Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-05560-x OPEN

Lexical richness in Chinese university students' EFL writing: a corpus-based comparison

Yang Yang[1] & Xubo He[2]✉

Lexical richness (LR) is widely recognized as a key indicator of proficiency among learners of English as a foreign language (EFL). With the rise of numerous software tools capable of automatically measuring LR in recent years, LR has received increasing attention as a focal area of research in the context of Chinese university students' (CUSs) EFL writing. However, these quantitative studies often reduce LR to numerical values, lacking a comparative framework for evaluating lexical proficiency in CUSs' EFL writing. This study adopts Foster and Tavakoli's approach, using native language proficiency as a comparative baseline. It examines LR in CUSs' EFL writing compared to English as a native language (ENL) writing, utilizing corpus-based data from *SWECCL 2.0* and *LOCNESS* across three dimensions: lexical density, sophistication, and variation. The results of Mann–Whitney U tests reveal that CUSs' lexical density in English writing is comparable to that of ENL writing, which may be attributed to the influence of their native language, Chinese. However, they exhibit lower mean ranks of lexical sophistication and variation, showing distinct patterns compared to ENL writing. These disparities may be attributed to factors like limited exposure to advanced vocabulary and cultural attitudes toward risk-taking in language use. Practical pedagogical implications include enhancing exposure to low-frequency vocabulary through enriched input, incorporating contextualized vocabulary instruction, and providing feedback-driven writing tasks that promote lexical variation. This study contributes to the understanding of LR in EFL contexts by emphasizing the need for targeted instructional strategies and offering insights into improving lexical proficiency among EFL learners.

[1] Southwest University, Chongqing, China. [2] Universiti Putra Malaysia, Seri Kembangan, Malaysia. ✉email: gs56743@student.upm.edu.my

Introduction

Lexis is widely regarded as a foundational component of English as a Foreign Language (EFL) writing (Caro and Mendinueta 2017; Mirzaei et al. 2016). Lexical studies have advanced significantly in recent years (e.g., Ibrahim et al. 2019; Kong et al. 2023; Kyle 2020; Shi and Lei 2022; Zhang et al. 2021) and lexical measurement software has enabled large-scale analysis of lexical features in corpora.

EFL writing research places strong emphasis on lexical richness (LR), defined as "the variety of lexis" (Malvern and Richards 2013, p. 3622), which reflects both the breadth and complexity of a learner's productive vocabulary. LR is recognized as a key indicator of EFL proficiency (Malvern and Richards 2013; Yang et al. 2023) and is incorporated into composition scoring systems such as the IELTS Writing band descriptors. For example, to achieve a high band score in IELTS writing, test takers must use "a wide range of vocabulary accurately and appropriately with very natural and sophisticated control of lexical features" (British Council 2023, p. 3).

LR can be measured from the following three dimensions: lexical density, lexical sophistication, and lexical variation[1] (Kyle 2020; Pan et al. 2024). Lexical density is conceptually defined as "the density with which the information is presented" (Halliday 1989, p. 62). Therefore, it primarily reflects the concentration of informational and conceptual content conveyed by content rather than function words. Operationally, it is calculated as the proportion of content words to the total number of words (Camiciottoli 2007; Castello 2008). Lexical sophistication refers to "the learner's use of sophisticated and advanced words" (Kim et al. 2018, p. 121) and is typically measured by the proportion of low-frequency words in a text. Higher lexical sophistication reflects a learner's ability to use rare and complex words effectively, demonstrating lexical depth and the ability to convey nuanced meaning. Finally, lexical variation, used interchangeably with the term lexical diversity (Jarvis, 2013), refers to "the range of a learner's vocabulary as displayed in his or her language use" (Lu 2012, p. 192), which assesses the diversity of lexical items within a text, representing the breadth of a language learner's vocabulary knowledge.

Lexical sophistication and variation represent distinct aspects of vocabulary use. Lexical sophistication focuses on the depth of vocabulary knowledge by measuring the use of infrequent and advanced words, whereas lexical variation emphasizes the breadth of vocabulary knowledge, reflecting the range of distinct words used in a text. Together, these dimensions provide a comprehensive and complementary view of learners' lexical richness.

Numerous software tools and systems are now available for automatically calculating LR in EFL writing, including *CohMetrix*[2](McNamara et al. 2014), *Lexical Complexity Analyzer* (Ai and Lu 2010; Lu 2012), *AntWordProfiler* (Anthony 2024), *VocabProfile* (Cobb 2018), *TAALED* and *TAALES* (Kyle et al. 2018; Kyle and Crossley 2015). These tools calculate various LR indices for EFL writing samples. However, the outputs are merely numerical values and, on their own, fail to fully capture the lexical proficiency of EFL writing. In this regard, Foster and Tavakoli (2009) emphasize that using English as a Native Language (ENL) writing as a baseline for comparison is essential to accurately evaluate the LR of similar EFL writing.

To address this, the study uses ENL writing from American and British university students as a reference to better understand the LR of Chinese university students' (CUSs) EFL writing. By comparing the two, this study aims to identify differences across LR dimensions, providing targeted recommendations for enhancing CUSs' EFL writing. The goal is not to set a benchmark, but to better understand lexical patterns in EFL writing and inform vocabulary instruction.

Guided by Kyle's (2020) framework of LR dimensions, this study addresses the following three research questions:

(1) What is the difference in lexical density between CUSs' EFL writing and native writer's ENL writing?
(2) What is the difference in lexical sophistication between CUSs' EFL writing and native writers' ENL writing?
(3) What is the difference in lexical variation between CUSs' EFL writing and native writers' ENL writing?

Literature review

The development of lexical richness dimensions. Laufer (1991) conducted one of the earliest systematic studies on LR, examining 47 first-year English majors who were native speakers of Hebrew or Arabic. She measured changes in LR over one to two semesters using four dimensions: lexical density, lexical sophistication, lexical variation, and lexical originality. Her findings indicated that lexical sophistication was the most effective dimension for distinguishing students' initial and later writing performance.

However, Laufer and Nation (1995) later questioned the reliability of lexical originality as a dimension of LR. Lexical originality refers to the proportion of words in a composition that do not appear in other texts from the same group. This method poses challenges because it depends on cross-text comparisons, making it difficult to ensure reliable results. Read (2000) also argued that lexical originality is not a suitable measure of LR in ESL learners' writing. As a result, lexical originality has gradually been abandoned in subsequent LR research due to its limited applicability and generalizability.

Building upon Laufer and Nation's framework, Engber (1995) examined the relationship between LR and writing quality in ESL compositions. Unlike Laufer and Nation, Engber (1995) introduced lexical errors as an additional dimension of lexical richness. She analyzed LR using four dimensions: lexical density, lexical variation with errors, lexical variation without errors, and the percentage of lexical errors.

Read (2000) further justified the inclusion of lexical errors as an essential dimension of LR measurement. He argued that effective writing should exhibit four features: (1) a high proportion of content words, (2) appropriate use of low-frequency words suited to the topic and register, (3) a diverse vocabulary that avoids excessive repetition, and (4) a minimal number of lexical errors. These features correspond to four proposed dimensions of LR: lexical density, lexical sophistication, lexical variation, and lexical errors.

Daller et al. (2007) introduced the concept of lexical space and elaborated on the dimensions of lexical knowledge. They identified three dimensions of lexical knowledge: breadth, depth, and fluency. Some of the previously discussed LR dimensions can be classified under lexical breadth and lexical depth. Lexical fluency, on the other hand, refers to the number of words a language learner can produce within a given time frame (Goodfellow et al. 2002). It reflects the learner's ability to quickly retrieve words from memory when needed (Daller et al. 2007; Šišková 2012). In time-constrained writing tasks, lexical fluency is typically measured by text length (i.e., total word count). However, as it applies only to timed tasks and lacks cross-context comparability, lexical fluency is limited as an LR measure.

Jarvis (2013, 2017) proposed a multidimensional model of lexical diversity, incorporating richness, evenness, dispersion, and disparity. He argued that lexical diversity is both statistical and perceptual, and should be measured using tools that combine computational modeling with human judgment. Kyle et al. (2021) further validated lexical diversity indices by showing their strong

alignment with human judgments. However, given this study's focus on quantitative comparisons of LR across learner and native corpora, a framework that aligns with widely used computational measures was adopted, ensuring consistency and comparability with prior EFL writing research.

