Contemporary Educational Technology, 2024, **16**(3), ep513
ISSN: 1309-517X (Online)

OPEN ACCESS

Research Article

A refined and concise model of indices for quantitatively measuring lexical richness of Chinese university students' EFL writing

Yang Yang [1,2*]
0000-0003-3114-0682

Ze Zheng [3]
0009-0006-3085-7453

[1] College of International Studies, Southwest University, Chongqing, CHINA
[2] Faculty of Modern Languages and Communication, Universiti Putra Malaysia, Serdang, Selangor, MALAYSIA
[3] Independent Researcher, CHINA
[*] Corresponding author: yangvictoryang@swu.edu.cn

Citation: Yang, Y., & Zheng, Z. (2024). A refined and concise model of indices for quantitatively measuring lexical richness of Chinese university students' EFL writing. *Contemporary Educational Technology, 16*(3), ep513. https://doi.org/10.30935/cedtech/14707

ARTICLE INFO

Received: 5 Jan 2024
Accepted: 30 May 2024

ABSTRACT

In the existing literature, scholars have proposed various indices to measure the lexical richness (LR) of English as a foreign language (EFL) writing. However, there are currently issues of redundant indices and inconsistent usage. Attempting to address the research question of which indices are the most sensitive and effective ones to distinguish between different grade levels of Chinese university students' EFL writing, this study aims to put forward a refined and concise model of indices that can truthfully reflect LR in EFL writing. A total of 180 compositions were selected from a Chinese EFL learner corpus: *Spoken and written English corpus of Chinese learners*. Scores of 28 LR indices of these compositions were computed using the software *Lexical Complexity Analyzer*, *MATTR*, and *Coh-Metrix*. One-way ANOVA or Welch's ANOVA, depending on the variable's homogeneity of variances, was conducted for each index. Two criteria were applied to determine which index of a measure should be included in the refined model: whether the difference of an index is significant among different grade levels and the effect size of ANOVA. Based on the quantitative results of ANOVAs and qualitative human judgment based on literature, six indices of the six LR measures were included in the refined model: lexical density, lexical sophistication-I, verb sophistication-II, number of different words-expected sequence 50, corrected TTR, and squared verb variation-I. This refined model addresses the issues of redundancy and inconsistency in previous studies, providing a more accurate and efficient tool for assessing LR in EFL writing.

Keywords: refined model, lexical richness, index, EFL writing, training and testing

INTRODUCTION

Lexical richness (LR) refers to the extent of sophistication in the productive vocabulary of a language user or learner. Evaluating writing proficiency through it has been widely regarded as highly effective (Fan et al., 2023). Within the framework of English as a foreign language (EFL) teaching and learning, LR is regarded as an important indicator of EFL proficiency (Malvern & Richards, 2013). Regarding the connection between LR and the quality of EFL writing, many researchers (e.g., Geng & Yang, 2021; Kojima & Yamashita, 2014; Treffers-Daller et al., 2018; Xie & Shen, 2015) have reported that there is a significant correlation between them.

Traditionally, there are four dimensions of LR: lexical originality, density, sophistication, and variation (Gregori-Signes & Clavel-Arroitia, 2015). Besides, lexical errors have also been regarded as another dimension of LR and it was claimed that measuring LR should consider lexical variation including errors, lexical variation

excluding errors, and the ratio of lexical errors (Zhang et al., 2021). Read (2000) claimed that good writing should have the following characteristics:

(1) a relatively high proportion of content words,

(2) utilizing less common vocabulary appropriate to the subject matter and tone,

(3) a rich vocabulary, not reusing a limited number of words, and

(4) fewer lexical errors.

These characteristics are the four dimensions for evaluating LR, namely lexical density, sophistication, variation, and errors. Based on these various LR dimensions, some related measures, indices, and corresponding calculation methods were proposed in the literature.

However, there are controversies about the dimensions, measures, and indices of LR. Read (2000) found that lexical originality was not suitable for evaluating learners' lexical development, while lexical errors were an effective measurement of LR, and learners' vocabulary acquisition could be observed through different types of errors. Although Read (2000) included lexical density as one of LR dimensions, Huang and Qian (2003) questioned the accuracy and validity of it as a measurement dimension; it was found that lexical density proved ineffective in discerning variations in vocabulary usage among distinct learners. Huang and Qian (2003) believed that vocabulary density is not sensitive to the development of learners' lexical ability and is not closely related to learners' overall language and writing proficiency.

What is more, there exist different calculation methods for the same LR measure. For instance, lexical sophistication is calculated as the proportion of sophisticated lexical terms compared to the overall count of lexical words, while it is also computed by dividing the count of sophisticated word types by the total count of word types. This has led to a bewildering array of LR indices in the literature. Within the scope of LR studies conducted in China, measures and indices of LR are used indiscriminately and inconsistently, especially for LR in Chinese university students' (CUSs) EFL writing. For example, Li (2021) studied LR in CUSs' EFL writing from the following four dimensions: lexical density, sophistication, variation, and originality, and she calculated the lexical variation using the index standardized TTR. Nonetheless, other scholars, such as Wan (2010) and Zhang (2021), investigated LR excluding measures of lexical density and originality but incorporating lexical errors, and employed the Uber index for assessing lexical variation. Therefore, it is necessary to sort out the indices in the literature and put forward a refined model of indices that is suitable to and can truthfully reflect EFL writing proficiency of CUSs (Yang et al., 2022). This paper reviews most of LR measures and indices in the literature, classifies them, and tries to find out which indices are sensitive and effective to the difference and development of EFL writing proficiency of CUSs.

LITERATURE REVIEW

Measures & Indices of Lexical Richness

The lexical knowledge can be assessed from two aspects (Yang et al., 2023). The first aspect is the lexical width, which describes how many words a learner can master. The measures of it are operationalized as lexical density, variation, and originality. The second dimension pertains to lexical depth, which delves into the extent of a learner's grasp of vocabulary. This facet is manifested through the level of lexical sophistication as well as the frequency of lexical errors present in their writing. In essence, lexical depth assesses the depth of a learner's lexical knowledge and proficiency in utilizing vocabulary accurately and effectively.

