were also excluded. This consideration is especially important for timed writing tasks, which typically produce short texts.

Lexical sophistication, according to Read (2000), refers to the use of advanced, less frequent, more specialized vocabulary. In this study, this dimension was operationalized using multiple TAALES indices (Kyle & Crossley, 2015; Kyle et al., 2018): word frequency (KF and BNC Written corpora), word range (BNC_Written_Range_AW), AWL proportion (All_AWL_Normed), and bigram frequency (BNC_Written_Bigram_Freq_Normed_Log). These indices jointly reflect different facets of lexical sophistication: AWL proportion reflects academic vocabulary use, while word frequency and bigram frequency reflect the conventionality and fluency of vocabulary and phrasal choices (Kyle & Crossley, 2015).

Lexical density refers to the proportion of lexical words (i.e., nouns, verbs, adjectives, and adverbs) relative to the total number of tokens (Laufer & Nation, 1995). In this study, it was automatically retrieved from TAALED (Kyle et al., 2021).

Furthermore, *text length*, which refers to the total number of tokens in each essay, served as both a descriptive variable and a control variable in the regression analysis. The TAALED tool (Kyle et al., 2021) automatically calculated text length and reported it as the metric basic_ntokens.

Data Analysis

All statistical analyses were conducted using SPSS (Version 26), and post hoc power analyses for the *t* tests were computed using G*Power 3.1 (Faul et al., 2009). The variables were first summarized using means, standard deviations, minima, and maxima to describe the overall characteristics and distribution of the data.

To address Objective 1, independent-samples *t* tests were performed to compare the ‘Below Standard’ and ‘Meets Standard’ groups on all lexical richness indicators. Before the *t* tests, Levene’s test for equality of variances was conducted to assess the homogeneity assumption. When the assumption was violated ($p < .05$), Welch’s *t* test with adjusted degrees of freedom was used; otherwise, pooled-variance *t* tests were employed. Effect sizes were measured by Cohen’s *d*.

To address Objective 2, Pearson’s correlation coefficients were calculated to examine the relationships between the lexical richness indicators and the writing scores. The Benjamini–Hochberg false discovery rate (FDR) procedure was applied to the *p* values from the *t* tests and correlation analyses (Benjamini & Hochberg, 1995). This procedure controlled the elevated risk of Type I error due to multiple statistical comparisons. The FDR threshold was 0.05. Both original and FDR-adjusted *p*-values are reported for transparency, and significance decisions are based on the FDR-adjusted values.

Regarding Objective 3, the study used two multiple linear regression models. Predictor selection was theory-based, following Read’s (2000) three-dimensional framework. One or more indicators were chosen to represent each dimension. To prevent multicollinearity, only one indicator was retained when two or more indicators showed high intercorrelations ($r \geq .80$), following established practices in lexical diversity research (McCarthy & Jarvis, 2010). Model 1 included the selected lexical indicators as predictors of writing scores. Model 2 added text length as a control variable to isolate the unique contribution of lexical features beyond textual productivity. A sensitivity model replaced the primary indicator in Model 2 with an alternative indicator to evaluate the robustness of the findings.

Ethical Considerations

This study received ethical approval from Khon Kaen University, Thailand (Ref. No. HE683159). Data were collected after approval and analyzed anonymously.

Results

Descriptive Statistics

Table 1 presents the means, standard deviations, minima, and maxima of all lexical richness indicators and essay length. These statistics summarize vocabulary distributions in KKU-AELT essays and identify the indicators with the greatest variation across the sample.

Table 1

Descriptive Statistics for the Full Sample (n = 103)

Indicator	<i>n</i>	Min.	Max.	<i>M</i>	<i>SD</i>
Text length	103	51	436	189.18	76.25
Lexical density	103	0.41	0.69	0.51	0.05
The Maas index	103	0.03	0.1	0.06	0.01
MATTR	103	0.57	0.89	0.75	0.05
HD-D	103	0.59	0.87	0.77	0.05
Word frequency (KF)	103	2.6	3.18	2.99	0.1
Word frequency (BNC)	103	-0.4	0.14	-0.09	0.12
Word range (BNC)	103	65.49	85.34	77.02	3.8
Bigram frequency	103	-2.11	-1.34	-1.6	0.12
AWL proportion	103	0.01	0.15	0.09	0.03

Note. All indicators were computed using TAALED and TAALES.

As shown in Table 1, mean text length was 189.18 tokens ($SD = 76.25$), with a range of 51–436 tokens. Lexical density averaged 0.51 ($SD = 0.05$). HD-D ranged from 0.59 to 0.87 ($M = 0.77$, $SD = 0.05$), indicating relatively limited variation in this measure. The mean AWL proportion was 9%, possibly because the general-topic prompt required limited disciplinary vocabulary.

The Characteristics of Lexical Richness in Academic Writing by Test Takers Across Different Proficiency Levels

Table 2 presents independent-samples *t* test results comparing lexical richness between the Below Standard and Meets Standard groups.