Unlike other dimensions, lexical errors are more suited for qualitative lexical research because they require manual identification and classification due to inconsistent error taxonomies across studies. By contrast, the other LR dimensions can be quantitatively measured using specific indices and computational methods through automated software. Considering these limitations, this study adopts the three most widely recognized LR dimensions—lexical density, lexical sophistication, and lexical variation—which align with those used in recent mainstream LR research (e.g., Kyle 2020; Li et al. 2025; Li and Zhang 2021; Pan et al. 2024; Tran and Ma 2025; Zheng 2025). These dimensions provide a robust and widely accepted framework for evaluating lexical richness in EFL writing while ensuring comparability across different studies.

Strands of lexical richness studies. Recent LR research has developed along three main lines. The first and most fundamental line of inquiry explores the relationship between LR and writing quality or language proficiency in EFL or English as a Second Language (ESL) contexts (e.g., Bestgen 2017; Ha 2019; Kojima and Yamashita 2014; Lan et al. 2019; Lemmouh 2010; Lu 2012; Muñoz Acevedo 2014; Shah et al. 2013; Treffers-Daller et al. 2018; Woods et al. 2023; Xie and Shen 2015). This research strand is pivotal, as it is based on the hypothesis of a positive correlation between LR and writing quality.

Numerous studies have established a significant relationship between different measures of LR and EFL writing quality (e.g., Kojima and Yamashita 2014; Lu 2012; Shah et al. 2013; Treffers-Daller et al. 2018; Xie and Shen 2015; Yu 2010). Using correlation and meta-analysis, Lu (2012) identified lexical variation as the strongest predictor of narrative quality among ESL learners, whereas lexical density had no effect and lexical sophistication exhibited only a minimal impact. Similarly, Bestgen (2017) reported that lexical variation, as measured by the number of different words, type-token ratio (TTR), and Guiraud's index, significantly accounted for the variance in overall EFL writing scores. Among these, the number of different words emerged as the best predictor for distinguishing between levels of the Common European Framework of Reference (CEFR). Besides, Woods et al. (2023) demonstrated that a multi-measure approach significantly improves the predictive validity of LR assessments. This finding reinforces the importance of using multiple, complementary indices in evaluating EFL writing quality.

Conversely, other studies have reported different conclusions. Lemmouh (2010), for example, investigated Swedish college students and found no correlation between LR and EFL writing scores. This led the researchers to suggest that LR might not be a central criterion in teachers' assessments of writing. Similarly, Muñoz Acevedo (2014) reported no correlation between LR and foreign language performance. Such discrepancies may stem from differences in tasks and participant populations. Additionally, it is crucial to examine different aspects of LR in EFL writing separately to better understand the relationship between its various dimensions and writing quality.

The second research focus examines the development of LR in EFL writing over time (e.g., Azodi et al. 2014; Shao 2019; Zheng 2015, 2016; Zheng and Feng 2017; Zhu and Wang 2013). This line of research examines how LR develops over time in EFL writing. The evolution of LR is closely tied to language input conditions (Azodi et al. 2014) and is further shaped by factors such as learners' language proficiency, learning background, and cognitive processing abilities. Consequently, differences in learning conditions and backgrounds result in distinct characteristics of lexical development in EFL writing across various groups.

The third research strand involves comparative studies on LR among learners with varying language backgrounds or proficiency levels (e.g., Eckstein and Ferris 2018; Geng and Yang 2021; Lei and Yang 2020; Mahardika 2015; McNamara et al. 2010). These studies typically compare LR across learners of different language backgrounds or proficiency levels using writing samples as corpora. For instance, Eckstein and Ferris (2018) compared the lexical complexity in compositions from native English speakers and second language (L2) students in the same composition classes, noting significant differences in lexical variation. They suggested that L2 students should be made aware of "the value of nuanced and purposeful lexical variety in their writing" (p. 137). In addition, Lei and Yang (2020) compared LR in research articles by Chinese Ph.D. candidates, native English-speaking university students, and native experts across dimensions of lexical density, lexical sophistication, and lexical variation, finding that the LR of Chinese Ph.D. candidates falls between that of native university students and native experts. Other comparative studies have investigated how LR varies across different genres in EFL learners' writing. For instance, Heng et al. (2023) demonstrated that Chinese EFL learners exhibited significant variations in lexical density, variation, and sophistication across argumentative and expository compositions, underscoring the impact of genre on lexical usage patterns.

Apart from Lei and Yang's research, there are few comparative studies on the LR of Chinese EFL learners and native speakers. Lei and Yang's findings suggest that Chinese Ph.D. candidates exhibit relatively high levels of LR; however, the LR of Chinese undergraduates remains underexplored. Additionally, previous studies on productive vocabulary in EFL writing often focus solely on differences in lexical errors, with limited attention given to other LR dimensions.

A review of previous research reveals that studies on LR in EFL writing, particularly those focused on development, are relatively abundant. However, comparative research, especially comparing the LR in CUSs' EFL writing with that of native English writers, is relatively scarce. Thus, this study investigates the differences in LR between CUSs' EFL writing and native writers' ENL writing, focusing on lexical density, sophistication, and variation.

Methods

Data collection. CUSs' EFL writing samples were sourced from the *Spoken and Written English Corpus of Chinese Learners Version 2.0* (SWECCL 2.0; Wen et al. 2008), a well-established EFL corpus. This corpus comprises 4950 compositions from students at 34 universities and colleges across China, providing a comprehensive representation of their writing (Mu 2011; Wen et al. 2008) and serving as a representative learner corpus of Chinese EFL students (Zhang and Yang 2019). The validity of SWECCL has been substantiated by numerous studies since its release (e.g., Gan 2024; Wei 2021; Yan and Li 2019; Zeng et al. 2023; Zhang and Yang 2021). The compositions were handwritten and later digitized without alteration. The corpus' diverse sources and variety of writing tasks ensure it accurately reflects the EFL writing proficiency of CUSs.

To analyze CUSs' EFL writing, the *Louvain Corpus of Native English Essays*[3](LOCNESS; Granger 1998) was employed as a comparative reference corpus. LOCNESS includes argumentative and literary essays by American and British university students, along with A-level student essays, representing authentic native English usage (Akbana 2015; Bozdağ 2019; Paradowski 2015).

Composed under diverse conditions, these essays provide a robust basis for comparison. The ENL corpus was used as a comparative reference due to its extensive documentation and frequent use in corpus-based linguistic research, allowing for a systematic comparison of LR characteristics.

Both SWECCL and LOCNESS are widely recognized for their authenticity and representativeness, with SWECCL reflecting the real writing proficiency of Chinese EFL learners and LOCNESS representing the authentic language use of American and British university students. Extensive research has validated their representativeness in examining EFL and ENL writing, respectively (Wu and Tissari 2021; Zhu and Pu 2024). Additionally, both corpora include metadata on variables such as writer demographics, task types, and situational contexts, enhancing their utility for controlled research. The comparability of the two corpora is well-documented (Chai et al. 2015), particularly regarding contributors' age, essay genre (predominantly argumentative), and writing conditions. Prior studies and evaluations confirm their high comparability, making them suitable for use as baseline and target corpora in comparative analyses of EFL and ENL writing (e.g., Ai and Lu 2013; Wu and Tissari 2021; Zhu and Pu 2024).

The sample includes 600 essays: 400 EFL compositions from SWECCL and 200 ENL essays from LOCNESS. The sample size aligned with Ai and Lu's (2013) methodology, where they utilized the same number of EFL and ENL writings as the present study for a comparative analysis of syntactic complexity.

EFL samples were selected using a combination of purposive and stratified sampling. Initially, purposive sampling was employed, focusing exclusively on the written component of SWECCL to ensure comparability with LOCNESS. Subsequently, stratified sampling was used to randomly select 100 compositions from each academic year group.

Purposive sampling and simple random sampling without replacement were employed to select essays from LOCNESS. To ensure comparability, essays written by British A-level students were excluded, focusing solely on essays by American and British university students. Subsequently, 200 essays were randomly selected using the *Random Generator*[4]. To date, 600 texts have been sampled from SWECCL and LOCNESS. Table 1 provides a summary of the writing samples.

Table 1 Summary of the writing samples in the present study.

Group		Number of texts	Number of words			
			Min	Max	Mean	SD
EFL group	EFL_Grade1	100	125	503	270.58	80.97
	EFL_Grade2	100	76	498	273.92	81.77
	EFL_Grade3	100	57	470	221.58	76.66
	EFL_Grade4	100	146	514	352.58	67.46
	Total	400	57	514	279.67	89.89
ENL group		200	56	1973	746.77	423.79
Total		600	56	1973	435.37	337.07

Table 2 Measures and indices of LR used in the present study.

