Journal of English Language Teaching and Applied Linguistics
ISSN: 2707-756X
DOI: 10.32996/jeltal
Journal Homepage: www.al-kindipublisher.com/index.php/jeltal

JELTAL
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

Lexical Richness of Chinese College Students' Spoken English

Fan Jiamin[1] ✉ Yang Caixin[2] and Huang Zheng[3]
[123]Shanghai University of International Business and Economics, Shanghai, China
Corresponding Author: Fan Jiamin, **E-mail**: 918948283@qq.com

| ABSTRACT
Lexical richness has been considered one of the most effective methods of assessing writing proficiency. However, the studies on spoken English lexical richness for EFL Chinese students are relatively few. By comparing low, middle, and high levels of Chinese college students' spoken English, based on Read's (2000) and Costa's (2005) framework, this study investigates the developmental features of lexical richness in terms of three dimensions: lexical sophistication, lexical variability, and lexical density. With the help of LCA, SpaCy, and Antwordprofiler, this quantitative study evaluates more than 150000 tokens and analyzes the data using SPSS. The findings are as follows: 1) Chinese college students' spoken English lexical variability increases significantly with the increase of the English levels; 2) Chinese college students' spoken English lexical sophistication increases with the increase in the English levels, but it has a critical point of growth rate, using the unique method--"avoidance"; 3) Chinese college students' spoken English lexical density firstly decreases and then increases with the increase of the English level, and the low-level learners use the fewer conjunctions which probably causes the higher lexical density. 4) Costa's (2005) psycholinguistic model can explain the language features in a non-specific language view. Based on the above conclusions, some suggestions are put forward for oral English teaching to improve the students' oral English ability.

| KEYWORDS

Lexical richness, psycholinguistic model, spoken English, lexical sophistication, lexical variability, lexical density

| ARTICLE INFORMATION

ACCEPTED: 20 March 2023 **PUBLISHED:** 31 March 2023 **DOI:** 10.32996/jeltal.2023.5.2.1

1. Introduction

As a language skill, learners' oral output and their lexical richness play an important role in the development of learners' second language proficiency, which has always been an important indicator of learners' language proficiency (Lu, 2012). And a psycholinguistic model is an approach to explaining language phenomena (Kroll & Tokowicz, 2005). However, in the field of lexical richness, previous studies haven't combined psycholinguistic models with lexical richness. Different from the previous studies, this article tentatively finds a new way to use a specific psycholinguistic model and explain the lexical richness with the help of the latest measuring tools. Therefore, we try to use data discovery and a psycholinguistic model to provide a new interpretation of lexical richness.

The study of vocabulary fountained in the 1990s. Since then, a large number of studies (e.g. Engber, 1995; Laufer, 1998; Nation, 1990; Schmitt, 2000.) have studied vocabulary from different perspectives. Lexical richness (LR), as an important aspect of second language acquisition, has become a hot field in the study of L2 vocabulary because it is not only an indicator of learners' productive vocabulary performance (Fletcher & Macwhinney, 1995;), but also the level of L2 writing (Wang & Zhou, 2012) and oral proficiency (Jarvis, 2002). Linnarud (1986) conducted a study on the vocabulary use of Swedish English learners and native speakers of the same age. The results show that the diversity and originality of the vocabulary of Swedish English learners are lower than that of native speakers, and there are great differences in the use of high-frequency vocabulary. Engber (1995:139-155) studied the lexical richness of 66 limited-time compositions written by Indian university students. The results showed that writing achievement was positively correlated with the degree of lexical change and negatively correlated with the rate of lexical errors. What's more, Kyle

& Crossley (2015) focus on a sub-dimensional construct of lexical richness to investigate the relationship between TOEFL oral marks and indicators. Eguchi & Kyle (2020), Kim et al. (2018), and Wang (2021) also take some studies on sub-dimensional lexical richness.