Table 2

Comparison of Lexical Richness Between Below Standard and Meets Standard Groups

Indicator	Below std. <i>M (SD)</i>	Meets std. <i>M (SD)</i>	<i>t</i>	<i>df</i>	<i>p</i>	Cohen's <i>d</i>	95% CI
Text length	153.53 (60.76)	247.67 (61.92)	-7.57	101	< .001**	1.54	-118.80, -69.47
Lexical density	0.52 (0.05)	0.50 (0.04)	1.81	97.81	.073	0.34	-0.001, 0.03

Table 2

Comparison of Lexical Richness Between Below Standard and Meets Standard Groups (Cont.)

Indicator	Below std. <i>M (SD)</i>	Meets std. <i>M (SD)</i>	<i>t</i>	<i>df</i>	<i>p</i>	Cohen's <i>d</i>	95% CI
Maas index	0.06 (0.01)	0.06 (0.01)	2.35	98.14	.021	0.41	0.001, 0.009
MATTR	0.74 (0.05)	0.77 (0.03)	-3.67	101	<.001**	0.75	-0.05, -0.015
HD-D	0.76 (0.05)	0.79 (0.02)	-4.40	93.03	<.001**	0.76	-0.047, -0.018
Word frequency (KF)	2.98 (0.11)	3.00 (0.08)	-1.00	98.71	.319	0.19	-0.057, 0.019
Word frequency (BNC)	-0.10 (0.12)	-0.08 (0.10)	-0.55	101	.582	0.11	-0.060, 0.034
Word range (BNC)	77.03 (4.31)	77.00 (2.85)	0.03	101	.974	0.01	-1.52, 1.57
Bigram frequency	-1.62 (0.13)	-1.56 (0.09)	-2.31	101	.023*	0.47	-0.10, -0.010
AWL proportion	0.09 (0.03)	0.09 (0.03)	0.35	101	.729	0.07	-0.01, 0.015

Note. std. = standard. Below-standard group: $n = 64$; meets-standard group: $n = 39$. Confidence intervals are for unadjusted mean differences. p -values reported are unadjusted. Asterisks indicate significance after Benjamini–Hochberg false discovery rate adjustment. * $p < .05$. ** $p < .001$.

Before FDR correction, differences were observed for five indicators: text length, the Maas index, MATTR, HD-D, and bigram frequency. After applying the Benjamini–Hochberg FDR correction (Benjamini & Hochberg, 1995), text length ($d = 1.54$, $FDR-p < .001$), MATTR ($d = 0.75$, $FDR-p < .001$), HD-D ($d = 0.76$, $FDR-p < .001$), and bigram frequency ($d = 0.47$, $FDR-p = .046$) remained significant. The Maas index, however, did not survive FDR correction (original $p = .021$; FDR -adjusted $p = .053$) and is therefore interpreted with caution. No significant differences were found for lexical density, word frequency (KF and BNC written), word range, or AWL proportion (all $FDR-p > .05$).

Post hoc power analysis using G*Power 3.1 (Faul et al., 2009) was conducted for the significant t test findings (two-tailed, $\alpha = .05$). Based on the observed effect sizes, achieved power was .99 for text length ($d = 1.54$, $n_1 = 64$, $n_2 = 39$), .96 for lexical diversity (MATTR: $d = 0.75$, $n_1 = 64$, $n_2 = 39$; HD-D: $d = 0.76$, $n_1 = 64$, $n_2 = 39$), and .63 for bigram frequency ($d = 0.47$, $n_1 = 64$, $n_2 = 39$). The power values for text length and lexical diversity surpassed the conventional threshold of .80 (Cohen, 1988), suggesting that the sample had sufficient power to detect these primary

effects. The lower power for bigram frequency suggests that this effect should be interpreted cautiously.

These findings suggest that, in the KKU-AELT context, higher-proficiency students wrote longer texts, showed higher lexical variation, and used more frequent bigram sequences than lower-proficiency students. However, no significant differences were found for lexical density, word frequency, word range, or AWL proportion.

The Relationship Between Lexical Richness Indicators and Academic Writing Scores

Pearson correlations were calculated to examine the relationships between the lexical richness measures and students' academic writing scores. Table 3 presents the results.

Table 3

Pearson Correlations Between Lexical Richness Indicators and Writing Scores (N = 103)

Indicator	<i>r</i>	95% CI for <i>r</i>	<i>p</i>
Text length	.775**	0.69, 0.84	< .001
Lexical density	-.197	-0.38, -0.004	.046
Maas index	-.156	-0.34, 0.04	.115
MATTR	.395**	0.22, 0.55	< .001
HD-D	.408**	0.23, 0.56	< .001
Word frequency (KF)	.119	-0.08, 0.31	.229
Word frequency (BNC)	.096	-0.10, 0.28	.336
Word range (BNC)	.033	-0.16, 0.26	.744
Bigram frequency	.321**	0.14, 0.48	.001
AWL proportion	.045	-0.15, 0.24	.649

Note. 95% CIs for correlation coefficients (Fisher's *z* transformation) are unadjusted for multiple comparisons. *p*-values reported are unadjusted. Significance was determined using Benjamini–Hochberg FDR correction at $\alpha = .05$. Asterisks indicate FDR-adjusted significance: * $p < .05$; ** $p < .01$.