Lexical density

The term "lexical density" was introduced by Ure (1971). It is determined by calculating the proportion of content words, also known as lexical words, within a given text, relative to the total number of words present. This measurement excludes functional or grammatical words, focusing solely on the content-bearing vocabulary.

Lexical sophistication

Lexical sophistication is the degree to which sophisticated or advanced words are used in a certain oral or written production of a language learner. It is measured by "the proportion of relatively unusual or advanced

words in the learner's text" (Read, 2000, p. 203). At the operational level, lexical sophistication can be quantified by assessing the proportion of sophisticated vocabulary, whether in terms of individual words, word types, lexical words, or lexical word types within the text, as indicated by Hyltenstam (1988). Hyltenstam (1988) and Linnarud (1986) computed lexical sophistication by determining the proportion of sophisticated lexical terms (N_{slex}) relative to the total count of lexical words (N_{lex}). Ai and Lu (2010) termed Linnarud's method in their tool *Lexical Complexity Analyzer* (Lu, 2012) as "lexical sophistication-I" (LS1).

Laufer (1994) and Laufer and Nation (1995) launched the lexical frequency profile (FLP). FLP also introduced a method for computing lexical sophistication, which involves comparing the count of sophisticated word types (T_s) to the total number of word types (T). This approach to measuring lexical sophistication is referred to as lexical sophistication-II (LS2) within the *Lexical Complexity Analyzer*. Other scholars utilize the proportion of sophisticated word types within specific parts of speech as an indicator of lexical sophistication, as exemplified by Harley and King's (1989) examination of verb sophistication. They derived their verb sophistication-I (VS1) index by determining the ratio of less common verb types, excluding those found among the most frequent 200 verbs, to the total count of verbs. To mitigate the impact of sample size, some modified and corrected indices of VS1 were proposed: corrected VS1 (CVS1; Chaudron & Parker, 1990) and verb sophistication-II (VS2; Wolfe-Quintero et al., 1998).

Lexical variation

Lexical variation signifies the width of vocabulary knowledge demonstrated by a language learner in their language usage. One common method of gauging lexical variation within a text is by counting the number of different words, also known as word types or NDW. However, a major drawback of relying solely on NDW is its susceptibility to the length of the text. To address this issue, several standardized adaptations of NDW have been suggested. For example, NDW-50 (NDW in the first 50 words) tallies the count of unique word types appearing within the initial 50 words of a text. NDW-ER50 (NDW in the expected random 50 words) computes the average count of word types across 10 random 50-word samples taken from the text. Lastly, NDW-ES50 (NDW in the expected sequence of 50 words) determines the average number of word types across 10 random 50-word sequences within the text.

Another method for assessing lexical variation is the type-token ratio (TTR), which represents the ratio of unique word types (T) to the total number of words (N) in a given text. Then, MSTTR (mean segmental TTR) was introduced, which divides a text into segments of a specified word count and computes the average TTR across all segments. While MSTTR effectively addresses the sample size issue encountered with TTR, it may result in data wastage as it discards any remaining words. To mitigate this concern, moving average TTR (MATTR; Covington & McFall, 2010) was developed. Other variations of TTR include corrected TTR (CTTR), bilogarithmic TTR (log TTR), root TTR (RTTR), and Uber index.

Other adaptations of TTR aim to assess the variation within specific word categories, such as lexical words and those belonging to particular parts of speech. Certain investigations have evaluated lexical word variation by calculating the ratio of unique lexical word types to the total count of lexical words within a text (Engber, 1995; Hyltenstam, 1988; Linnarud, 1986). Harley and King (1989) investigated verb variation by determining the ratio of unique verb types to the total count of verbs within a text, identified as verb variation-I in *Lexical Complexity Analyzer*. To address the influence of varying sample sizes, Chaudron and Parker (1990) introduced the squared VV1 (SVV1) index, while Wolfe-Quintero et al. (1998) proposed the corrected VV1 (CVV1) index. In contrast to Harley and King's (1989) approach to quantifying verb variation, McClure (1991) determined it as the ratio of unique verb types to the total count of lexical words. Similarly, she explored variation in nouns, adjectives, adverbs, and modifiers using the indices verb variation-II (VV2), noun variation (NV), adjective variation (AdjV), adverb variation (AdvV), and modifier variation (ModV).

In addition to the above indices, a curve-fitting approach was employed to measure lexical variation. As an illustration, Malvern and Richards (1997) introduced the D measure to signify the extent of lexical variation within a text. However, D measure was soon replaced by a more solid and reliable measure of its adaption, and a computer program called *vocd* (McKee et al., 2000) was developed to automatically calculate its value. To differentiate D measure proposed by Malvern and Richards (1997), Mckee et al.'s (2000) D measure is called vocd-D in the literature (McCarthy & Jarvis, 2007, 2010; Šišková, 2012). So far, none of the aforementioned measures of lexical variation considers the structure of a text (Šišková, 2012). To fill this gap, McCarthy and

Jarvis (2010) presented the measure of textual lexical diversity (MTLD), which is computed as the average length of consecutive word sequences with a specified TTR value.

Lexical originality

The concept of lexical originality was initially introduced by Laufer (1991). It assesses the proficiency of a language learner or user relative to their peers within the same writing cohort (Laufer & Nation, 1995). It represents the count of words unique to an individual writer, quantified as the percentage of distinct words in a specific piece of writing that are absent in other compositions from the same group. Laufer and Nation (1995) contended that lexical originality lacks reliability as an LR measure, as it is influenced not only by the language usage of the individual but also by that of their peers within the same group.

Similarly, Read (2000) asserted that lexical originality is an inadequate indicator of the lexical proficiency of English as a second language (ESL) learners. Thus, it is not strong on the practicability and generalizability of this measure. Besides, the compositions of the present study are not from the same group, so lexical originality is not applicable here. Thus, it will not be considered a measure of LR in this study.

Lexical errors

Engber (1995) introduced the term lexical errors as another element of measuring LR. In his study, Read (2000) incorporated lexical errors as one of the measures of LR. Hawkey and Barker (2004) have similarly observed that lexical errors serve as a significant indicator of writing quality. In Read's (2000) model of LR, lexical error is one of the important dimensions of LR besides the other three: lexical density, sophistication, and variation.