Studies of the relationship between lexical richness and sample quality have also been conducted on the L2 Chinese corpus. At the end of the 1990s, mainly ten years later, lexical richness was still one of the main focuses of L2 vocabulary research, which has attracted the attention of Chinese researchers and practitioners, focusing on the vertical and horizontal developmental characteristics of Chinese English learners' productive vocabulary and the relationship between vocabulary richness and other language abilities, such as writing ability (e.g., Bao,2008; Wan, 2010; Wang & Zhou, 2012; Lei & Yang, 2020; Zhang et al., 2021; Li & Zhang, 2021). Bao (2008:38-44) uses wordsmith and range to compare the development patterns of lexical richness in English learners' timed compositions. Wan's (2010:40-46) studies of English major students show that the richness and diversity of vocabulary in their second language learning have improved significantly with the improvement of English proficiency. Wang & Zhou (2012:40-43) conducted a dynamic follow-up study on the development and change of non-English Majors' lexical richness in second language learning and its relationship with second language learning quality from the four dimensions of lexical diversity, lexical sophistication, lexical density and lexical errors. What's more, Lei & Yang (2020) used a corpus to find the differences among Chinese EFL learners, English native beginner students and experts in written forms. Zhang et al. (2021) compared the English compositions of Chinese beginner learners by using four lexical richness measures. Li & Zhang (2021) turned to the object of the L3 ethnic minority of Chinese learners with their writing compositions. And as to the perspective of the oral corpus, lexical richness for EFL Chinese students is relatively few (e.g., Lu, 2012; Zhang & Daller, 2020; Shi & Lei, 2021). Lu (2012) used large corpus and multidimensional measures to test which measures are correlated to Chinese students' oral proficiency. Zhang & Daller (2020) tested 60 participants from Grades 2, 5 & 7 in an international examination and tried to see the validity of the measures of oral proficiency. What's more, Shi & Lei (2021) examined speakers from different social classes and tried to find out what are the differences in lexical use. However, these studies are seldom involved in psychological interpretations, and few studies focus on the developmental features of both Chinese learners and oral corpora.

This study focuses on oral English with the latest tools and explores the differences in lexical richness at different levels of spoken English. And this study attempts to explore the developmental characteristics of Chinese learners' oral English in lexical richness from three dimensions: lexical density, lexical sophistication and lexical diversity, and then explain the discoveries by a certain psycholinguistic model. In this article, chapter one is the introduction; chapter two is the literature review of lexical richness and psycholinguistic model; chapter three is the method and data; chapters four and five refer to the results and discussion; the last chapter is the conclusion. The paper mainly answers the following three questions:

(1) What are the developmental features of Chinese college students' spoken English lexical variability?

(2) What are the developmental features of Chinese college students' spoken English lexical sophistication?

(3) What are the developmental features of Chinese college students' spoken English lexical density?

(4) What are the reasons for these features of lexical richness?

2. Literature review

Read (2000:203-205) summarized the dimensions of lexical richness as lexical variability, lexical sophistication, lexical density and lexical error. Since then, most studies have adopted this framework, such as Lu (2012). Lexical variability is the ratio of different lexemes / total lexemes, and it can judge the ability of word production. Lexical sophistication is the proportion of low-frequency words, and it can judge whether the oral context is sophisticated or not. Lexical density is the ratio of the notional lexis in the oral context, and it can judge the amount of information and the ability to express it.

Lexical variability is "the scope of the expression". Read believes that if one is a proficient learner, he will use fewer synonyms, superordinates and other related words (Read 2000:200). Lexical variability refers to whether the words used in language expression are rich in variety, which is used to measure the richness of the words used by learners in language production. It mainly examines the repetition rate of the words used. The lower the repetition rate of words, the wider the scope of words used and the higher the degree of lexical change. What's more, Type-token Ratio (TTR) is an important index to evaluate lexical variability; however, this standardized index is not satisfied (Covington & McFall, 2010) because longer text sample will have lower TTR, making the statistics unrealistic (Lu, 2012; Thordardottir & Weismer, 2001; Shi & Lei, 2022). However, Lu (2012) calculated several measurements of TTR, and he found that CTTR (types $/\sqrt{2\text{tokens}}$) is correlated with learners' level, and text length has a small influence on it. So we choose CTTR as the index of lexical variability.