After FDR correction, text length showed the strongest positive correlation with writing scores ($r = .775$, FDR- $p < .01$), followed by HD-D ($r = .408$, FDR- $p < .01$), MATTR ($r = .395$, FDR- $p < .01$), and bigram frequency ($r = .321$, FDR- $p < .01$). Lexical density showed a weak negative correlation ($r = -.197$), but this did not

reach significance after FDR correction (FDR-adjusted $p = .092$). No significant correlations were found for the Maas index, word frequency measures (KF and BNC written), word range, or AWL proportion (all FDR- $p > .05$). These findings suggest that text length, lexical diversity (HD-D, MATTR), and bigram frequency were positively associated with writing scores, while the other lexical indicators were not.

To examine whether the relationship between lexical richness and writing scores was driven primarily by the Language Use dimension, correlations were calculated between key lexical indicators and the three subscores (Topic Development, Organization, and Language Use).

As shown in Table 4, HD-D correlated positively with all three subscores: Language Use ($r = .40, p < .01$), Topic Development ($r = .41, p < .01$), and Organization ($r = .32, p < .01$). MATTR showed a similar pattern. Bigram frequency also correlated with Language Use ($r = .29, p < .01$) and Topic Development ($r = .30, p < .01$), but its correlation with Organization did not reach significance ($r = .19, p = .06$). Lexical density showed a weak negative correlation with Topic Development ($r = -.20, p < .05$) and no significant correlations with Organization or Language Use.

Table 4

Correlations Between Key Lexical Richness Indicators and Writing Sub-scores (N = 103)

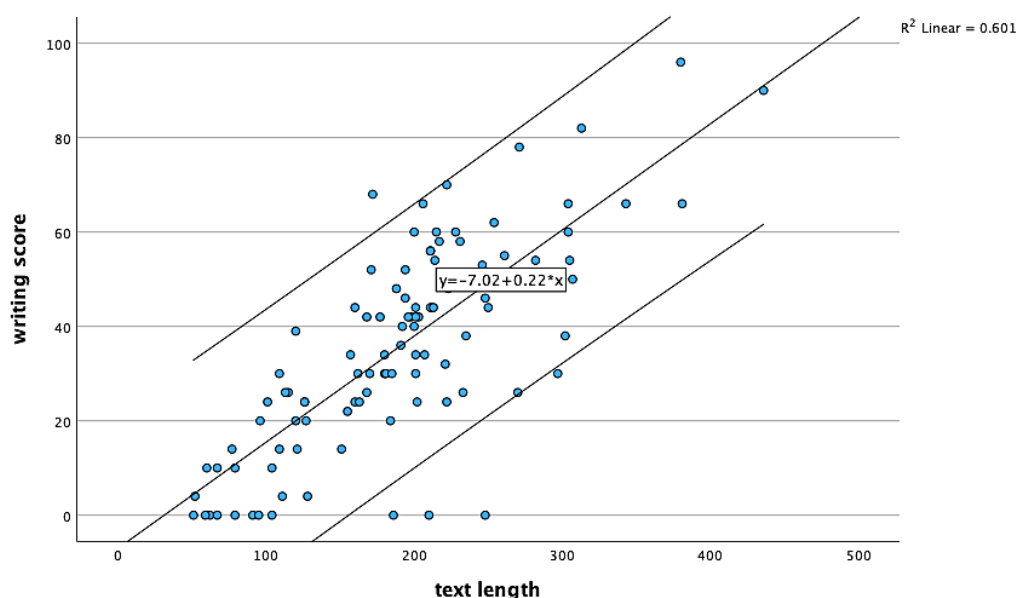
Lexical Indicator	TD score	Org score	LU score
HD-D	.41**	.33**	.40**
MATTR	.38**	.32**	.39**
Bigram frequency	.30**	.19	.29**
Lexical density	-.20*	-.08	-.19

Note. TD = Topic Development; Org = Organization; LU = Language Use. ** $p < .01$; * $p < .05$ (two-tailed).

These results show that the observed associations between lexical diversity and total writing scores extended across all three scoring dimensions rather than being confined to Language Use. To visually display the relationship between writing scores and text length, Figure 1 presents a scatter plot.

Figure 1

Scatterplot of Writing Scores Against Text Length (N = 103)



The scatterplot above reveals a clear positive linear trend; the fitted regression line ($y = -7.02 + 0.22x$) suggests that each additional token was associated with an increase of approximately 0.22 points.

The Predictive Power of Lexical Richness Indicators for Academic Writing Proficiency, Independent of Text Length

Following the theory-based selection described in the Methods section, the variable-selection procedure for the regression models produced the following set of predictors. For lexical variation, initial correlation analysis showed strong collinearity between HD-D and MATTR ($r = .869$, $p < .01$) and between HD-D and the Maas index ($r = -.860$, $p < .01$), suggesting that these indicators reflect the same underlying construct; accordingly, only HD-D was retained. For lexical sophistication, three indicators were retained to reflect its distinct facets: BNC Written Frequency for word-level frequency (with KF frequency and BNC range excluded due to strong collinearity with BNC Written Frequency, $r = .887$ and $.851$, respectively, both $p < .01$), AWL proportion for academic vocabulary, and bigram frequency for n-gram frequency. Lexical density was retained as the sole representative of this construct.