However, what is different from the other three dimensions is that there is no relevant software to automatically identify the error types and numbers in a text. Lexical errors need to be qualitatively analyzed and manually identified. The current study seeks to clarify and develop a refined model of LR indices that can be automatically and quantitatively calculated by the software. Besides, numerous researchers (e.g., Housen & Kuiken, 2009; Housen et al., 2012; Michel, 2017) have identified accuracy, complexity, and fluency as the three key dimensions of second language (L2) performance and proficiency and based on this division, linguistic errors are under the dimension of accuracy. Thus, the dimension of lexical errors is excluded as a measure of LR in this study.

To conclude, all the measures and indices investigated in this study are summarized in **Table 1**.

"Training & Testing" Method

The "training and testing" method is originally from the field of machine learning and data mining. In the realm of machine learning, when the objective is to construct a model for predicting test data, it is common practice to partition the data into a training set and a testing set. The training dataset is used to fit the model and the testing dataset is used to test the model. The "training and testing" method is borrowed to be applied in the field of language studies (see McNamara et al., 2014, p. 167). For example, if a researcher wants to investigate the verb sophistication in the writing of EFL learners, the researcher may be confused as to which index of verb sophistication to choose: VS1, CVS1, or VS2 (see **Table 1**). Based on the notion of "training and testing", the researcher can use part of the data to test which index is the best one to investigate verb sophistication, and then use this index to complete the research.

Inspired by the notion of "training and testing", this study selects representative CUSs' EFL writing as the data to verify which of the different indices under the same LR measure can truly reflect CUSs' performance of this measure in their EFL writing. The expectation is that the findings of this study will offer a dependable model of indices for the future study of CUSs' LR in their EFL writing.

Research Question

Based on the gap mentioned before and the notion of "training and testing", the research question of the present study is, as follows: Which indices, among different indices under the same LR measures, are the most sensitive and effective ones to distinguish between different grade levels of CUSs' EFL writing?

Table 1. LR measures & indices investigated in this study

Dimension	Measure	Code	Index
Lexical density	Lexical density	LD	Lexical density
Lexical sophistication	Lexical sophistication	LS1	Lexical sophistication-I
		LS2	Lexical sophistication-II
	Verb sophistication	VS1	Verb sophistication-I
		CVS1	Corrected VS1
		VS2	Verb sophistication-II
Lexical variation	Number of different words	NDW	Number of different words
		NDWZ-50	NDW (first 50 words)
		NDW-ER50	NDW (expected random 50 words)
		NDW-ES50	NDW (expected sequence 50 words)
	Type/token ratio	TTR	Type/token ratio
		MSTTR-50	Mean segmental TTR (50 words)
		MATTR-50	Mean average TTR (50 words)
		CTTR	Corrected TTR
		RTTR	Root TTR
		LogTTR	Bilogrithmic TTR
		Uber	Uber Index
		MTLD	Measure of textual lexical diversity
		vocd-D	vocd-D
	Lexical word variation	LWV	Lexical word variation
		NV	Noun variation
		VV1	Verb variation-I
		SVV1	Squared VV1
		CVV1	Corrected VV1
		VV2	Verb variation-II
		AdjV	Adjective variation
		AdvV	Adverb variation
		ModV	Modifier variation

METHODOLOGY

Research Design

This study employed a mixed-method research design that integrates both quantitative and qualitative analyses. The quantitative aspect involved using one-way ANOVA and Welch's ANOVA to statistically analyze LR indices of EFL writing samples from CUSs across four grade levels. The qualitative component included a thorough examination of the collected data to interpret the context and implications of the quantitative findings. Data collection involved stratified sampling from the *spoken and written English corpus of Chinese learners version 2.0* (SWECCL 2.0; Wen et al., 2008), followed by data processing with various LR analyzing software. The subsequent data analysis applied ANOVA to determine the effectiveness and sensitivity of different LR indices in distinguishing between grade levels.

Data Collection

In this study, the writing samples of CUSs were from EFL learner corpus: SWECCL 2.0. The stratified sampling method was adopted for sampling CUSs' writing samples. Four strata were classified with the SWECCL: grades 1, 2, 3, and 4, and compositions were randomly sampled from each stratum. Statistical power analysis software, *G*Power*[1] (Faul et al., 2007, 2009), was utilized to determine the appropriate sample size.

Corpus description

SWECCL 2.0 is a learner corpus that includes compositions of students from a total of 34 distinct universities situated in China. It includes two sub-corpora: spoken English corpus of Chinese learners (SECCL) and written English corpus of Chinese learners (WECCL). WECCL consists of 4,950 compositions penned by English majors and a portion of non-English majors hailing from over 20 universities with different types and levels all over the country. The writing tasks were completed on paper; then the compositions were collected

[1] https://www.psychologie.hhu.de/arbeitsgruppen/allgemeine-psychologie-und-arbeitspsychologie/gpower

and recorded on the computer without any changes to their content. The corpus has a wide range of sources, which can accurately reflect the real situation of students' compositions (Wen et al., 2008). In addition, the variety of writing tasks makes sure that the corpus can well reflect CUSs' EFL writing performance.

Sampling method

This study utilized both non-probability and probability sampling methods. For non-probability sampling, the purposive sampling method was employed: the sub-corpus SECCL was excluded in this study since only the written EFL compositions were needed. Then the probability sampling method was used, including a stratified sampling method and a simple random sampling method without replacement. In SWECCL corpus, the software *Sub-Corpus Generator* was provided. It can be used to generate specific sub-corpus of WECCL with specific variables, such as prompt and genre of the writing, as well as major and grade level of the writer. In this study, the variable of grade was controlled, and four sub-corpora were generated: writing samples of grades 1, 2, 3, and 4, which were written by freshmen, sophomores, juniors, and seniors of CUSs, respectively. There were 1,054, 2,075, 1,108, and 121 compositions, respectively in these four sub-corpora. The stratified sampling method was applied to randomly sample compositions in each sub-corpus. When conducting the stratified sampling method, firstly, the compositions in these four sub-corpora were coded with numbers. For example, the compositions in group grade 1 were coded with the numbers one to 1054. Then certain random numbers were generated by using *Random Generator*[2] for each sub-corpus. The corresponding compositions were chosen according to the random numbers.