Lexical sophistication, in general, computes the proportion of words in a text that are "relatively uncommon or superior" (Read, 2000). Lexical sophistication is shared with Laufer (1998). He introduced the Lexical Frequency Profile, including the first 1000 most frequent words (GSL1K—General Service List 1000), the second 1000 most frequent words (GSL2K—General Service List 2000), and the 570 most frequent academic words (AWL—Academic Word List). Read (2000) believed that lexical sophistication should reflect the proportion of low-frequency words. Moreover, Zheng (2016) found that low-frequency words will help students do better in academic work, as well as Higginbotham & Reid (2019). Since Anthony (2014) created the AntWordProfiler and which can process the Lexical Frequency Profile, we chose this software to calculate lexical sophistication.

Lexical density is shared by Ure (1971) and later by Engber (1995). They all claimed that lexical density is defined as the ratio of lexical words to the total number of words in texts. Here, lexical words are content words. Because the meaning of notional lexis is more specific than that of function words, the content of notional lexis (quantifier, verb, pronoun, adverb, adjective, noun, numeral, interjection and onomatopoeia) is more abundant (Zheng, 2016; Gregori-Signes & Clavel-Arroitia, 2015). If the proportion of notional lexis in the total number of words is high, the lexical density of the text content will be high, and the information will be rich, too.

After introducing the terms of Lexical Richness, we should explain the phenomenon which shows the characteristics of lexical richness from its sub-dimensions. A great number of researchers, like Lu (2012), explained the phenomenon from the perspective of Statistics and concluded the characteristics of L2 learners. However, few studies concentrated on the explanation for lexical richness's characteristics from a psycholinguistics perspective. From the linguistic perspective, lexical richness is the indicator to measure L2 learners' ability of language production. And from the field of psychology, the characteristics of lexical richness are explained by theory in the process of lexical production, and the characteristics of second language learners can be explained by lexical access. Lexical Access refers to the process of lexicalization, that is, the process of transforming thinking into word expression and further into sound. The process of lexical access mainly includes two stages: the first stage of lexical access is semantic activation and lexical selection, and the second stage is phonological coding (Forster & Davis, 1984). Semantic activation is the process of making the thinking transform into lexical meaning and then selecting the suitable vocabulary. The second stage prepares the speech plan so that the speaker can extract the phonological form of the vocabulary.

Inspired by the monolingual speech production model of Levelt (1993) and the bilingual speech production model of De Bot (2020), some scholars (Costa, 2005; Kroll & Tokowicz, 2005) proposed the lexical access model in the process of bilingual speech production. Scholars agree that speech production includes three levels of representation and selection: concept, vocabulary and phoneme, and the ongoing study will use the psycholinguistic model to give some explanations. In the process of speech production, there are differences in representation types and activation levels, and if the activation level of a certain type of representation is high, it is easy to extract; on the contrary, it is not easy to extract. In addition, they also discuss the extraction process of bilingual vocabulary, which has the characteristics of non-specific language at different representation levels. They believe that not only the target language but also the vocabulary and phonetic nodes of non-target language related to semantics are activated at the same time, which has an impact on the processing of the target language.