A total of five lexical indicators were included in the regression models as predictors: HD-D, lexical density, BNC Written Frequency, AWL proportion, and

bigram frequency. In Model 1, text length was not included, while in Model 2, text length was included as a control variable to isolate the unique contribution of lexical features beyond textual productivity. The results are presented in Table 5.

Table 5

Summary of Multiple Linear Regression Models Predicting Writing Scores (N = 103)

Model 1 (without text length)					
Predictor	B (SE)	β	<i>p</i>	VIF	95% CI for B
Constant	-8.21 (49.89)	—	.870	—	-107.23, 90.81
HD-D	202.44 (42.89)	.41	< .001**	1.036	117.31, 287.57
Lexical density	-73.70 (59.46)	-.15	.218	2.096	-191.71, 44.31
BNC Written Freq	-14.15 (24.76)	-.074	.569	2.251	-63.30, 34.99
AWL proportion	39.66 (67.32)	.055	.557	1.186	-93.95, 173.28
Bigram frequency	50.10 (19.91)	.27	.013*	1.562	10.59, 89.62

Note. $R^2 = .281$; Adjusted $R^2 = .244$; $F(5, 97) = 7.596$, $p < .001$; Durbin-Watson = 1.416

Model 2 (with text length)					
Predictor	B (SE)	β	<i>p</i>	VIF	95% CI for B
Constant	-11.67 (35.09)	—	.740	—	-81.32, 57.98
HD-D	85.09 (32.36)	.174	.010**	1.192	20.85, 149.33
Lexical density	-47.46 (41.90)	-.099	.260	2.105	-130.63, 35.71
BNC Written Freq	-30.17 (17.49)	-.157	.088	2.270	-64.89, 4.54
AWL proportion	-7.77 (47.58)	-.011	.871	1.198	-102.22, 86.68
Bigram frequency	21.56 (14.30)	.116	.135	1.63	-6.80, 49.92
Text length	0.20 (0.02)	.691	< .001**	1.30	0.16, 0.24

Note. $R^2 = .648$; Adjusted $R^2 = .626$; $F(6, 96) = 29.487$, $p < .001$; Durbin-Watson = 1.813. ** $p \leq .01$; * $p < .05$. Standard errors are shown in parentheses. 95% CIs for B are reported in the final column. VIF values below 5 suggest no multicollinearity concerns.

As shown in Table 5, Model 1 was significant and explained 28.1% of the variance in writing scores ($F(5, 97) = 7.596$, $p < .001$). HD-D ($\beta = .41$, $p < .001$) and bigram frequency ($\beta = .27$, $p = .013$) were found to be significant predictors, though lexical density, BNC Written Frequency, and AWL proportion were not. In Model 2, text length was included as a control variable. The model explained 64.8% of the variance ($F(6, 96) = 29.487$, $p < .001$). Text length served as the dominant predictor

($\beta = .691$, $p < .001$). HD-D retained a small but significant independent association ($\beta = .174$, $p = .010$), while lexical density, BNC Written Frequency, AWL proportion, and bigram frequency were no longer significant. The substantial increase in R^2 from Model 1 to Model 2 (.281 to .648) indicates that much of the variance attributed to lexical features overlaps with that explained by text length. All VIF values were below 5, suggesting no multicollinearity concerns.

To evaluate whether the selection of the lexical diversity indicator affected the substantive conclusions, a sensitivity model was estimated by replacing HD-D with MATTR while keeping all other predictors the same. A consistent pattern—with text length as the dominant predictor and the diversity indicator retaining a small but significant independent association—would support the robustness of the findings. The results are shown in Table 6.

Table 6

Summary of Sensitivity Regression Model (MATTR Replacing HD-D, N = 103)

Predictor	B (SE)	β	p	VIF	95% CI for B
(Constant)	-7.81 (33.56)	—	.817	—	-74.43, 58.81
MATTR	85.51 (31.16)	.182	.007**	1.201	23.67, 147.36
Lexical density	-39.16 (41.58)	-.082	.349	2.085	-121.69, 43.37
BNC Written Freq	-28.47 (17.51)	-.148	.107	2.289	-63.23, 6.29
AWL proportion	-26.52 (47.40)	-.037	.577	1.196	-120.61, 67.57
Bigram frequency	24.37 (14.25)	.132	.090	1.626	-3.91, 52.65
Text length	0.20 (0.02)	.691	< .001**	1.288	0.16, 0.24

Note. $R^2 = .650$, adjusted $R^2 = .629$, $F(6, 96) = 29.761$, $p < .001$, Durbin-Watson = 1.890. ** $p < .01$. 95% CIs for B are reported in brackets. VIF values below 5 suggest no multicollinearity concerns.

The sensitivity model using MATTR instead of HD-D produced consistent results. Text length remained the dominant predictor ($\beta = .691$, $p < .001$), and MATTR retained a small but significant independent association ($\beta = .182$, $p = .007$). The other lexical indicators were nonsignificant. The R^2 (.650) was virtually identical to that of the main model (.648), confirming robustness. All VIF values were below 5.

Discussion

The Characteristics of Lexical Richness in Academic Writing by Test Takers Across Different Proficiency Levels

Regarding Objective 1, before FDR correction, differences were found in five metrics: text length, the Maas index, MATTR, HD-D, and bigram frequency. Following FDR correction, text length, MATTR, HD-D, and bigram frequency remained significant, while the Maas index did not.