Sample size determination

To find out which index under an LR measure is more sensitive and statistically effective in distinguishing between CUSs' EFL writing of the four Grade levels, a one-way ANOVA or Welch's ANOVA was conducted. A priori power analysis using *G*Power* was performed to establish the sample size for ANOVA with a given significance level, power, and effect size. By choosing a conventionally medium effect size of 0.25[3] (Cohen, 1969), a significance level of 0.05, a conventionally high enough power of 0.8 (Cohen, 1988), and the number of groups as four, the result of priori power analysis showed that a total sample size of 180 is needed to reach aforementioned effect size and power, which means that 45 compositions are needed for each grade level.

Data Processing

The 180 writing samples were uploaded to some LR analyzing software or systems, and the values of LR indices in **Table 1** can be calculated and generated.

All LR indices in **Table 1** can be calculated with the system web-based *Lexical Complexity Analyzer*[4] (Ai & Lu, 2010; Lu, 2012), except for MATTR, MTLD, and vocd-D. MATTR can be calculated with software *MATTR* (Covington & McFall, 2010); MTLD and vocd-D can be calculated with the system *Coh-Metrix*[5] (McNamara et al., 2014). It is important to note that the web-based *Lexical Complexity Analyzer* provides two settings: one for British English and one for American English. In his study, American English mode is chosen since it is more preferred and commonly used in the English instruction system in mainland China. 180 compositions were uploaded to LR analyzing systems or software, and the scores of 28 LR indices were automatically calculated.

Table 2 displays the average values and the standard deviations of these indices of the four grade levels.

Data Analysis

After the scores of the 28 LR indices of the four grade levels were obtained, ANOVAs were conducted to find out if the indices can significantly distinguish writing samples of the four grade levels and which one under the same measure does so with the largest effect size. There are some assumptions for conducting one-way ANOVA:

[2] https://www.random.org/

[3] Cohen's guidelines for interpreting effect size of ANOVA changed in the second edition (Cohen, 1988). The latest version of G*Power 3.1 and its manual applies Cohen's (1969) effect size conventions, so Cohen's (1969) effect size convention is followed when determining the effect size in this study.

[4] https://aihaiyang.com/software/lca/

[5] http://www.cohmetrix.com/

Table 2. Descriptive statistics of LR indices by grade

Measure	Index	Mean (standard deviation)			
		Grade 1	Grade 2	Grade 3	Grade 4
Lexical density	LD	0.50 (0.04)	0.51 (0.04)	0.50 (0.04)	0.53 (0.04)
Lexical sophistication	LS1	0.17 (0.05)	0.20 (0.05)	0.24 (0.09)	0.19 (0.04)
	LS2	0.16 (0.03)	0.18 (0.04)	0.16 (0.04)	0.18 (0.04)
Verb sophistication	VS1	0.09 (0.05)	0.10 (0.06)	0.08 (0.05)	0.08 (0.05)
	CVS1	0.35 (0.18)	0.43 (0.24)	0.29 (0.18)	0.38 (0.24)
	VS2	0.31 (0.28)	0.48 (0.46)	0.23 (0.23)	0.40 (0.41)
Number of different words	NDW	129.00 (31.20)	140.20 (31.50)	109.47 (22.20)	161.91 (28.89)
	NDWZ-50	37.93 (2.90)	37.4 (3.22)	38.13 (3.17)	38.31 (3.51)
	NDW-ER50	38.18 (1.87)	39.08 (1.56)	38.59 (1.94)	39.10 (1.70)
	NDW-ES50	37.90 (1.78)	38.56 (1.45)	37.74 (2.12)	38.92 (1.81)
Type/token ratio	TTR	0.47 (0.05)	0.50 (0.06)	0.52 (0.05)	0.46 (0.04)
	MSTTR-50	0.76 (0.04)	0.77 (0.03)	0.77 (0.04)	0.78 (0.03)
	MATTR-50	0.78 (0.03)	0.79 (0.02)	0.78 (0.04)	0.80 (0.03)
	CTTR	5.47 (0.62)	5.85 (0.59)	5.32 (0.59)	6.08 (0.68)
	RTTR	7.74 (0.87)	8.28 (0.84)	7.53 (0.83)	8.59 (0.96)
	LogTTR	0.87 (0.02)	0.88 (0.02)	0.88 (0.02)	0.87 (0.01)
	Uber	18.22 (2.01)	19.86 (2.07)	19.25 (2.55)	19.25 (2.18)
	MTLD	71.01 (16.16)	76.25 (12.94)	73.67 (18.93)	80.87 (17.45)
	vocd-D	76.14 (16.46)	81.6 (13.38)	77.42 (17.91)	87.93 (19.11)
Lexical word variation	LWV	0.68 (0.10)	0.76 (0.13)	0.75 (0.11)	0.66 (0.09)
	NV	0.59 (0.10)	0.63 (0.07)	0.59 (0.13)	0.58 (0.09)
	VV1	16.32 (5.82)	19.85 (6.58)	13.87 (5.46)	20.39 (6.86)
	SVV1	2.81 (0.51)	3.11 (0.53)	2.58 (0.54)	3.14 (0.55)
	CVV1	0.64 (0.07)	0.68 (0.06)	0.67 (0.08)	0.62 (0.06)
	VV2	0.17 (0.03)	0.18 (0.03)	0.17 (0.04)	0.16 (0.03)
	AdjV	0.12 (0.03)	0.13 (0.02)	0.12 (0.03)	0.12 (0.02)
	AdvV	0.10 (0.03)	0.10 (0.02)	0.11 (0.03)	0.10 (0.02)
	ModV	0.22 (0.04)	0.23 (0.03)	0.23 (0.04)	0.22 (0.02)

(1) continuous dependent variable,

(2) categorical independent variable with more than two levels,

(3) independence of observations,

(4) normal distribution of the dependent variable within each category of the independent variable, and

(5) equality of variances.