The processing mechanism of bilingual vocabulary is the major question in lexical access, and we discuss it in three steps in Costa's (2005) framework. (1) Word selection: according to the non-specific language view, in the process of bilingual vocabulary production, the activation of the conceptual system spreads to the lexical representation linked by the two languages at the same time, and the target language and non-target language compete with each other (Green, 1998). As for the correct choice of target words, scholars also put forward two hypotheses: first, the activation degrees of target words and non-target words are different, and the former is higher than the latter (Hermans et al., 1998); second, the target language has an inhibitory effect on the lexical representation of non-target language. Target words and non-target words are activated and compete at the same time, but the former usually has a higher activation level and has an inhibitory effect on the latter (Green, 1998). (2) Speech coding: The cascade model (Dell, 1986) holds that words with similar semantics and speech are activated simultaneously in the process of monolingual vocabulary extraction. Similarly, words with similar semantics and pronunciation in the two languages will be activated at the same time. In other words, the lexical representations of the target language and the non-target language of bilinguals will transfer the activated information to their corresponding phonetic representations (Costa et al., 2000:1293). Research shows that the difficulty and speed of target vocabulary speech activation are affected by the phonetic characteristics of the non-target language (Olson, 2013). (3)Grammatical processing: As for the grammatical features of bilingual speech production, the research mainly focuses on the linear rules of word order in code switching and emphasizes the legitimacy of discourse structure. The precondition of code switching is that the word order before and after the switch point must be consistent in the two languages. For example, Poplack (1980: 586) believes that code switching often occurs between monolingual and L2 parallel elements, which does not violate the syntactic rules of any language.

Several variables that affect the degree of lexical access are listed as follows (Carroll, 2007) : (1)Word frequency is the major variable to consider, and it is found by Foss (1969). High-frequency words need a shorter time to activate and manipulate. (2) Phonological

variable shows that word recognition is influenced by prosodic factors such as stress and intonation pattern. (3) Syntactic category contains different types of words like closed words and open words that can influence the effect of frequency and thus affect the extract speed. (4) lexical sophistication shows that reflection time varies with the complexity of word derivation. Mackay (2012) has found that Words with affixes take longer to react than words without affixes. (5) Semantic priming refers to the activation of another related word by a first presented word, and it influences the reaction time. (6) Lexical ambiguity means that a word can be interpreted as containing multiple meanings, and lexical ambiguity will increase the processing burden. What's more, several variables that affect word choice are listed as follows: (1) Individual differences: Language proficiency plays a key role in the activation and extraction of bilingual vocabulary. Researchers (Costa & Sentesteban, 2004) found that low-level bilinguals adopt a non-specific language selection mechanism to extract target words by suppressing the interference of non-target language. With the improvement of their language level, bilinguals gradually form a specific language selection mechanism, adjust the concept, and can extract the target words directly without the help of non-target language. (2) Linguistic features: The type similarity of language plays a significant role in the activation and extraction of bilingual vocabulary. If the two languages are similar at a certain level, the choice of target language vocabulary will be affected by non-target language at this level (De Angelis & Selinker, 2001). (3) Contextual factor: Contextual information such as participants, location, degree of formality and theme all regulate the activation and extraction of target vocabulary. Dewaele's (2001) research shows that bilinguals show more specific language representation in formal situations and show more fluent, frequent and diversified code switching in communication with bilingual friends. In this study, the characteristics of lexical richness are explained by these factors, and these language phenomena are viewed from the perspective of psychology. These factors will be considered in the following discussion.

3. Research method and data

3.1 Corpus description

This study uses the corpus of the COLSEC. COLSEC (College Learners' Spoken English Corpus) contains the corpus of CET-4 and CET-6, which reflects the characteristics and level of some learners' spoken English. The corpus is built by the China Foreign Language Education Research Center of Beijing Foreign Studies University and is the first large-scale spoken and written English Corpus for English learners in China.

The corpus is derived from the video of CET-4 and CET-6 from 2000 to 2003. CET-4 and CET-6 is a national tests organized by the Ministry of Chinese Education. It contains more than 1000 precious oral phonetic samples, 1-million-word phonetic transcribed texts, 1-million-word written composition samples and a corpus introduction. In this corpus, three levels can be divided. The wide distribution of candidates ensures the representativeness of the corpus. Table 1 shows the data of this study.

Table 1 Data of this study

Level of Oral Grade	Number of texts	Types	Tokens	Notional words
A&A+	52	49556	15704	21131
B&B+	96	85056	25536	34754
C&C+	75	39225	18750	16475
Total	223	173837	59990	72360

Note: A&A+ means a high level of oral English. B&B+ and C&C+ mean the middle and low level, respectively (henceforth, A &A+ equal high level, B&B+ equal middle level and C&C+ equal low level).