Measure	Index	Code
Lexical density	Lexical Density	LD
Lexical sophistication	Lexical Sophistication-I	LS1
	Verb Sophistication-II	VS2
Lexical variation	Corrected Type/Token Ratio	CTTR
	Lexical Word Variation	LWV

Despite the observed differences in average text length between the EFL and ENL groups, these discrepancies are not expected to influence the outcomes of this study. As shown in Table 2 and the accompanying formula descriptions, all LR indices adopted in this study are ratio-based measures, which normalize for text length and thus remain stable across samples of varying size (Ai and Lu 2013; Lu and Ai 2015; Wang and Slater 2016; Yang et al. 2022). This rationale aligns with previous comparative studies on EFL and ENL writing that have also utilized SWECCL and LOCNESS as representative corpora (e.g., Lu and Ai 2015; Wu and Tissari 2021; Zhu and Pu 2024). Therefore, the differences in word count between the two groups do not compromise the validity or comparability of the lexical richness analyses conducted in this study.

Data processing. The 600 writing samples were analyzed using the *Lexical Complexity Analyzer* (Ai and Lu 2010; Lu 2012), which calculates 25 indices of lexical density, sophistication, and variation. Although the software includes a broad set of literature-based indices, some have been found to be redundant or highly correlated (Yang et al. 2022). The measures and indices used in this study were adapted from Yang and Zheng's (2024) research. Using the framework of training and testing, they employed ANOVA to develop a refined and concise model for quantitatively measuring the LR of CUSs' EFL writing.

The validity of this refined index model is reinforced by existing research. For instance, this study selects Corrected Type/Token Ratio (CTTR) as the index of lexical variation, as it has been widely recognized for its methodological robustness. Traditional TTR is widely criticized for its sensitivity to text length, making it unreliable for comparative studies (Covington and McFall 2010; Lu 2012). Alternatives such as Mean Segmental TTR (MSTTR) and Mean Average TTR (MATTR) address this issue but introduce limitations, including data loss or dependence on window size, reducing cross-study comparability (Lei and Yang 2020; Richards and Malvern 1997). CTTR offers a more stable alternative, mitigating text-length effects while preserving sensitivity to lexical diversity, making it a widely accepted and methodologically sound choice for this study.

The LR indices utilized in this study are presented in Table 2 along with their corresponding formulas.

$$\text{Lexical Density(LD)} = \frac{Number\ of\ lexical\ words}{Number\ of\ words}$$

$$\text{Lexical Sophistication} - \text{I(LS1)} = \frac{Number\ of\ sophisticated\ lexical\ words}{Number\ of\ lexical\ words}$$

$$\text{Verb Sophistication} - \text{II(VS2)} = \frac{(Number\ of\ sophisticated\ verb\ types)^2}{Number\ of\ verbs}$$

$$\text{Corrected Type/Token Ratio(CTTR)} = \frac{Number\ of\ word\ types}{\sqrt{2(Number\ of\ words)}}$$

$$\text{Lexical Word Variation(LWV)} = \frac{Number\ of\ lexical\ word\ types}{Number\ of\ lexical\ wrods}$$

To ensure clarity in the interpretation of the formulas presented, it is necessary to define and explain certain key terms used in the calculation of lexical richness indices.

First, in the lexical density formulas, the term lexical word requires clarification. A lexical word, also known as a content word, refers to words that carry semantic meaning in contrast to function words, which primarily serve grammatical purposes. Lexical words include nouns, verbs, adjectives, and adverbs.

Second, the dimension lexical sophistication was measured by the two indices: Lexical Sophistication-I (LS1) and Verb Sophistication-II (VS2). LS1 refers to the proportion of relatively sophisticated lexical words (Lu 2012; Read 2000); VS2, by contrast, specifically measures the proportion of sophisticated verbs, thereby capturing the use of advanced verbal constructions in writing (Kyle et al. 2018; Kyle and Crossley 2015). The two indices involve the term "sophisticated". In these calculations, words, lexical words, and verbs were categorized as sophisticated if they did not appear in the list of the 2000 most frequently used words in the British National Corpus (Lu 2012). This classification reflects the extent to which a writer employs advanced or less commonly used vocabulary. A limitation of this approach to defining sophisticated words is that spelling errors and arbitrary letter sequences may also be classified as sophisticated words. To address this issue, an inspection of the corpus used in this study indicated that such instances were rare, confirming that their influence on the results would be negligible.

Finally, in the lexical variation dimension, the indices involve the terms type and token, which are crucial for understanding lexical variation. A token refers to an individual occurrence of a word in a text, meaning that every instance of a word, including repetitions, is counted separately. In contrast, a type represents a unique word form, where repeated words are counted only once. When counting word types, variations of the same lemma (e.g., *write*, *writes*, *writing*, *wrote*, and *written*) were treated as a single type, ensuring a more accurate assessment of lexical variation.

These definitions help clarify the methodological framework for measuring lexical richness in the study, ensuring consistency and transparency in index calculations.

Data analysis. To examine group differences in the three LR dimensions, independent samples *t*-tests or Mann–Whitney U tests were selected based on data distribution and assumption checks. The *t*-test assesses whether the means of two groups differ significantly. When the assumption of normality is violated, the Mann–Whitney U test, a nonparametric alternative, would be applied.

The t-test requires the following assumptions: (1) the dependent variable is continuous, (2) the independent variable is categorical, (3) observations are independent, (4) no significant outliers are present, (5) the dependent variable follows a normal distribution, and (6) homogeneity of variance is maintained. In this study, the dependent variables, score of LR index, were continuous, while the independent variable, group membership, was categorical with two levels: the EFL group and the ENL group. Each text was produced by an individual student and randomly sampled, satisfying the assumption of independence. Therefore, the first three assumptions were satisfied.

However, other assumptions were not fully satisfied. Boxplot analysis identified outliers in all LR indices except Lexical Word Variation. Moreover, histogram analysis showed that most LR indices did not follow a normal distribution. The Kolmogorov–Smirnov test confirmed these findings, with only the EFL group's CTTR showing normal distribution ($p > 0.05$). Moreover, Levene's test for homogeneity of variances indicated significant differences in variance between the EFL and ENL groups ($p < 0.05$) for several variables, including Lexical Density, Lexical Sophistication-I, Verb Sophistication-II, and Corrected TTR.

In summary, three critical assumptions for independent samples t-test were violated: significant outliers, non-normal data distribution, and unequal variances. Therefore, Mann–Whitney U tests were conducted in SPSS to assess group differences across the five LR indices. This test compares mean ranks rather than raw means between groups.

The null and alternative hypotheses for this study are defined as follows:

H_0: There are no significant differences in the mean ranks of LR indices between CUSs' EFL writing and ENL writing.

H_1: There are significant differences in the mean ranks of LR indices between CUSs' EFL writing and ENL writing.

To determine statistical significance, the threshold was set at 0.05. However, given that multiple dependent variables were analyzed, and several Mann–Whitney U tests were conducted simultaneously, the Bonferroni correction (Bonferroni 1935, 1936; Haynes 2013) was applied to mitigate the risk of false positives associated with multiple comparisons. This is a standard procedure for adjusting the significance level to reduce the likelihood of committing Type I errors when performing multiple tests (Yang et al. 2022). As a result, a more stringent threshold was applied to establish statistical significance. For instance, with two lexical sophistication indices, Lexical Sophistication-I and Verb Sophistication-II (see Table 2), being compared between the EFL and ENL groups, the adjusted significance level was set at 0.025 (0.05 divided by 2).

Lastly, since SPSS does not automatically calculate and report the effect size after running the Mann–Whitney U test, it was manually computed using the Z score obtained from the test results and presented in the following section. The effect size for the Mann–Whitney U test was calculated using the following formula:

$$Effect\ Sizer = \frac{|z|}{\sqrt{n}}$$

In this formula, "Z" represents the Z score from the Mann–Whitney U test, and "n" refers to the total sample size.

Results and discussion

This section presents the descriptive statistics for the LR indices of both the EFL and ENL groups, including their mean ranks, distribution illustrated by histograms, and the results of the Mann–Whitney U tests. The findings are interpreted and discussed in relation to relevant previous studies. In the Mann–Whitney U tests, ranks are assigned such that the smallest value is given rank one, the second smallest rank two, and so forth. Thus, a higher mean rank indicates a larger value and superior performance on the corresponding LR index.

Lexical density. Figure 1 depicts the mean ranks and distributions of lexical density for the EFL and ENL groups. The mean rank for the ENL group (312.18) slightly exceeds that of the EFL group (294.66), suggesting that the ENL group demonstrates greater lexical density. Furthermore, the histograms indicate that the mode for the ENL group is higher than that of the EFL group. However, the results of the Mann–Whitney U test (see Table 3) indicate that this difference in lexical density is not statistically significant ($U = 37663.5$, $Z = -1.17$, $p > 0.05$). Moreover, based on Cohen's (1988) guidelines for interpreting effect size, the effect size ($r = 0.05$) is negligible, suggesting minimal difference between the two groups in terms of lexical density.