The most apparent difference between the two groups was found in text length. Essays written by the Meets Standard Group were longer than those by the Below Standard Group, with a large effect size. Writing longer texts is indeed a recognized characteristic of better L2 writers, and the current findings corroborate this pattern (Hyland, 2016; Weigle, 2002). However, the observed difference in text length is substantial, suggesting that the ability to develop more extensive arguments in timed argumentative writing may serve as a strong indicator of writing performance. This can be reasonably attributed to the nature of the KKU-AELT writing task, which requires test-takers to develop arguments and support their claims, naturally leading more skilled writers to generate more detailed texts.

For lexical variation, both HD-D and MATTR indicated moderate-to-large differences, consistently distinguishing the two proficiency groups. This aligns with previous findings that lexical diversity distinguishes higher- from lower-scoring essays (Engber, 1995; Zhang & Ren, 2018). The consistency between the two length-robust indices supports the interpretation that lexical diversity is a genuine distinguishing feature, though neither measure is completely independent of text length (HD-D: $r = .325$; MATTR: $r = .299$). In the Below Standard group, the standard deviation for HD-D was twice that of the Meets Standard group, indicating greater heterogeneity among lower-proficiency writers—some used diverse vocabulary, whereas others relied on a limited range of words. This finding offers pedagogical implications for targeted vocabulary instruction.

Bigram frequency also differed between the two groups, with the Meets Standard Group using more frequent two-word combinations. This finding may be related to research on multiword sequences, which suggests that awareness of multiword units is an important indicator of L2 proficiency (Simpson-Vlach & Ellis, 2010). The production of conventional collocations or multiword sequences has often been linked to greater fluency and native-like writing ability, and the present findings suggest that it can distinguish candidates at different proficiency levels writing short, timed essays as well.

Notably, no significant differences between the two groups were found in lexical density, word frequency, word range, or AWL proportion. As a well-known feature of academic writing, lexical density in the L2 writing literature is commonly recognized as being higher than in spoken discourse (Biber & Gray, 2016). The absence of a difference in lexical density between the higher- and lower-proficiency groups was therefore unexpected. It may be due to the similar proportion of content words in the essays by L2 writers in both groups, or alternatively, it may be a consequence of task conditions when completing argumentative writing on a common topic. Similarly, since no distinctions were detected between higher- and lower-proficiency writing in word frequency or word range, these results indicate that both groups relied heavily on a similar set of high-frequency words, given that this type of writing only required writers to access words from common English vocabulary to express their stance rather than specific technical terms (Wang, 2011).

In contrast to Srisawat and Poonpon's (2023) findings, academic word use was not significantly correlated with writing scores from the same English test (KKU-AELT). This may be because Srisawat and Poonpon counted tokens of academic words, whereas the present study used a normed proportion index, which controls for text length effects. This difference underscores the importance of selecting lexical indices carefully, as the relationship between academic word use and writing performance may be less robust than previously believed when considering the effect of text length on raw counts.

The Relationship Between Lexical Richness Indicators and Academic Writing Scores

With respect to Objective 2, Pearson's correlation analysis showed that text length had the strongest positive correlation with writing scores ($r = .775$), followed by HD-D ($r = .408$), MATTR ($r = .395$), and bigram frequency ($r = .321$). Lexical density showed a weak negative correlation ($r = -.197$), but this did not meet significance after FDR correction ($\text{FDR-}p = .092$).

This suggests that writing assessments typically reward writers who delve deeply into meaningful content and generate longer texts (Weigle, 2002). The KKU-AELT rubric evaluates development, organization, and language use—dimensions closely related to length. Although these dimensions represent substantive aspects of writing quality, the strong association with text length also indicates that textual productivity may influence scores.

The finding that lexical variation indices (HD-D and MATTR) showed only a moderate positive correlation with writing scores is consistent with research showing that lexical variation measures correlate with assessments of writing quality.

For example, McCarthy and Jarvis (2010) found that for a dataset of L1 and L2 student writing, HD-D was strongly correlated with native-speaker judgments of stylistic lexical diversity. The weaker correlation in this study could be explained by several factors. First, unlike judgments of lexical diversity, the writing task scores used in this study include judgments of style as well as judgments of content and organization. The present study does not isolate word use but instead reflects the overall quality of writing (McCarthy & Jarvis, 2010). Second, as the writing task is timed, it imposes time limitations on the test taker and is based on general topics—this makes it difficult for students to demonstrate their full lexical repertoire. In addition, automated metrics provide consistent, objective measures by quantifying lexical patterns, while human evaluation is subjective and can be influenced by stylistic interpretations. Human evaluation and automated metrics serve as complementary approaches; correlation strength can vary depending on the scoring criteria and task.