As shown in Table 1, the corpus contains a total of 223 texts, of which B&B+ occupies the most. What's more, it contains oral pretend sentences which cannot be seen in the composition, like "Er", "I think". So it is credible to use this corpus. And the total tokens of each text are about 850.

The scoring rating criteria for the oral test are shown in Table 2. The oral performance represents the holistic oral rating, and three levels are selected.

Table 2 Rating Criteria for Oral Test

Ranking	Description
A&A+	Can talk about familiar topics in English, basically without difficulty. Be able to consistently express opinions and opinions on familiar topics. Be able to clearly and fluently describe or describe general events and phenomena.
B&B+	Can talk about familiar topics in English, but it does not affect communication. Be able to make more coherent speeches on familiar topics. Be able to describe or describe general events and phenomena clearly and fluently.
C&C+	Be able to have a simple conversation on familiar topics in English. Be able to make short speeches on familiar topics. Can simply describe or describe general events and phenomena.
D	Oral communication skills are not yet available.

Table 2 describes the ranking criteria of CET4 and CET6 oral performance. It generally differs in accuracy, range, discourse length, coherence, flexibility and appropriateness. And this research takes A&A+, B&B+ and C&C+ as samples to construct corpus data.

This corpus includes three types of data: teacher-student interview, student-student free discussion and teacher-student discussion. What's more, it contains 39 specific topics in the college oral English test, which can be divided into three categories: personal life and learning, social concerns and campus life. At present, only the oral test corpus from 2001 to 2004 is selected in this corpus, and the time span is 4 years. As a result, it is a synchronic corpus (Wei et al., 2007).

3.2 Data analysis

In this paper, quantitative methods are used to test the differences in oral English learners' lexical richness, observing and analyzing the quantitative relationship in lexical richness. Then discuss the data on the basis of the corpus. What's more, take a thorough look at the relationship between learners of different grades.

As to the study on lexical variability and lexical density, this study uses Spring & Johnson's (2022) updated version from Lu's (2012) LCA (Lexical Complexity Analyzer) based on SpaCy to measure indicators. LCA's CTTR and LD are the indicators of this study because Lu (2012) has verified that they correlate to the learners' levels.

As to the study on lexical sophistication, this study uses the Antwordprofiler (Anthony, 2014) to achieve low-frequency word retrieval and lexical sophistication statistics. Laufer & Nation (1995) designed the Lexical Frequency Profile, which consists of three sub thesauri. The formula of low-frequency words is the total number of words minus two sub thesauri words (GSL1K & GSL2K) (Zheng, 2016). Table 3 is the lexical frequency profile.

Table 3 Lexical Frequency Profile

Name	Abbreviation	explanation
General Service List 1000	GSL1K	the first 1000 most frequent words
General Service List 2000	GSL2K	the second 1000 most frequent words
Academic Word List	AWL	the 570 most frequent academic words

Note: Three word-frequency lists are in Table 3, General Service List 1000 (GSL1K), General Service List 2000 (GSL2K) and Academic Word List (AWL). GSL1K and GSL2K are considered the lower level vocabulary because they are more frequent words than others, and other words that are not involved in these word lists are considered uncommon words or advanced words.

This study uses SPSS to calculate the mean value and standard deviation, then uses a homogeneity test of variance, one-way ANOVA and post-hoc analysis to test the relationship. Using this method can easily find whether two different oral levels have a relationship.

4. Results

Table 4 shows that the mean scores for lexical richness increase with higher grades. This holds for Tokens, Types, GSL1K, GSL2K, AWL, Low Frequency Words, Lexical sophistication and Lexical density. Table 4 shows the post-hoc comparison (Tukey) between the levels.