Lexical density is defined as the ratio of lexical words to the total number of words in a text. This suggests that CUSs produce a proportion of lexical words in their EFL writing comparable to that of their ENL counterparts. Waller (1993) argued that a lexical density of 0.5 suggests that a composition may have been written by a native speaker or by a writer who has achieved a level of lexical density comparable to that typically observed in native English writing. In this study, the mean lexical density for CUSs is 0.51, with mean ranks closely matching those of the ENL group.

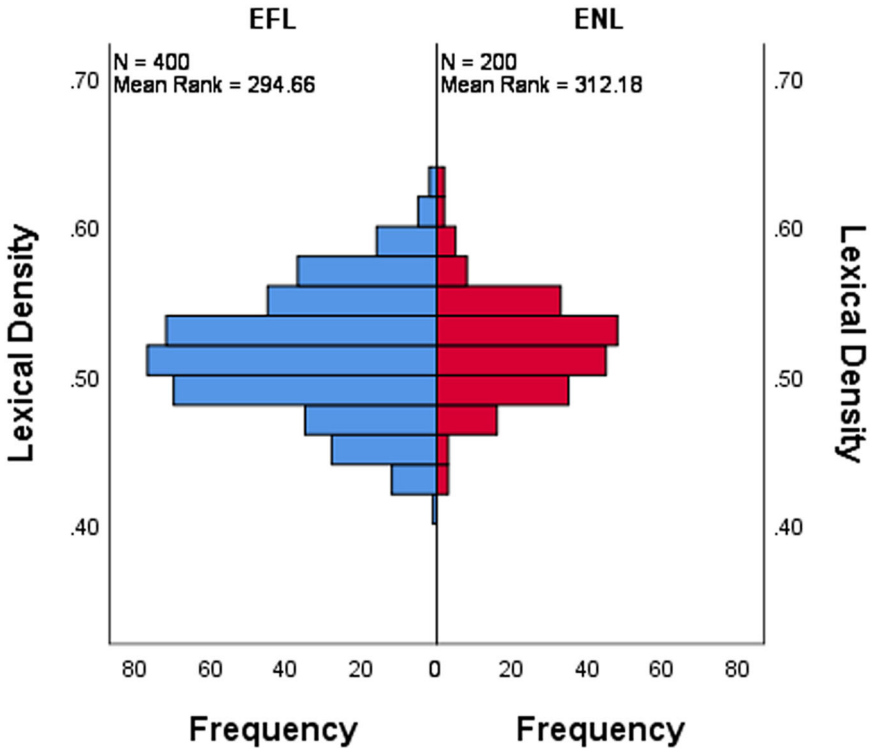

Fig. 1 Mean ranks of Lexical Density of EFL and ENL groups.

Table 3 Ranks and test statistics of Mann-Whitney U test for lexical density.

Index	Group	*N*	(Mean) Mean rank	Mann-Whitney U	Z	Asymp. Sig. (2-tailed)	Effect size r
LD	EFL	400	(0.510) 294.66	37,663.50	−1.17	0.242	0.05
	ENL	200	(0.514) 312.18				

Thus, Chinese university EFL learners appear to achieve lexical density levels comparable to ENL writers.

This finding aligns with prior research (e.g., Eckstein and Ferris 2018; Lei and Yang 2020), which similarly concluded that there are no significant differences in lexical density between EFL and ENL writing of the same genre or among writers with comparable age or proficiency levels. For instance, Lei and Yang (2020) noted that the lexical density in research articles authored by Chinese Ph.D. candidates is comparable to that of ENL experts and even surpasses that of novice ENL students.

Many prior studies have reported similar findings but offered limited explanation, especially regarding Chinese EFL learners. A closer analysis suggests that this result may be influenced by linguistic and educational factors unique to Chinese EFL learners. One possible explanation involves cross-linguistic influence from Chinese and its typological characteristics. The proportion of lexical words in Chinese is higher than in English. Liu (2021) found that in a random sample of news articles, Chinese nouns and verbs in *People's Daily* constituted 67.58% of the content, compared to only 47.11% in the English *New York Times*. Similarly, Ye and Xin (2018) reported that the English translation of China's *Report on the Work of the Government* (政府工作报告) contains a higher proportion of lexical words than its American counterpart, the *State of the Union Addresses*. Chinese EFL learners tend to use more content words in their Chinese writing, and this tendency may transfer to their EFL writing, resulting in similar proportions of lexical words, even though other LR dimensions, such as lexical sophistication and variation, may be lower compared to ENL writers.

Pedagogical emphasis on accuracy and clarity in Chinese EFL instruction may also contribute to this result. Chinese EFL learners are often trained to write concise and information-rich texts to align with standardized testing requirements, such as the CET-4 (College English Test Band 4) and CET-6, where lexical density is indirectly rewarded. This instructional emphasis might result in students producing texts with high lexical density, despite potential limitations in their lexical sophistication and variation.

Originally, lexical density was used to distinguish written from spoken language (Read 2000). A lexical density greater than 0.4 is typically associated with written language, as spoken language usually has a lower lexical density. A higher lexical density suggests that a text conveys more information. However, based on the findings of this study and previous research (e.g., Lei and Yang 2020; Linnarud 1986), it can be concluded that lexical density may not be a reliable measure for distinguishing between EFL and ENL writing. The findings underscore the need to consider additional LR dimensions, such as lexical sophistication and variation, to capture a fuller picture of language learners' proficiency.

Lexical sophistication. Figure 2 presents the mean ranks and distribution of Lexical Sophistication-I and Verb Sophistication-II for both the EFL and ENL groups. The histograms suggest that the ENL group exhibits higher values in both indices compared to the EFL group. Specifically, the mean rank for Lexical Sophistication-I is higher in the ENL group (426.16) than in the

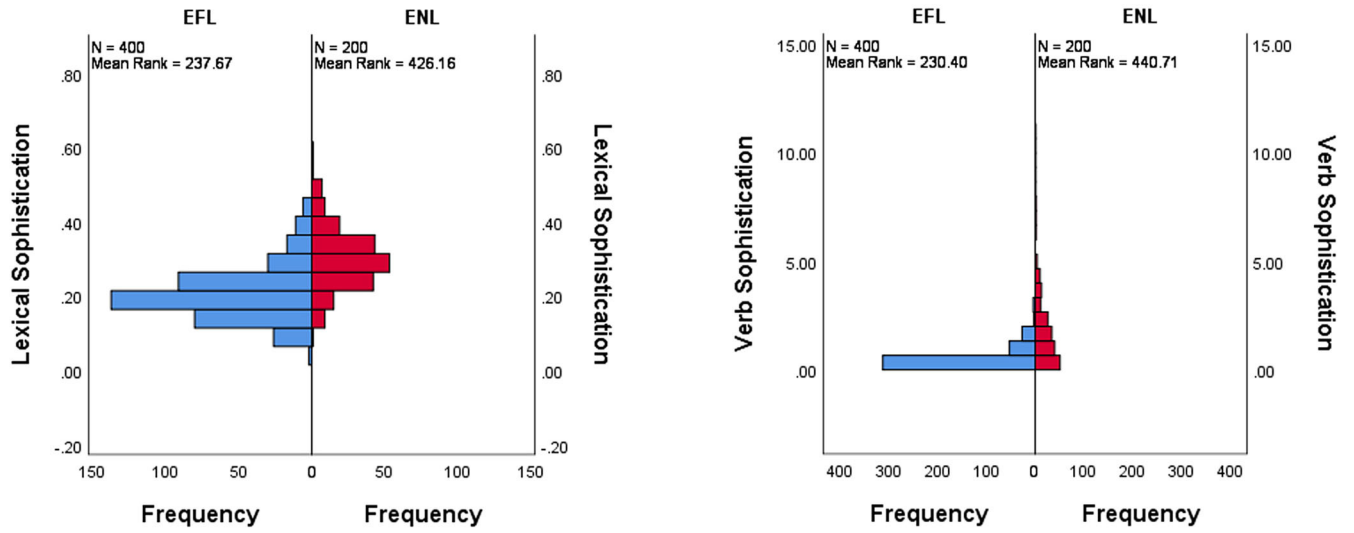

Fig. 2 Mean ranks of lexical sophistication indices of EFL and ENL groups.

Table 4 Ranks and test statistics of Mann-Whitney U test for lexical sophistication.