In this study, the dependent variables are scores of LR indices, so they are continuous data. The independent variable is the grade with four levels: grades 1, 2, 3, and 4. Each writing sample is written by an independent CUS. Thus, the first three assumptions were met in this study.

Besides, according to the central limit theorem (Pólya, 1920), If the sample size exceeds 30, the sample distribution can be considered approximately normal (Chang et al., 2006; Kwak & Kim, 2017). Therefore, the fourth assumption was met.

Finally, to assess variance equality, Levene's test for homogeneity of variances was administered, revealing that the following indices did not have equal variances among the grade levels: LS1, VS2, MATTR-50, MTLD, NV, VV2, AdjV, AdvV, and ModV. For these indices that did not have equal variances, Welch's ANOVA, instead of classic one-way ANOVA, was conducted (Moder, 2007, 2010).

As an equivalent of classic one-way ANOVA, Welch's ANOVA does not assume the homogeneity of variances of the data, since it is not sensitive to heterogeneous variances.

After the one-way ANOVA or Welch's ANOVA was conducted, there were two criteria used to determine which index of an LR measure should be retained.

The first one was the significance of ANOVA result: if an LR index could not significantly distinguish the four grade levels, it would be discarded. The second criterion was the effect size of the test: the largest effect size of an index indicates that this index is the most sensitive, effective, and robust one of an LR measure that can distinguish among EFL writing of different grade levels.

Table 3. Results of ANOVA for LR measure: Lexical density

Index	df	F	Sig.	Effect size
LD	3	3.813	0.011*	0.061

Note. *Mean difference is significant at level of 0.05

Table 4. Results of ANOVA for LR measure: Lexical sophistication

Index	df	F	Sig.	Effect size
LS1	3	9.371	0.000*	0.138
LS2	3	6.493	0.000*	0.100

Note. *Mean difference is significant at level of 0.05

Table 5. Results of ANOVA for LR measure: Verb sophistication

Index	df	F	Sig.	Effect size
VS1	3	2.152	0.095	0.035
CVS1	3	3.694	0.013*	0.059
VS2	3	4.292	0.006*	0.068
VS1	3	2.152	0.095	0.035

Note. *Mean difference is significant at level of 0.05

RESULTS & DISCUSSION

Lexical Density

The result of ANOVA in **Table 3** shows that F(3, 176)=3.813, p<0.05, which indicates that the index LD can significantly distinguish CUSs' EFL writing of different grades. Since there is only one index under this LR measure, LD will be kept in the refined model. Though some researchers (e.g., Engber, 1995; Huang & Qian, 2003) questioned the validity of LD, it is an indispensable dimension of LR and is an indicator of lexical and language proficiency (Bulté & Housen, 2014; Lu, 2012; Nasseri & Thompson, 2021) as well as language complexity (Halliday & Matthiessen, 2004). Halliday and Matthiessen (2004) argued that "written language typically becomes complex by being lexically dense: it packs a large number of lexical items into each clause" (Halliday & Matthiessen, 2004, p. 654).

Lexical Sophistication

Lexical sophistication

The results of ANOVA for two lexical sophistication indices in **Table 4** show that there are statistically significant differences (p<0.05) among the four grade levels for both indices. However, the effect size of LS1 (η^2=0.138) is much larger than that of LS2 (η^2=0.100). Therefore, the index LS1 remains in the refined model of LR indices.

It has been reported that there is a strong correlation between lexical sophistication and writing quality (Crossley & McNamara, 2017; Crossley et al., 2016; Kyle & Crossley, 2016). LS1 was introduced by Linnarud (1986), while LS2 was proposed by Laufer (1994). In determining lexical sophistication, Laufer concerns with word types of all types of parts of speech, while Linnarud (1986) only pays attention to the lexical words, which are more reflective of lexical sophistication, since grammatical words are more stable in writing.

Laufer and Nation (1995) asserted that LS2 serves as a dependable LR index, yet Meara (2005) contested this notion, suggesting that it might not be sensitive enough to detect subtle alterations in vocabulary breadth. In addition, using natural language toolkits (NLTK) and spaCy[6] to calculate LS1 and LS2, Spring and Johnson (2022) reported that LS1 has a higher correlation to the human-rating score of EFL writing than LS2.

Verb sophistication

The results of ANOVA for verb sophistication indices show that there are statistically significant differences (p<0.05) in both CVS1 and VS2 among the four grade levels. However, the effect size of VS2 (η^2=0.068) is larger than that of CVS1 (η^2=0.059) (**Table 5**).

[6] https://spacy.io/

Table 6. Results of ANOVA for LR measure: Number of different words

Index	df	F	Sig.	Effect size
NDW	3	26.218	0.000*	0.309
NDWZ-50	3	0.681	0.565	0.011
NDW-ER50	3	2.786	0.042	0.045
NDW-ES50	3	4.250	0.006*	0.068

Note. *Mean difference is significant at level of 0.05

Table 7. Results of ANOVA for LR measure: Type/token ratio

Index	df	F	Sig.	Effect size
TTR	3	16.087	0.000*	0.215
MSTTR-50	3	2.058	0.108	0.034
MATTR-50	3	5.254	0.002*	0.082
CTTR	3	14.082	0.000*	0.194
RTTR	3	14.032	0.000*	0.193
LogTTR	3	7.953	0.000*	0.119
Uber	3	4.276	0.006	0.068
MTLD	3	2.905	0.036	0.047
vocd-D	3	4.471	0.005*	0.071

Note. *Mean difference is significant at level of 0.05

Thus, the index VS2 remains in the refined model of LR indices. CVS1 and VS2 are transformations of VS1. They are designed to reduce the effect of sample size (Dewi, 2017; Lu, 2012). It is reported that VS2 generated by *Lexical Complexity Analyzer* is more correlated to human-rating scores of EFL writing than the index CVS1 (Spring & Johnson, 2022). In addition, Wang (2018) used both CVS1 and VS2 to analyze the verb sophistication of two college English textbooks, and the result shows that VS2 has a higher degree of difference between the two textbooks than CVS1, which indicates that VS2 is more sensitive than CVS1 as an index of verb sophistication.