Table 4 Lexical Richness mean scores between the different levels (ANOVA and post-hoc comparison Tukey)

Measure	Level	N	Mean	SD	ANOVA			Post-hoc comparison (Tukey)		
					F	df	p	low vs middle	low vs high	middle vs high
Tokens	low	52	523	77.5	9.657	2	<0.01	<0.01	<0.01	<0.01
	middle	96	886	83.4						
	high	75	956	97.9						
Types	low	52	250	49.6	4.715	2	<0.01	<0.01	<0.01	<0.01
	middle	96	266	56.3						
	high	75	302	52.8						
GSL1K	low	52	81.9	20.6	10.365	2	<0.01	<0.01	<0.01	0.681
	middle	96	79.8	28.2						
	high	75	76.47	31.5						
GSL2K	low	52	9.49	5.5	58.641	2	<0.01	<0.01	<0.01	<0.01
	middle	96	9.76	5.9						
	high	75	10.55	7.6						
AWL	low	52	7.41	9.6	63.258	2	<0.01	<0.01	<0.01	<0.01
	middle	96	7.12	4.8						
	high	75	6.84	5.9						
Low Frequency Words	low	52	1.2	1.6	21.664	2	<0.01	<0.01	<0.01	0.357
	middle	96	3.32	2.9						
	high	75	6.14	4.6						
Lexical sohpistication	low	52	8.61	10.8	18.624	2	<0.01	<0.01	<0.01	0.503
	middle	96	10.44	15.8						
	high	75	12.98	17.3						
Lexical density	low	52	42.64	50.8	52.351	2	<0.01	<0.01	0.354	0.089
	middle	96	40.86	36.5						
	high	75	42	48.1						

4.1 Lexical Variability

Lexical variability is "the scope of the expression", and the more proficient a learner is, the less repetition he uses synonyms, superordinates, and other related words.

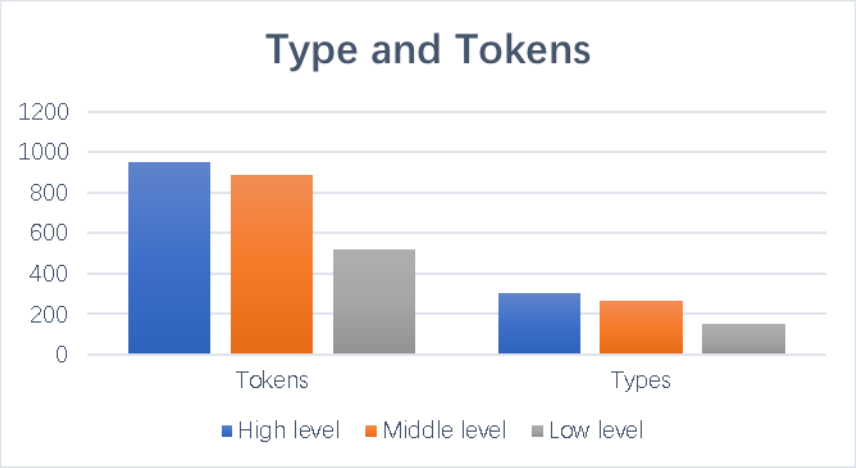

Figure 1 Type and Tokens

As shown in Figure 1, the types and tokens of each grade are302, 953; 266, 886; 250, 523, respectively

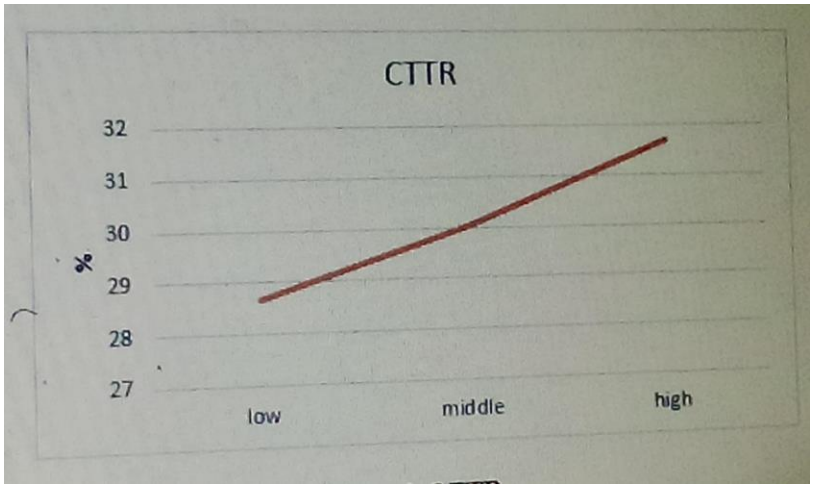

Figure 2 CTTR

From Figure 2, we can see the CTTR increases from the low level to the high level (28.65<30.07<31.74).