Index	Group	*N*	(Mean) Mean rank	Mann-Whitney U	Z	Asymp. Sig. (2-tailed)	Effect size r
LS1	EFL	400	(0.206) 237.67	14,867.50	−12.57	0.000[a]	0.51
	ENL	200	(0.300) 426.16				
VS2	EFL	400	(0.433) 230.40	11,959.00	−14.01	0.000[a]	0.57
	ENL	200	(2.054) 440.71				

[a]Difference is statistically significant at the level of 0.025 after Bonferroni Correction.

EFL group (237.67). Similarly, the mean rank for Verb Sophistication-II is also greater in the ENL group (440.71) compared to the EFL group (230.4).

Mann–Whitney U tests revealed significant differences between the EFL and ENL groups in both Lexical Sophistication-I ($U = 14867.50$, $Z = -12.57$, $p = 0.000$) and Verb Sophistication-II ($U = 11959.00$, $Z = -14.01$, $p = 0.000$). Effect sizes were medium ($r = 0.51$ and 0.57, respectively), according to Cohen's (1988) guidelines. These findings suggest that CUSs use considerably fewer sophisticated words and verbs in their EFL writing compared to ENL students.

Lexical sophistication reflects a writer's use of low-frequency and advanced vocabulary. In this study, it is assessed through two indices: Lexical Sophistication-I and Verb Sophistication-II. As shown in Table 4, approximately 20% of lexical words in CUSs' EFL writing are sophisticated, compared to 30% in ENL writing. Similarly, the values for Verb Sophistication-II indicate a substantial gap between CUSs and ENL writers in their use of sophisticated verbs. This is particularly relevant given the central role of verbs in academic writing, especially reporting verbs, which are essential for source integration, conveying authorial stance, and persuasion (Hyland 2014).

The pronounced gap in sophisticated word use may stem from several interrelated factors. First, exposure to low-frequency words in authentic contexts is a key determinant. ENL writers, immersed in an English-dominant environment, have abundant exposure to advanced vocabulary through daily interactions and academic materials. In contrast, CUSs rely heavily on textbook-based learning, which often emphasizes high-frequency vocabulary (Qin and Wen 2007). This limited exposure reduces opportunities for learners to internalize and use low-frequency words.

Second, teaching practices in Chinese EFL classrooms often prioritize grammatical accuracy and basic vocabulary over the development of sophisticated language use. Many teachers adopt a prescriptive approach focused on avoiding errors, which discourages students from experimenting with unfamiliar or complex vocabulary (Zhu and Wang 2013). This accuracy-first approach results in overuse of high-frequency vocabulary.

Finally, limited writing practice targeting lexical sophistication and cultural attitudes toward risk-taking may further contribute to the issue. Exam-style prompts in Chinese universities prioritize clarity over lexical variety, while fear of errors often discourages students from experimenting with low-frequency words, hindering the expansion of their lexical repertoire (Zheng 2016).

To address these limitations, EFL instruction in China should adopt targeted strategies to enhance students' exposure to and use of low-frequency words. Introducing vocabulary-enriched reading materials, such as academic journals and literature, can provide contextual exposure to advanced vocabulary (Hyland 2014). Tasks that integrate sophisticated vocabulary into meaningful contexts, such as argumentative essays, and explicit instruction on word families and collocations can further support lexical development (Lu 2012). Workshops using vocabulary substitution and corpus-based tools (e.g., the *Lexical Complexity Analyzer*) can help students replace overused words with advanced alternatives (Kyle and Crossley 2015). Additionally, fostering a classroom culture that values experimentation with less familiar vocabulary and incorporating lexical sophistication into grading rubrics may encourage students to expand their lexical repertoire. These strategies collectively aim to balance accuracy with creativity, promoting deeper engagement with sophisticated vocabulary.

Lexical variation. Figure 3 presents the mean ranks and distributions for the two lexical variation indices, Corrected TTR

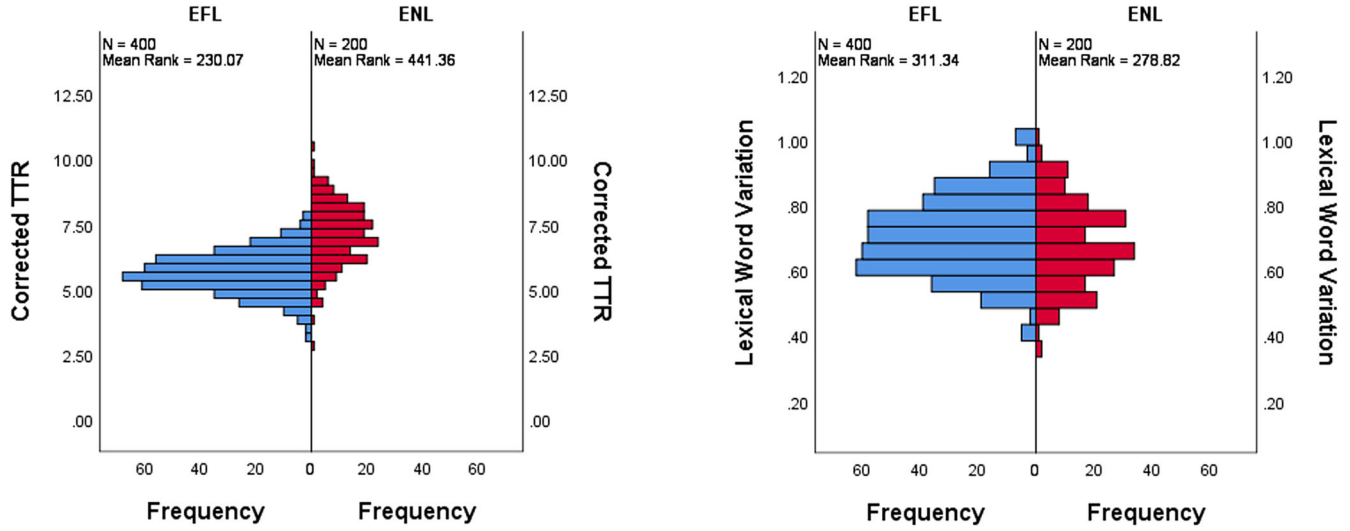

Fig. 3 Mean ranks of lexical variation indices of EFL and ENL groups.

Table 5 Ranks and test statistics of Mann-Whitney U test for lexical variation.

Index	Group	*N*	(Mean) Mean rank	Mann-Whitney U	Z	Asymp. Sig. (2-tailed)	Effect size r
CTTR	EFL	400	(5.627) 230.07	11,828.50	−14.07	0.000[a]	0.57
	ENL	200	(7.123) 441.36				
LWV	EFL	400	(0.704) 311.34	35,664.00	2.17	0.030	0.09
	ENL	200	(0.679) 278.82				

[a]Difference is statistically significant at the level of 0.025 after Bonferroni Correction.

and Lexical Word Variation, for both the EFL and ENL groups. The histograms indicate that the ENL group outperforms the EFL group in terms of Corrected TTR, while the performance of both groups is comparable in terms of Lexical Word Variation.

The results in Table 5 show that the mean rank of Corrected TTR for the ENL group (441.36) is significantly higher than that for the EFL group (230.07), as confirmed by the Mann–Whitney U test results ($U = 11828.5$, $Z = -14.07$, $p = 0.000$) at the 0.025 significance level with a substantial effect size ($r = 0.57$). This result indicates that ENL writing contains a significantly higher proportion of distinct words, reflecting greater lexical diversity compared to CUSs' EFL writing. The Lexical Word Variation index shows no significant difference between the EFL group (mean rank = 311.34) and the ENL group (mean rank = 278.82), $U = 35664.00$, $Z = 2.17$, and $p = 0.030$ with a negligible effect size (r = 0.09), suggesting that CUSs and ENL writers use a comparable proportion of distinct lexical words in their total lexical output.

While Lexical Word Variation showed no significant difference, Corrected TTR was significantly lower in EFL writing. These findings suggest that EFL writing may exhibit lower overall lexical variation than ENL writing, particularly as indicated by the significant difference in Corrected TTR. This conclusion is further supported by the fact that Corrected TTR serves as a more comprehensive and fundamental index of lexical variation. While Lexical Word Variation reflects the degree of variation specifically within lexical words, Corrected TTR captures a broader scope of lexical variation by considering the overall distribution of unique words relative to text length. Given its ability to account for variations across the entire lexical repertoire rather than a subset of words, Corrected TTR provides a more robust and representative assessment of lexical variation in writing.