Lexical Variation

Number of different words

There are four indices under LR measure, number of different words: NDW, NDWZ-50, NDW-ER50, and NDW-ES50. The latter three are corrections and standardized versions of NDW since NDW relies heavily on text length and it is not comparable between EFL writings with different numbers of tokens. Given this, NDW is excluded in the refined model of LR indices, though there exists a notable discrepancy in NDW across the four Grade levels in the present study.

Among the standardized versions of NDW, the results of ANOVA show that only NDW-ES50 can significantly ($p<0.05$) distinguish EFL writings of different grade levels. Besides, based on Cohen's (1988) recommendations for interpreting the effect magnitude of ANOVA, there is an acceptable medium effect size ($\eta^2=0.068$) for NDW-ES50. Thus, it remains in the refined model of LR indices. It is believed that NDW-ES50 can reduce the influence of text length on EFL writing (Cheung & Jang, 2019; Lu, 2012). Unlike NDWZ-50, which encompasses the initial 50 words of a text, and NDW-ER50, which entails random 50-word samples, NDW-ES50 captures random 50-word sequences, which does not waste data and can preserve the integrity of the sentences (**Table 6**).

Type/token ratio

Table 7 shows the results of ANOVAs of TTR and its transformations. For a similar reason to NDW, TTR is excluded first though it can significantly distinguish different grade levels since it has been documented that TTR is "an unsatisfactory measure" (Covington & McFall, 2010, p. 94) of lexical variation because it is influenced by the text length. This renders it unreliable as a measure of lexical variation (Heaps, 1978; Lei & Yang, 2020; Lu, 2012). The reality is that the TTR value decreases as the text length increases (Hess et al., 1986; Richards & Malvern, 1997). Except for TTR, the following indices can significantly ($p<0.05$) distinguish EFL writing of different grade levels: MATTR-50, CTTR, RTTR, LogTTR, and vocd-D. Among these indices, CTTR has the largest effect size ($\eta^2=0.194$). According to Cohen's (1988) guidelines on effect size conventions, this represents a very large effect size. Thus, CTTR is chosen in the refined model of LR indices.

Table 8. Results of ANOVA for LR measure: Lexical word variation

Index	df	F	Sig.	Effect size
LWV	3	8.580	0.000*	0.128
NV	3	2.013	0.114	0.033
VV1	3	11.070	0.000*	0.159
SVV1	3	11.218	0.000*	0.161
CVV1	3	6.501	0.000*	0.100
VV2	3	2.464	0.064	0.040
AdjV	3	1.208	0.308	0.020
AdvV	3	1.538	0.206	0.026
ModV	3	0.818	0.485	0.014

Note. *Mean difference is significant at level of 0.05

Table 9. Refined model of LR indices

Dimension	Measure	Code	Index
Lexical density	Lexical density	LD	Lexical density
Lexical sophistication	Lexical sophistication	LS1	Lexical sophistication-I
	Verb sophistication	VS2	Verb sophistication-II
Lexical variation	Number of different words	NDW-ES50	NDW (expected sequence 50 words)
	Type/token ratio	CTTR	Corrected TTR
	Lexical word variation	SVV1	Squared VV1

MSTTR solved TTR's problem of text length sensitivity, while it may cause a waste of data. MATTR is effective for texts of any length and ensures no data is wasted. Nevertheless, a shortcoming of both MATTR and MSTTR is that results in different research with different window sizes are not comparable. **Table 7** shows that the effect sizes of CTTR and RTTR are almost the same. As can be seen from their calculational formulas defined before, they are corrections of TTR with litter differences in the denominators of their formulas. Besides, they are highly correlated with each other (Chung & Ahn, 2019; Lu, 2012). In recent years, some research investigated both CTTR and RTTR together as indices of lexical variation (e.g. Kovacevic, 2019; Pyo, 2020; Ströbel et al., 2020; Wang & Jin, 2022), while one of them is excluded in some research because of the high correlation between them (e.g., Chung & Ahn, 2019). In addition, in research investigating both indices together, the results of comparisons, correlations, or regressions for the two indices are almost the same. Therefore, it should be noted that CTTR and RTTR are almost interchangeable in most research, though CTTR is included in the refined model of LR indices in the present study because of its slight edge on effect size.

Lexical word variation

Table 8 shows that four indices of lexical word variation can significantly ($p<0.05$) distinguish EFL writing of different grade levels: LWV, VV1, SVV1, and CVV1. Among them, SVV1 has the largest effect size of the ANOVA. Therefore, SVV1 is chosen for the refined model of LR indices.

It can also be seen from **Table 8** that except for verb variation, none of the lexical variations of other Parts of Speech can significantly distinguish EFL writing of different grade levels ($p>0.05$), including NV, AdjV, AdvV, and ModV. Besides verb variation, the holistic lexical word variation index LWV can also significantly distinguish EFL writing of different grade levels ($p<0.05$), but it can be inferred that this is the contribution of the variation of verbs included in the lexical words. It can also be inferred from lexical word variation that it is not necessary to investigate noun, adjective, adverb, and modifier sophistication because lexical sophistication and verb sophistication are enough.

Similar to corrected versions of VS1 and TTR, SVV1, CVV1, and VV2 are transformations of VV1 made to minimize the impact of text length. Among the indices of verb variation: VV1, SVV1, CVV1, and VV2, it is reported that only the value of SVV1 has a high correlation to the human-rating score of EFL writing (Spring & Johnson, 2022).

CONCLUSIONS

To conclude, the six indices in **Table 9** are included in the refined and concise model of indices for quantitatively measuring LR.

It should be noted that the indices in the refined model are not chosen simply based on the quantitative results of ANOVA, but also based on qualitative human judgment. For example, TTR and NDW are excluded because of their extensive proven shortcomings. With this concise and refined model, researchers do not have to investigate all highly correlated indices under the same construct in future LR research.