The subscore correlations further explained the pattern of these associations. HD-D correlated most strongly with Topic Development ($r = .41$) and Language Use ($r = .40$), and slightly less strongly with Organization ($r = .32$). This pattern shows that lexical diversity is not simply a reflection of the rubric's Language Use criterion but is also associated with broader aspects of writing quality, such as content elaboration and argument development. This is particularly noteworthy given that the rubric's Topic Development criterion emphasizes the provision of “explanations, exemplifications, and/or details”—features that may naturally require a wider range of vocabulary to express fully. The similar pattern observed for MATTR further supports this interpretation. For bigram frequency, the correlations with Topic Development ($r = .30$) and Language Use ($r = .28$) were comparable, while the nonsignificant correlation with Organization suggests that conventional multiword sequences may be more relevant to content and language than to structural coherence. These findings collectively indicate that lexical richness, especially lexical diversity, is associated with multiple dimensions of writing quality beyond the vocabulary-focused Language Use criterion.

The significant correlation between bigram frequency and writing scores indicates that test takers who used frequent two-word patterns also tended to have higher scores. Research on multiword sequences provides a possible context for understanding this finding. The use of conventional two-word combinations has been associated with fluency and proficiency in language production (Kyle & Crossley, 2015; Simpson-Vlach & Ellis, 2010). As discussed in the next section, this relationship appears to be largely attributable to text length.

It is worth noting that the KKU-AELT rubric's Language Use dimension explicitly rewards vocabulary range (Thongyoi & Poonpon, 2019). The positive correlation between HD-D and writing scores is not confined to the Language Use component. It is consistent with, but does not independently establish, raters' implementation of the rubric. However, the finding that this correlation is only moderate ($r = .408$) suggests that raters may attend to multiple features simultaneously when scoring, with lexical diversity being only one.

Lexical density showed a weak negative correlation with writing scores, although this correlation did not survive FDR correction ($FDR-p = .092$). Despite its non-significance after correction, the pattern is worth brief comment. One possible interpretation is that low-scoring writers tend to produce shorter essays, in which the proportion of content words may be relatively higher due to the limited space for function words. Alternatively, higher-proficiency writers may use more grammatical words to achieve cohesion and coherence, resulting in slightly lower lexical density (Biber & Gray, 2016). Given that the correlation was not significant after FDR correction, however, this pattern should be interpreted with caution and warrants further investigation.

No significant correlations were found for word frequency, word range, or AWL proportion. These indices may not capture meaningful differences in writing quality in the current context. The results of word frequency and word range suggest that test takers at both proficiency levels drew words from the same pool of high-frequency words. When a writing task does not require specialized vocabulary, candidates tend to use high-frequency words for their general-topic essays. Again, the nonsignificant finding of AWL proportion could be explained by the general nature of the prompt, which did not require disciplinary content; it may also reveal the limitations of using the Academic Word List created by Coxhead (2000) as an index for argumentative essays on general topics written under time pressure. The list was first developed for academic texts written across various contexts, but its validity for short, timed essays may therefore be limited. Academic vocabulary is often most prominent and relevant in writing tasks that require addressing subject-specific topics, as suggested by previous research (e.g., Hyland & Tse, 2007).

The Predictive Power of Lexical Richness Indicators for Academic Writing Proficiency, Independent of Text Length

For Objective 3, regression analysis examined whether lexical richness indicators retained predictive power after controlling for text length.

Model 1 showed that, when text length was not controlled for, HD-D and bigram frequency were significant predictors of writing scores, and the model explained

28.1% of the variance. Without controlling for length, this result might suggest that lexical richness is predictive of writing scores. However, in Model 2, when text length was included as a control variable, the explained variance increased substantially to 64.8%, with text length appearing as the dominant factor ($\beta = .691$). The marked increase in R^2 from .281 to .648 suggests that a large portion of the variance previously attributed to lexical features overlaps with that explained by text length.

The influence of the lexical indicators varied. Bigram frequency was no longer significant, suggesting that its association largely overlapped with essay length. Once length was included in the model, bigram frequency provided little additional information about writing quality. The other lexical indicators—lexical density, BNC Written Frequency, and AWL proportion—were not significant in either model.

Even after controlling text length, lexical diversity remained significantly associated with writing scores. The decrease in its coefficient—from $\beta = .41$ in Model 1 to $\beta = .174$ in Model 2—shows that text length and lexical variation share much of their association with writing scores, though a smaller but still significant independent association remains. In practical terms, with text length held constant, an increase of one standard deviation in HD-D ($SD = 0.05$) is associated with an increase of approximately 4.3 score points ($B = 85.09 \times 0.05 \approx 4.3$). The association is modest but statistically detectable, suggesting that lexical variation is related to writing quality beyond simply writing more words. This is noteworthy given that the rubric explicitly rewards vocabulary range (Thongyoi & Poonpon, 2019). Yet the small independent contribution of lexical diversity beyond length reinforces the interpretation that textual productivity may exert a disproportionately strong influence on scoring decisions in practice.

A sensitivity analysis was conducted to evaluate whether the selection of the lexical diversity indicator affected the conclusions. As stated in the Results section, replacing HD-D with MATTR while keeping all other predictors the same produced consistent results: text length remained the dominant predictor and MATTR retained a small but significant independent association. The R^2 (.650) and adjusted R^2 (.629) were nearly identical to those of the main model (.648 and .626, respectively), confirming that the substantive conclusions are robust to the choice of lexical diversity indicator.