Lexical variation reflects vocabulary breadth and the degree of lexical diversity in writing. The gap in lexical variation underscores the limited lexical repertoire of EFL learners. This finding aligns with previous research (e.g., Eckstein and Ferris 2018; Nakamaru 2010; Staples and Reppen 2016). For instance, Eckstein and Ferris (2018) reported similar outcomes when comparing L1 and L2 texts, showing that ENL writing consistently had a higher Corrected TTR than EFL/ESL writing, while differences in Lexical Word Variation were not statistically significant.

This disparity may arise from several interrelated factors. Limited exposure to authentic lexical input constrains CUSs' ability to internalize and apply broader vocabulary, as textbooks often prioritize high-frequency words over those found in authentic texts like academic articles or literature (Hyland 2014). Writing tasks in EFL classrooms typically emphasize coherence and accuracy, discouraging lexical experimentation and variety (Hernández 2022). Additionally, vocabulary instruction often focuses on rote memorization rather than strategies like using synonyms or word families, further limiting lexical variation (Kyle and Crossley 2015).

To enhance lexical variation in Chinese EFL instruction, targeted interventions are essential. For example, exposure to diverse lexical items through enriched input, such as academic texts and authentic media, provides meaningful contexts for vocabulary acquisition (Calvo-Ferrer and Belda-Medina 2021; Myers and Chang 2009). Writing tasks should promote varied vocabulary use, including lexical substitution and creative writing. Explicit instruction on synonyms, paraphrasing, and word families equips students with practical tools for diversification (Kyle and Crossley 2015). Moreover, a classroom culture that fosters risk-taking can encourage students to experiment with broader vocabulary.

Another noteworthy point is that EFL and ENL writing show no significant difference in Lexical Word Variation. This suggests

that CUSs can produce a comparable proportion of distinct lexical words, aligning with studies showing that formulaic sequences and high-frequency chunks dominate both EFL and ENL writing (Duan and Shi 2024; Staples and Reppen 2016). However, similar Lexical Word Variation does not imply equal lexical proficiency, as it overlooks lexical sophistication and contextual appropriateness (Crossley et al. 2012). Thus, Lexical Word Variation should be interpreted alongside other LR indices to better assess EFL vocabulary use.

In summary, the results of this study highlight significant differences between EFL and ENL writing across LR dimensions. While CUSs demonstrate a lexical density comparable to that of ENL writing, their writing exhibits notable gaps in lexical sophistication, underscoring limited vocabulary depth. For lexical variation, although no significant difference was found between EFL and ENL writing in terms of Lexical Word Variation, the mean rank of Corrected TTR for the EFL group was significantly lower than that of the ENL group. Corrected TTR is considered more stable across texts of varying lengths and has been shown to yield more reliable results in previous studies on lexical variation (Covington and McFall 2010; Lu 2012). Given these methodological advantages, CTTR was emphasized in interpreting group differences in lexical variation. Based on this, it can be cautiously concluded that, overall, EFL writing may exhibit lower lexical variation than ENL writing. These findings emphasize the need for targeted pedagogical interventions, including enriched input, explicit vocabulary instruction, and feedback-driven strategies, to enhance EFL learners' lexical development. Addressing these gaps is critical for improving their overall writing proficiency and advancing research on LR in EFL contexts.

Conclusion

This study examined LR in CUSs' EFL writing by comparing it with an ENL corpus. The research aimed to identify key differences in lexical choices and the overall LR in EFL learners compared to native speakers, focusing on quantitative measures of LR. Through corpus-based analysis, this study sought to address the existing gap in understanding the specific areas where Chinese EFL learners struggle with lexical usage in their writing.

Findings show that although CUSs match ENL writers in lexical density, they exhibit notable differences in both lexical sophistication and lexical variation. The high lexical density of CUSs' writing may stem from linguistic transfer from Chinese, a language characterized by a higher proportion of lexical words. In contrast, the gaps in lexical sophistication and variation highlight limited vocabulary depth and breadth among EFL learners, likely influenced by teaching practices that prioritize grammatical accuracy over advanced vocabulary development and task designs that do not incentivize lexical diversity. These disparities underline the need for more targeted instructional interventions to address these specific deficiencies.

This study enhances the understanding of LR in EFL contexts by highlighting key distinctions among lexical dimensions and their pedagogical implications. Its findings emphasize the importance of designing instructional strategies that integrate enriched input, explicit teaching of advanced vocabulary, and feedback-driven writing tasks to foster greater lexical depth and diversity. Beyond its practical implications, this study contributes theoretically by refining and empirically validating a multidimensional framework of LR tailored to EFL writing contexts. By adopting three computationally robust and pedagogically interpretable dimensions, lexical density, sophistication, and variation, the study advances a more nuanced conceptualization of LR that aligns with EFL learners' actual linguistic output. The corpus-based comparative design, incorporating matched genres and task conditions, offers a replicable methodological model for future research on productive vocabulary use. These contributions narrow the gap between quantitative lexical indices and theoretical frameworks of lexical proficiency in second language writing.

However, the present study is limited by its cross-sectional design and focuses on a single genre, argumentative writing, which may constrain the generalizability of its findings. Future research could address these limitations by exploring longitudinal changes in CUSs' LR, examining the effects of pedagogical innovations, and utilizing advanced corpus-based tools to further refine the measurement and understanding of LR in EFL writing.

Data availability

The primary data applied in this research are from the published corpora SWECCL 2.0 (Wen et al. 2008) and LOCNESS (Granger 1998). The processed data will be available upon reasonable request from the corresponding author.

Received: 9 October 2024; Accepted: 14 July 2025;
Published online: 29 July 2025

Notes

1 Kyle (2020) uses the term "lexical diversity" to denote "lexical variation". Since the term "lexical diversity" in other studies (e.g., Jarvis 2013) refers to a broader concept that encompasses lexical density and sophistication, this paper adopts the term "lexical variation".
2 http://www.cohmetrix.com/.
3 https://uclouvain.be/en/research-institutes/ilc/cecl/locness.html.
4 https://www.random.org/.

References

Ai H, Lu X (2010) A web-based system for automatic measurement of lexical complexity. In: 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10), Amherst, MA

Ai H, Lu X (2013) A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. In Díaz-Negrillo A, Thompson P, Ballier N (eds) Automatic treatment and analysis of learner corpus data. Amsterdam: John Benjamins B.V., pp. 249–264, https://doi.org/10.1075/scl.59.15ai

Akbana YE (2015) A Contrastive Interlanguage Analysis of the Highest-Frequency Vocabulary in Advanced and Native English. In: The European Conference on Language Learning 2015, Brighton, United Kingdom

Anthony L (2024) AntWordProfiler (Version 2.2.1) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software/antwordprofiler/

Azodi N, Karimi F, Vaezi R (2014) Measuring the lexical richness of productive vocabulary in Iranian EFL University Students' writing performance. Theory Pract Lang Stud 4(9):1837–1849. https://doi.org/10.4304/tpls.4.9.1837-1849

Bestgen Y (2017) Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. System 69:65–78. https://doi.org/10.1016/j.system.2017.08.004

Bonferroni CE (1935) Il calcolo delle assicurazioni su gruppi di teste [The calculation of insurance on groups of heads]. Studi in Onore del Professore Salvatore Ortu Carboni [Studies in honor of Professor Salvatore Ortu Carboni], 13–60

Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilita [Statistical theory of classes and probability calculation] Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commericiali di Firenze [Publications of the R Upper Institute of Economic and Commercial Sciences of Florence] 8:3–62

Bozdağ FÜ (2019) The use of vertically prepositions in argumentative essays of Turkish EFL Learners. In: Can C, Patsala P, Tatsioka Z (eds) Language in Focus: Contemporary Means and Methods in ELT and Applied Linguistics. Cappadocia: Lychnos-Printhouse, pp. 105–120

British Council (2023) Writing Band Descriptors. British Council Global. Retrieved September 9, 2023 from https://takeielts.britishcouncil.org/sites/default/files/ielts_writing_band_descriptors.pdf

Calvo-Ferrer JR, Belda-Medina J (2021) The effect of multiplayer video games on incidental and intentional L2 vocabulary learning: the case of Among Us. Multimodal Technol Interact 5(12):80. https://doi.org/10.3390/mti5120080

Camiciottoli BC (2007) The Language of Business Studies Lectures: A Corpus-Assisted Analysis. Amsterdam: John Benjamins Publishing. https://doi.org/10.1075/pbns.157

Caro K, Mendinueta NR (2017) Lexis, lexical competence and lexical knowledge: a review. J Lang Teach Res, 8(2):205-213. https://doi.org/10.17507/jltr.0802.01