Shoes are not one size fits all. It should also be noted that some indices of this model are replaceable. For example, CTTR and RTTR are interchangeable because they are highly correlated to each other. This study has limitations, the first being its generalizability. The refined model is only applicable to the writing of Chinese EFL learners at advanced proficiency levels, particularly university students since the data of this study are collected from CUSs' EFL writing. Then, the refined model does not cover all the dimensions of LR. For example, lexical error is one of the important LR dimensions (Read, 2000), but its index is not included in the model because lexical errors need to be manually identified without well-developed automatic identification software. In future LR studies, sufficient emphasis should be placed on in-depth qualitative error analysis.

Author contributions: YY: conceptualization, methodology, validation, formal analysis, investigation, data curation, writing–original draft, writing–review & editing; **ZZ:** conceptualization, formal analysis, data curation, writing–original draft, writing–review & editing. Both authors approved the final version of the article.

Funding: The authors received no financial support for the research and/or authorship of this article.

Acknowledgements: The authors would like to thank the editors and anonymous reviewers for their suggested revisions to this article, which have played a crucial role in improving its quality.

Ethics declaration: The authors declared that the study does not involve human or animal subjects. The authors further declared that the study does not require any ethical approval.

Declaration of interest: The authors declare no competing interest.

Data availability: Data generated or analyzed during this study are available from the authors on request.

REFERENCES

Ai, H., & Lu, X. (2010). A web-based system for automatic measurement of lexical complexity. In *Proceedings of the 27th Annual Symposium of the Computer-Assisted Language Consortium*.

Bulté, B., & Housen, A. (2014). Conceptualizing and measuring short-term changes in L2 writing complexity. *Journal of Second Language Writing, 26*, 42-65. https://doi.org/10.1016/j.jslw.2014.09.005

Chang, H., Huang, K., & Wu, C. (2006). Determination of sample size in using central limit theorem for Weibull distribution. *International Journal of Information and Management Sciences, 17*(3), 31-46.

Chaudron, C., & Parker, K. (1990). Discourse markedness and structural markedness: The acquisition of English noun phrases. *Studies in Second Language Acquisition, 12*(1), 43-64. https://doi.org/10.1017/S0272263100008731

Cheung, Y. L., & Jang, H. (2019). Effects of task structure on young learners' writing quality. *INTESOL Journal, 16*(1), 52-78. https://doi.org/10.18060/23193

Chung, E. S., & Ahn, S. (2019). Examining cloze tests as a measure of linguistic complexity in L2 writing. *Language Research, 55*(3), 627-649. https://doi.org/10.30961/lr.2019.55.3.627

Cohen, J. (1969). *Statistical power analysis for the behavioral sciences*. Academic Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.

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

Crossley, S. A., & McNamara, D. S. (2017). *Adaptive educational technologies for literacy instruction*. Routledge. https://doi.org/10.4324/9781315647500

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*, 1227-1237. https://doi.org/10.3758/s13428-015-0651-7

Dewi, R. (2017). Lexical complexity in the introductions of undergraduate students' research articles. *Jurnal Pendidikan Bahasa dan Sastra Inggris* [*Journal of English Language and Literature Education*]*, 6*(2), 161-172. https://doi.org/10.26618/exposure.v6i2.1179

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

Fan, J., Yang, C., & Huang, Z. (2023). Lexical richness of Chinese college students' spoken English. *Journal of English Language Teaching and Applied Linguistics, 5*(2), 1-14. https://doi.org/10.32996/jeltal.2023.5.2.1

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

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175-191. https://doi.org/10.3758/BF03193146

Geng, H., & Yang, Y. (2021, October 4-5). *Lexical richness in English travel guidebooks by EEL and ENL writers* [Paper presentation]. The 7th Malaysia International Conference on Foreign Languages 2021 (MICFL2021), Kuala Lumpur, Malaysia. http://micfl2021.upm.edu.my

Gregori-Signes, C., & Clavel-Arroitia, B. (2015). Analyzing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences, 198*, 546-556. https://doi.org/10.1016/j.sbspro.2015.07.477

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2004). *An introduction to functional grammar*. Edward Arnold.

Harley, B., & King, M. L. (1989). Verb lexis in the written compositions of young L2 learners. *Studies in Second Language Acquisition, 11*(4), 415-439. https://doi.org/10.1017/S0272263100008421

Hawkey, R., & Barker, F. (2004). Developing a common scale for the assessment of writing. *Assessing Writing, 2*(9), 122-159. https://doi.org/10.1016/j.asw.2004.06.001

Heaps, H. S. (1978). *Information retrieval, computational and theoretical aspects*. Academic Press.

Hess, C. W., Sefton, K. M., & Landry, R. G. (1986). Sample size and type-token ratios for oral language of preschool children. *Journal of Speech, Language, and Hearing Research, 29*(1), 129-134. https://doi.org/10.1044/jshr.2901.129

Housen, A., & Kuiken, F. (2009). Complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 461-473. https://doi.org/10.1093/applin/amp048

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Definitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32, pp. 1-20). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.32.01hou

Huang, L., & Qian, X. (2003). An inquiry into Chinese learners' knowledge of productive vocabulary: A quantitative study. *Chinese Language Learning, 24*(1), 56-61. https://doi.org/10.3969/j.issn.1003-7365.2003.01.010

Hyltenstam, K. (1988). Lexical characteristics of near-native second-language learners of Swedish. *Journal of Multilingual & Multicultural Development, 9*(1-2), 67-84. https://doi.org/10.1080/01434632.1988.9994320

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

Kovacevic, E. (2019). The relationship between lexical complexity measures and language learning beliefs. *Jezikoslovlje* [*Linguistics*]*, 20*(3), 555-582. https://doi.org/10.29162/jez.2019.20

Kwak, S. G., & Kim, J. H. (2017). Central limit theorem: The cornerstone of modern statistics. *Korean Journal of Anesthesiology, 70*(2), 144-156. https://doi.org/10.4097/kjae.2017.70.2.144

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing, 34*, 12-24. https://doi.org/10.1016/j.jslw.2016.10.003

Laufer, B. (1991). The development of L2 lexis in the expression of the advanced learner. *The Modern Language Journal, 75*(4), 440-448. https://doi.org/10.2307/329493

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25*(2), 21-33. https://doi.org/10.1177/003368829402500202

Laufer, B., & Nation, I. (1995). Lexical richness in L2 written production: Can it be measured. *Applied Linguistics, 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. *Journal of English for Academic Purposes, 47*, 100894. https://doi.org/10.1016/j.jeap.2020.100894

Li, X. (2021). A corpus based study on lexical richness of flipped classroom model in college English writing. *Journal of Bengbu University, 10*(6), 71-78. https://doi.org/10.3969/j.issn.2095-297X.2021.06.016

Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*. CWK Gleerup.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190-208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Malvern, D. D., & Richards, B. J. (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray (Eds.), *Evolving models of language* (pp. 58-71). Multilingual Matters.