It is informative to compare these results with Srisawat and Poonpon (2023), who worked on the same English test (KKU-AELT) and examined lexical variation and writing scores using TTR. They found a negative correlation between lexical variation and writing scores. This study, using HD-D and MATTR instead, yielded positive correlations. TTR is highly sensitive to text length, whereas HD-D and

MATTR are less length-dependent (Covington & McFall, 2010; McCarthy & Jarvis, 2007, 2010). Because of the effect of length on writing scores, the negative correlation obtained by Srisawat and Poonpon using TTR may have been more related to length than to lexical variation itself. This comparison thus supports the use of length-robust lexical variation indices. Indeed, the length of timed essays varies among test takers; hence, length is a potentially confounding variable. Using indices that are less sensitive to text length therefore improves the interpretability of the results.

The finding that text length accounts for a substantial portion of the variance in writing scores can be understood by considering the nature of timed argumentative writing tasks. Higher-proficiency writers tend to produce longer texts on these tasks. This may be because they have more to articulate, and because they have access to richer linguistic resources to elaborate on their ideas, provide evidence, and construct coherent arguments (Hyland, 2016; Weigle, 2002). Therefore, length can serve as a proxy for overall writing performance, including content elaboration, organizational complexity, and linguistic fluency. This indicates that measures of lexical richness—which tend to increase in longer texts—are related to text length. The relationship between vocabulary use and writing quality may be related, at least in part, to the ability to produce extended discourse. Researchers who do not control text length risk conflating vocabulary knowledge with discourse-level writing performance.

Conclusion

The findings indicate that higher-proficiency writers produced longer texts and demonstrated greater lexical diversity than lower-proficiency writers. Among all lexical richness indicators, text length showed the strongest correlation with writing scores, followed by lexical diversity (HD-D). Regression analysis revealed that the model including text length explained 64.8% of the variance, with text length as the dominant predictor. After controlling for text length, lexical diversity (HD-D) retained a small but statistically significant independent association, whereas bigram frequency and the other lexical indicators did not. A sensitivity analysis replacing HD-D with MATTR confirmed that these conclusions were robust to the choice of lexical diversity indicator.

The study makes contributions in the following ways. Methodologically, it extends previous validation of length-robust lexical diversity indices (HD-D, MATTR) to the Thai EFL context. It also offers a template for future research by comparing regression models with and without text length as a control variable to isolate the unique contribution of lexical features. In practice, it demonstrates that lexical diversity (as measured by HD-D) had a consistent association with writing performance in

this context, whereas most sophistication indicators and lexical density did not. The strong association between text length and scores also highlights the importance of content development and organization, which are already emphasized in the KKU-AELT rubric. However, a small but significant independent association with lexical diversity remained; therefore, the rubric's existing attention to vocabulary range should be retained without making lexical diversity a dominant criterion. For pedagogy, the findings support integrating vocabulary instruction with content development. Instead of teaching vocabulary separately, instructors could guide students in expanding short paragraphs into more elaborated ones while naturally infusing the text with different expressions. Feedback on the level of elaboration, alongside vocabulary variety, might have a greater effect on students than isolated exercises.

For KKU-AELT test development, given the small independent effect of lexical diversity, the existing attention to vocabulary range in the higher bands of the Language Use dimension should be retained, but lexical diversity should not become a primary scoring criterion. The finding that bigram frequency did not retain predictive power after controlling for text length suggests that collocation-specific criteria are unnecessary. Additionally, the finding that AWL proportion did not correlate with scores implies that general-topic prompts may not adequately elicit academic vocabulary; if the test aims to assess such vocabulary, prompts with more discipline-specific content could be explored.

Several limitations should be acknowledged. First, the study is based on a single-institution, one-prompt corpus, limiting generalizability; the high-proficiency subsample was small ($n = 8$, Bands 5–7), precluding a three-way analysis. Second, the exclusion of essays under 50 tokens disproportionately removed lower-scoring scripts (all 14 excluded essays were from the Below Standard group), which may have reduced variability in the lower proficiency range and led to a conservative estimate of the true gap. Third, the cross-sectional design prevents causal claims about the relationship between lexical richness and writing performance. Fourth, as noted in the rubric descriptors (Thongyoi & Poonpon, 2019), the Language Use criterion explicitly rewards vocabulary range. While the positive correlation between HD-D and writing scores is consistent with raters' implementation of this criterion, the subscore correlations showed that HD-D was associated with Topic Development and Organization as well as Language Use, suggesting that the effect is not confined solely to the vocabulary-related sub-score. Nevertheless, the construct overlaps between the lexical indicators and the Language Use criterion means that part of the observed association may still be definitional rather than fully independent. Fifth, analyzing unedited texts with spelling errors may have inflated lexical diversity

values for lower-proficiency writers, as misspelled words are treated by TAALED as novel types (Kyle et al., n.d.). Conversely, TAALES-based indices (AWL proportion, frequency, bigram frequency) may be underestimated due to exact matching. The true proficiency gap may therefore be larger than reported. Future research could apply spelling correction to validate these findings.