Castello E (2008) Text complexity and reading comprehension tests, vol. 85. New York: Peter Lang. https://doi.org/10.3726/978-3-0351-0276-5
Chai N, Wannaruk A, Lian A (2015) A corpus-based study on Chinese EFL learners' use of transitive constructions with neutral participants. Theory Pract Lang Stud 5(9):1778–1790. https://doi.org/10.17507/tpls.0509.03
Cobb T (2018) Web VocabProfile. http://www.lextutor.ca/vp/
Cohen J (1988) Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale: Erlbaum
Covington MA, McFall JD (2010) Cutting the Gordian knot: the moving-average type–token ratio (MATTR). J Quant Linguist 17(2):94–100. https://doi.org/10.1080/09296171003643098
Crossley SA, Salsbury T, McNamara DS (2012) Predicting the proficiency level of language learners using lexical indices. Lang Test 29(2):243–263. https://doi.org/10.1177/0265532211419331
Daller H, Milton J, Treffers-Daller J (2007) Modelling and assessing vocabulary knowledge. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511667268
Duan S, Shi Z (2024) A longitudinal study of formulaic sequence use in second language writing: Complex dynamic systems perspective. Lang Teach Res 28(2):497–530. https://doi.org/10.1177/13621688211002942
Eckstein G, Ferris D (2018) Comparing L1 and L2 texts and writers in first-year composition. TESOL Q 52(1):137–162. https://doi.org/10.1002/tesq.376
Engber CA (1995) The relationship of lexical proficiency to the quality of ESL compositions. J Second Lang Writ 4(2):139–155. https://doi.org/10.1016/1060-3743(95)90004-7
Foster P, Tavakoli P (2009) Native speakers and task performance: comparing effects on complexity, fluency, and lexical diversity. Lang Learn 59(4):866–896. https://doi.org/10.1111/j.1467-9922.2009.00528.x
Gan Q (2024) Different registers, different grammars in second language production? The dative alternation in spoken and written Chinese learner English. Lingua 309:103790. https://doi.org/10.1016/j.lingua.2024.103790
Geng H, Yang Y (2021) Lexical richness in english travel guidebooks by EEL and ENL writers. In: The 7th Malaysia International Conference on Foreign Languages 2021 (MICFL2021), Kuala Lumpur, Malaysia. http://micfl2021.upm.edu.my/
Goodfellow R, Lamy M-N, Jones G (2002) Assessing learners' writing using lexical frequency. ReCALL 14(1):133–145. https://doi.org/10.1017/S0958344002001118
Granger S (1998) The computer learner corpus: a versatile new source of data for SLA research. In: Granger S (ed), Learner English on Computer. London: Routledge, pp. 3–18. https://doi.org/10.4324/9781315841342-1
Ha HS (2019) Lexical richness in EFL undergraduate students' academic writing. Engl Teach 74(3):3–28. https://doi.org/10.15858/engtea.74.3.201909.3
Halliday MAK (1989) Spoken and written language, 2ed ed. Oxford: Oxford University Press
Haynes W (2013) Bonferroni correction. In: Dubitzky W, Wolkenhauer O, Cho KH, Yokota H (eds), Encyclopedia of systems biology. New York: Springer, pp. 154–154, https://doi.org/10.1007/978-1-4419-9863-7_1213
Heng R, Pu L, Liu X (2023) The effects of genre on the lexical richness of argumentative and expository writing by Chinese EFL learners. Front Psychol 13:1082228. https://doi.org/10.3389/fpsyg.2022.1082228
Hernández AIM (2022) Using a free corpus tool for time-efficient feedback on english as a foreign language reports. Miscelánea J Engl Am Stud 66:13–39. https://doi.org/10.26754/ojs_misc/mj.20227356
Hyland K (2014) Activity and evaluation: reporting practices in academic writing. In: Flowerdew J (ed), Academic discourse. London: Routledge, pp. 115–130
Ibrahim EHE, Muhamad AJ, Esa Z (2019) A comparison of lexical richness in L2 written productions. Int J Emerg Technol Learn, 14(20):174–181. https://doi.org/10.3991/ijet.v14i20.11467
Jarvis S (2013) Capturing the diversity in lexical diversity. Lang Learn 63(s1):87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x
Jarvis S (2017) Grounding lexical diversity in human judgments. Lang Test 34(4):537–553. https://doi.org/10.1177/0265532217710632
Kim M, Crossley SA, Kyle K (2018) Lexical sophistication as a multidimensional phenomenon: relations to second language lexical proficiency, development, and writing quality. Mod Lang J 102(1):120–141. https://doi.org/10.1111/modl.12447
Kojima M, Yamashita J (2014) Reliability of lexical richness measures based on word lists in short second language productions. System 42:23–33. https://doi.org/10.1016/j.system.2013.10.019
Kong M, Ungsitipoonporn S, Patpong P (2023) Lexical richness in Thai textbooks published in China. Eurasia J Educ Res 106(106):211–231. https://doi.org/10.14689/ejer.2023.106.013
Kyle K (2020) Measuring lexical richness. In: Webb S (ed) The Routledge handbook of vocabulary studies. London: Routledge, pp. 454–476. https://doi.org/10.4324/9780429291586-29
Kyle K, Crossley S, Berger C (2018) The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. Behav Res methods 50(3):1030–1046. https://doi.org/10.3758/s13428-017-0924-4
Kyle K, Crossley SA (2015) Automatically assessing lexical sophistication: indices, tools, findings, and application. TESOL Q 49(4):757–786. https://doi.org/10.1002/tesq.194
Kyle K, Crossley SA, Jarvis S (2021) Assessing the validity of lexical diversity indices using direct judgements. Lang Assess Q 18(2):154–170. https://doi.org/10.1080/15434303.2020.1844205
Lan, G, Lucas, K, Sun, Y (2019) Does L2 writing proficiency influence noun phrase complexity? A case analysis of argumentative essays written by Chinese students in a first-year composition course. System, 85:102116. https://doi.org/10.1016/j.system.2019.102116
Laufer B (1991) The development of L2 lexis in the expression of the advanced learner. Mod Lang J 75(4):440–448. https://doi.org/10.2307/329493
Laufer B, Nation I (1995) Lexical richness in L2 written production: can it be measured. Appl Linguist 16(3):307–322. https://doi.org/10.1093/applin/16.3.307
Lei S, Yang R (2020) Lexical richness in research articles: corpus-based comparative study among advanced Chinese learners of English, English native beginner students and experts. J Engl Acad Purp 47:100894. https://doi.org/10.1016/j.jeap.2020.100894
Lemmouh Z (2010) The relationship among vocabulary knowledge, academic achievement and the lexical richness in writing in Swedish university students of English [Doctoral dissertation, Stockholm University]. Stockholm. https://su.diva-portal.org/smash/record.jsf?pid=diva2%3A351417anddswid=6867
Li R, Liu K, Cheung AK (2025) Exploring the impact of intermodal transfer on simplification: Insights from signed language interpreting, subtitle translation, and native speech in TED talks. Lang Sci 110:101726. https://doi.org/10.1016/j.langsci.2025.101726
Li X, Zhang H (2021) Developmental features of lexical richness in English writings by Chinese L3 beginner learners. Front Psychol 12:752950. https://doi.org/10.3389/fpsyg.2021.752950
Linnarud M (1986) Lexis in composition: a performance analysis of Swedish learners' written English. Malmö: CWK Gleerup
Liu N (2021) Verb bias in Chinese and its effect on L2 Chinese learners' writing. Int J Chin Lang Teach 2(2):32–43. https://doi.org/10.46451/ijclt.2021.10.03
Lu X (2012) The relationship of lexical richness to the quality of ESL learners' oral narratives. Mod Lang J 96(2):190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x
Lu X, Ai H (2015) Syntactic complexity in college-level English writing: differences among writers with diverse L1 backgrounds. J Second Lang Writ 29:16–27. https://doi.org/10.1016/j.jslw.2015.06.003
Mahardika R (2015) Comparing lexical richness and lexical cohesion of descriptive essays written by student with different exposures to English/Riska Mahardika [Doctoral dissertation, Universitas Negeri Malang]. Malang
Malvern D, Richards B (2013) Measures of lexical richness. In: Chapelle, C (ed) The encyclopedia of applied linguistics. Hoboken, NJ: John Wiley and Sons, Inc., pp. 3622–3627 https://doi.org/10.1002/9781405198431
McNamara DS, Crossley SA, McCarthy PM (2010) Linguistic features of writing quality. Writ Commun 27(1):57–86. https://doi.org/10.1177/0741088309351547
McNamara DS, Graesser AC, McCarthy PM, Cai Z (2014) Automated evaluation of text and discourse with Coh-Metrix. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511894664
Mirzaei A, Domakani MR, Rahimi S (2016) Computerized lexis-based instruction in EFL classrooms: using multi-purpose LexisBOARD to teach L2 vocabulary. ReCALL 28(1):22–43. https://doi.org/10.1017/S0958344015000129
Mu H (2011) Corpus-based research on content validity of CET-4 cloze. Comput Assisted Foreign Lang Educ 4, 66–70, https://doi.org/10.3969/j.issn.1001-5795.2011.04.012
Muñoz Acevedo, DO (2014) An analysis of gains in lexical richness in the writing of instructed intermediate EFL learners, Doctoral dissertation, University of Reading. https://ethos.bl.uk/OrderDetails.do?uin=uk.bl.ethos.632857
Myers JL, Chang S-F (2009) A multiple-strategy-based approach to word and collocation acquisition. Int Rev Appl Linguist Lang Teach 47(2):179–207. https://doi.org/10.1515/iral.2009.008
Nakamaru S (2010) Lexical issues in writing center tutorials with international and US-educated multilingual writers. J Second Lang Writ 19(2):95–113. https://doi.org/10.1016/j.jslw.2010.01.001
Pan K, Yang L, Lu X (2024) Lexical fluency and richness in L2 attrition among Chinese college English learners. Foreign Lang Teach Res 56(03):394-404+479-480. https://doi.org/10.19923/j.cnki.fltr.2024.03.008
Paradowski MB (2015) Productive foreign language skills for an intercultural world. Lausanne: Peter Lang Publishing Group. https://doi.org/10.3726/978-3-653-03913-9
Qin, X, Wen, Q (2007) EFL writing of college English majors in China: a developmental perspective. Beijing: China Social Science Press
Read J (2000) Assessing vocabulary. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511732942
Richards B, Malvern D (1997) Quantifying lexical diversity in the study of language development. In: The New Bulmershe Papers, University of Reading