Malvern, D., & Richards, B. (2013). Measures of lexical richness. In C. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 3622-3627). John Wiley and Sons, Inc. https://doi.org/10.1002/9781405198431

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

McClure, E. (1991). A comparison of lexical strategies in L1 and L2 written English narratives. *Pragmatics and Language Learning, 2*, 141-154.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and Linguistic Computing, 15*(3), 323-338. https://doi.org/10.1093/llc/15.3.323

McNamara, D. S., Graesser, A. C., McCarthy, P. M., & Cai, Z. (2014). *Automated evaluation of text and discourse with Coh-Metrix*. Cambridge University Press. https://doi.org/10.1017/CBO9780511894664

Meara, P. (2005). Lexical frequency profiles: A Monte Carlo analysis. *Applied Linguistics, 26*(1), 32-47. https://doi.org/10.1093/applin/amh037

Michel, M. (2017). Complexity, accuracy and fluency in L2 production. In S. Loewen, & M. Sato (Eds.), *The Routledge handbook of instructed second language acquisition* (pp. 68). Taylor & Francis. https://doi.org/10.4324/9781315676968-4

Moder, K. (2007). How to keep the type I error rate in ANOVA if variances are heteroscedastic. *Austrian Journal of Statistics, 36*(3), 179-188. https://doi.org/10.17713/ajs.v36i3.329

Moder, K. (2010). Alternatives to F-test in one way ANOVA in case of heterogeneity of variances (a simulation study). *Psychological Test and Assessment Modeling, 52*(4), 343.

Nasseri, M., & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing, 47*, 100511. https://doi.org/10.1016/j.asw.2020.100511

Pólya, G. (1920). Über den zentralen Grenzwertsatz der Wahrscheinlichkeitsrechnung und das Momentenproblem [About the central limit of the probability calculation and the moment problem]. *Mathematische Zeitschrift* [*Mathematical Magazine*]*, 8*(3-4), 171-181. https://doi.org/10.1007/BF01206525

Pyo, H. (2020). The effects of dictionary app use on college-level Korean EFL learners' narrative and argumentative writing. *Journal of Asia TEFL, 17*(2), 580. https://doi.org/10.18823/asiatefl.2020.17.2.17.580

Read, J. (2000). *Assessing vocabulary*. Cambridge University Press. https://doi.org/10.1017/CBO9780511732942

Richards, B., & Malvern, D. (1997). Quantifying lexical diversity in the study of language development. In B. J. Richards (Ed.), *Quantifying lexical diversity in the study of language development: New Bulmershe papers*. University of Reading.

Šišková, Z. (2012). Lexical richness in EFL students' narratives. *Language Studies Working Papers, 4*, 26-36.

Spring, R., & Johnson, M. (2022). The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK and SpaCy tools. *System, 106*, 102770. https://doi.org/10.1016/j.system.2022.102770

Ströbel, M., Kerz, E., & Wiechmann, D. (2020). The relationship between first and second language writing: Investigating the effects of first language complexity on second language complexity in advanced stages of learning. *Language Learning, 70*(3), 732-767. https://doi.org/10.1111/lang.12394

Treffers-Daller, J., Parslow, P., & Williams, S. (2018). Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics, 39*(3), 302-327. https://doi.org/10.1093/applin/amw009

Ure, J. (1971). Lexical density: A computational technique and some findings. In M. Coulthard (Ed.), *Talking about text* (pp. 27-48). English Language Research, University of Birmingham.

Wan, L. (2010). An empirical investigation into lexical diversity of Chinese English majors' TEM writings. *Foreign Language World, 31*(1), 40-46.

Wang, L., & Jin, C. (2022). Effects of task complexity on linguistic complexity for sustainable EFL writing skills development. *Sustainability, 14*(8), 4791. https://doi.org/10.3390/su14084791

Wang, Z. (2018). The analysis of lexical complexity of two college English textbooks. In *Proceedings of the 3rd International Conference on Education and Management Science*. https://doi.org/10.12783/dtssehs/icems2018/20109

Wen, Q., Liang, M., & Yan, X. (2008). *Spoken and written English corpus of Chinese learners (version 2.0)*. Foreign Language Teaching and Research Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy, & complexity*. University of Hawaii Press.

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. In *Proceedings of the 2015 International Conference on Social Science, Education Management and Sports Education*. https://doi.org/10.2991/ssemse-15.2015.419

Yang, Y., Yap, N. T., & Mohamad Ali, A. (2023). Predicting EFL expository writing quality with measures of lexical richness. *Assessing Writing*, *57*, 100762. https://doi.org/10.1016/j.asw.2023.100762

Yang, Y., Zhang, F., & Zhang, S. (2022). Yīngyǔ xiězuò zhōng cíhuì fēngfù xìng cèliáng wéidù, fāngfǎ yǔ zhǐbiāo yánjiū zòngshù [An overview on dimensions, measures, and indices of lexical richness in English writing]. *Wàiyǔ yǔ fānyì* [*Foreign Languages and Translation*], *29*(4), 80-85. https://doi.org/10.19502/j.cnki.2095-9648.2022.04.006

Zhang, H., Chen, M., & Li, X. (2021). Developmental features of lexical richness in English writings by Chinese beginner learners. *Frontiers in Psychology, 12*, 665988. https://doi.org/10.3389/fpsyg.2021.665988

Zhang, Y. (2021). A study on the lexical richness development in online writing for English learners. *Journal of Gansu Normal Colleges, 26*(3), 31-34. https://doi.org/10.3969/j.issn.1008-9020.2021.03.007