Future studies should include multiple prompts and cross-institutional samples, so that the results can be applied to different settings. Longitudinal designs would allow tracking of lexical richness development over time and provide stronger evidence for causal relationships. Formal mediation analysis could more directly test the role of text length in the relationship between lexical diversity and writing scores. Additionally, applying automated spelling correction tools (e.g., pylats) to learner corpora prior to analysis would help validate whether the observed lexical patterns are robust to spelling-related noise. Finally, if the test aims to assess academic vocabulary use, prompts with more discipline-specific content could be explored to better elicit such vocabulary.

References

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1), 289–300. <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- Biber, D., & Gray, B. (2016, May 5). *Grammatical complexity in academic English: Linguistic change in writing*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511920776>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.
- Covington, M. A., & McFall, J. D. (2010). Cutting the Gordian knot: The moving-average type-token ratio (MATTR). *Journal of Quantitative Linguistics*, 17(2), 94–100. <https://doi.org/10.1080/09296171003643098>
- Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, 34(2), 213–238. <https://doi.org/10.2307/3587951>
- Coxhead, A. (2012). Academic vocabulary, writing and English for academic purposes: Perspectives from second language learners. *RELC Journal*, 43(1), 137–145. <https://doi.org/10.1177/0033688212439323>
- Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, 4(2), 139–155. [https://doi.org/10.1016/1060-3743\(95\)90004-7](https://doi.org/10.1016/1060-3743(95)90004-7)

- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Field, A. (2018). *Discovering statistics using IBM SPSS statistics* (5th ed.). SAGE.
- Hasko, V. (2022). The categorization of lexical errors in written L2 English: Measuring lexical accuracy. In A. Leńko-Szymańska & S. Götz (Eds.), *Complexity, accuracy and fluency in learner corpus research* (pp. 159–180). John Benjamins. <https://doi.org/10.1075/scl.104.07hof>
- Hyland, K. (2003). *Second language writing*. Cambridge University Press.
- Hyland, K. (2016). *Teaching and researching writing* (3rd ed.). Routledge.
- Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? *TESOL Quarterly*, 41(2), 235–253. <https://doi.org/10.1002/j.1545-7249.2007.tb00058.x>
- Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, 19(1), 57–84. <https://doi.org/10.1191/0265532202lt220oa>
- Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly*, 49(4), 757–786. <https://doi.org/10.1002/tesq.194>
- Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods*, 50(3), 1030–1046. <https://doi.org/10.3758/s13428-017-0924-4>
- Kyle, K., Crossley, S. A., & Jarvis, S. (2021). Assessing the validity of lexical diversity using direct judgments. *Language Assessment Quarterly*, 18(2), 154–170. <https://doi.org/10.1080/15434303.2020.1844205>
- Kyle, K., Crossley, S. A., & Jarvis, S. (n.d.). *TAALED: Tool for the automatic analysis of lexical diversity* [Computer software]. <https://pypi.org/project/taaled/>
- Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, 16(3), 307–322. <https://doi.org/10.1093/applin/16.3.307>
- McCarthy, P. M., & Jarvis, S. (2007). vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4), 459–488. <https://doi.org/10.1177/0265532207080767>
- McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, 42(2), 381–392. <https://doi.org/10.3758/BRM.42.2.381>

- Read, J. (2000). *Assessing vocabulary*. Cambridge University Press.
- Sermsook, K., Liamnimitr, J., & Poonpon, K. (2017). An analysis of errors in English sentences: A case study of Thai EFL students. *English Language Teaching*, 10(3), 101–110. <https://doi.org/10.5539/elt.v10n3p101>
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(4), 487–512. <https://doi.org/10.1093/applin/amp058>
- Srisawat, C., & Poonpon, K. (2023). Revision of an academic English writing rubric for a graduate school admission test. *PASAA Journal*, 65, 234–262. <https://doi.org/10.58837/CHULA.PASAA.65.1.9>
- Thongyoi, K., & Poonpon, K. (2019). An investigation of syntactic complexity in essays written by test takers of Khon Kaen University Academic English Language Test (KKU-AELT). In *Proceedings of the International Conference on English Language Studies 2019 (ICELS 2019)* (pp. 27–34). Khon Kaen University.
- Torruella, J., & Capsada, R. (2013). Lexical statistics and typological structures: A measure of lexical richness. *Procedia – Social and Behavioral Sciences*, 95, 447–454. <https://doi.org/10.1016/j.sbspro.2013.10.668>
- Wang, S. (2011). Analyzing lexical richness in CET-4 Range Finder. *Journal of Chongqing University of Posts and Telecommunications (Social Science Edition)*, 23(6), 104–110. <https://doi.org/10.3969/J.ISSN.1673-8268.2011.06.021>
- Weigle, S. C. (2002). *Assessing writing*. Cambridge University Press.
- Zenker, F., & Kyle, K. (2021). Investigating minimum text lengths for lexical diversity indices. *Assessing Writing*, 47, Article 100505. <https://doi.org/10.1016/j.asw.2020.100505>
- Zhang, X. (2022). The relationship between lexical use and L2 writing quality: A case of two genres. *International Journal of Applied Linguistics*, 32(3), 371–396. <https://doi.org/10.1111/ijal.12420>
- Zhang, X., & Ren, J. (2018). A comparative study in lexical richness between high and low score L2 impromptu compositions. *Journal of Beijing International Studies University*, 40(4), 41–52. <https://doi.org/10.12002/j.bisu.156>