Shah SK, Gill AA, Mahmood R, Bilal M (2013) Lexical richness, a reliable measure of intermediate L2 learners' current status of acquisition of English language. J Educ Pract 4(6):42–47
Shao C-Z (2019) A corpus-based exploration into the sentence development in EFL writing in China. In: 2019 International Conference on Applied Mathematics, Modeling, Simulation and Optimization (AMMSO 2019). https://doi.org/10.12783/dtcse/ammso2019/30149
Shi Y, Lei L (2022) Lexical richness and text length: an entropy-based perspective. J Quant Linguist 29(1):62–79. https://doi.org/10.1080/09296174.2020.1766346
Šišková Z (2012) Lexical richness in EFL students' narratives. Lang Stud Working Pap 4:26–36
Staples S, Reppen R (2016) Understanding first-year L2 writing: a lexico-grammatical analysis across L1s, genres, and language ratings. J Second Lang Writ 32:17–35. https://doi.org/10.1016/j.jslw.2016.02.002
Tran TTT, Ma Q (2025) Technology-enhanced self-regulation training: a dynamic training model to facilitate second language Vietnamese learners' self-regulated writing skills. System 130:103625. https://doi.org/10.1016/j.system.2025.103625
Treffers-Daller J, Parslow P, Williams S (2018) Back to basics: how measures of lexical diversity can help discriminate between CEFR levels. Appl Linguist 39(3):302–327. https://doi.org/10.1093/applin/amw009
Waller T (1993) Characteristics of near-native proficiency in writing. In: Ringbom H (ed) Near-native proficiency in English, vol. 2. Turku: Akademi University, pp. 183–293
Wang S, Slater T (2016) Syntactic complexity of EFL Chinese students' writing. Engl Lang Lit Stud 6(1):81–86. https://doi.org/10.5539/ells.v6n1p81
Wei Y (2021) Use of English phrasal verbs of Chinese students across proficiency levels: a corpus-based analysis. Int J TESOL Stud 3(4):25–41. https://doi.org/10.46451/ijts.2021.12.03
Wen Q, Liang M, Yan X (2008) Spoken and written English corpus of Chinese learners (Version 2.0). Beijing: Foreign Language Teaching and Research Press
Woods K, Hashimoto B, Brown EK (2023) A multi-measure approach for lexical diversity in writing assessments: considerations in measurement and timing. Assess Writ 55:100688. https://doi.org/10.1016/j.asw.2022.100688
Wu J, Tissari H (2021) Intensifier-verb collocations in academic English by Chinese learners compared to native-speaker students. Chin J Appl Linguist 44(4):470–487. https://doi.org/10.1515/CJAL-2021-0030
Xie Y, Shen Y (2015) A study of the relationships between lexical richness and writing quality: taking the English majors at Guangxi University as an example 2015 International Conference on Social Science, Education Management and Sports Education, https://doi.org/10.2991/ssemse-15.2015.419
Yan H, Li Y (2019) Beyond length: investigating dependency distance across L2 modalities and proficiency levels. Open Linguist 5(1):601–614. https://doi.org/10.1515/opli-2019-0033
Yang Y, Yap NT, Mohamad Ali A (2022) A corpus-based comparative study on syntactic complexity in University Students' EFL writing in Southwestern China: a case of Pu'er University. World J Engl Lang 12(8):172–180. https://doi.org/10.5430/wjel.v12n8p172
Yang Y, Yap NT, Mohamad Ali A (2023) Predicting EFL expository writing quality with measures of lexical richness. Assess Writ 57:100762. https://doi.org/10.1016/j.asw.2023.100762
Yang Y, Zhang F, Zhang S (2022) An overview on dimensions, measures, and indices of lexical richness in English writing. Foreign Lang Transl 29(4):80–85. https://doi.org/10.19502/j.cnki.2095-9648.2022.04.006
Yang Y, Zheng Z (2024) A refined and concise model of indices for quantitatively measuring lexical richness of Chinese university students' EFL writing. Contemp Educ Technol 16(3):ep513. https://doi.org/10.30935/cedtech/14707
Ye C, Xin H (2018) A corpus-based study of stylistic features of English versions of the report on the work of the government in recent ten years. Mod Linguist 6(4):588–598. https://doi.org/10.12677/ml.2018.64069
Yu G (2010) Lexical diversity in writing and speaking task performances. Appl Linguist 31(2):236–259. https://doi.org/10.1093/applin/amp024
Zeng X, Shirai Y, Chen X (2023) A corpus-based study of the acquisition of the English progressive by L1 Chinese learners: from prototypical activities to marked statives. Linguistics 61(3):749–778. https://doi.org/10.1515/ling-2020-0199
Zhang B, Yang L (2019) Construction of CivDEAP corpus and its application in academic research and teaching practice in civil engineering. Int J Soc Policy Educ 1(2):135–142
Zhang H, Chen M, Li X (2021) Developmental features of lexical richness in English writings by Chinese beginner learners. Front Psychol 12:665988. https://doi.org/10.3389/fpsyg.2021.665988
Zhang X, Yang H (2021) Gender voices in Chinese university students' English writing: a corpus study. Linguist Educ 64:100935. https://doi.org/10.1016/j.linged.2021.100935
Zheng W (2025) Lexical richness viewed through lexical diversity, density, and sophistication. Digital Scholarship Humanit, fqaf023, https://doi.org/10.1093/llc/fqaf023
Zheng Y (2015) The long-term development of foreign language vocabulary in the framework of dynamic systems theory. Shanghai: Fudan University Press
Zheng Y (2016) The complex, dynamic development of L2 lexical use: a longitudinal study on Chinese learners of English. System 56:40–53. https://doi.org/10.1016/j.system.2015.11.007
Zheng Y, Feng Y (2017) A Dynamic Systems study on Chinese EFL learners' syntactic and lexical complexity development. Mod Foreign Lang 40(01):57-68+146
Zhu H, Wang J (2013) Developmental features of lexical richness in English writing: a self-built corpus-based longitudinal study. Foreign Language World 33(6):77–86
Zhu Z, Pu X (2024) A corpus-based study on the development of lexical diversity and lexical sophistication of argumentative writings. Acad J Humanit Soc Sci 7(8):1–8. https://doi.org/10.25236/AJHSS.2024.070801

Author contributions

YY: Conceptualization, Methodology, Formal analysis, Investigation, Resources, Data Curation, Writing - Original Draft, Writing - Review and Editing. XH: Conceptualization, Methodology, Validation, Formal analysis, Data Curation, Writing - Original Draft, Writing - Review and Editing, Supervision.

Competing interests

The authors declare no competing interests.

Ethical approval

This article does not contain any studies with human participants performed by any of the authors.

Informed consent

This study did not involve human participants. The analysis was based on publicly available corpora. Therefore, informed consent was not applicable.

Additional information

Correspondence and requests for materials should be addressed to Xubo He.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.