The homogeneity test of variance and one-way ANOVA showed that there is a significant difference between groups ($F = 4.715$, $P < 0.010$). And further post-hoc analysis shows that there are significant differences among the groups of each grade ($p < 0.010$), and the degree of lexical variability increases.

4.2 Lexical Sophistication

Lexical sophistication reflects the proportion of low-frequency words, which are considered advanced vocabulary (Read, 2000), and using lower-frequency words or more sophisticated words indicates higher language ability (Lu, 2012).

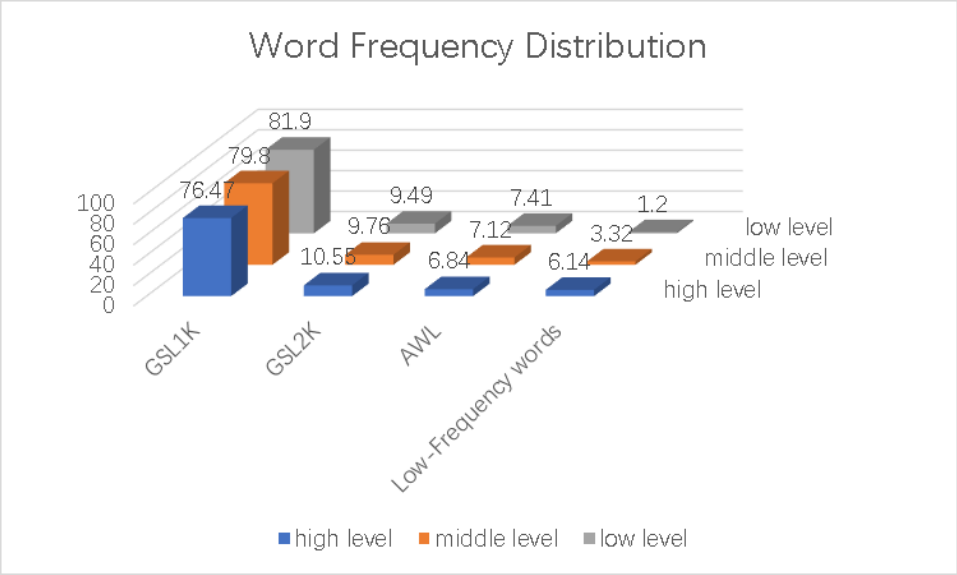

Figure 3 Word Frequency Distribution

Figure 3 illustrates the proportion of each list respectively. According to this figure, the frequency distribution of low-level spoken English is 81.90%-9.49%-7.41%-1.20%; the middle-level is 79.80%-9.76%-7.12%-3.32%; the high-level is 76.47%-10.55%-6.84%-6.14%.

By comparing the word frequency distribution of different grades, we can see that with the increase of the grades, the proportion of GSL1K decreased (81.90>79.80>76.47), the proportion of GSL2K increased (9.49<9.76<10.55), the proportion of AWL decreased (7.41<7.12<6.84), and the proportion of the Low-Frequency Words increased (1.20<3.32<6.14). The test of homogeneity of variance and one-way ANOVA showed there were significant differences among groups ($p<0.010$).

Further post-hoc analysis was conducted, and as to GSL1K and Low-frequency Words, there are significant differences between low and middle grades and low and high levels ($p<0.010$), but there is no significant difference between middle and high grades (for GSL1K, $p = 0.681 > 0.050$; for Low-Frequency Words, $p=0.357 >0.050$). And as to AWL, there are significant differences among the three levels ($p < 0.010$). In other words, with the development of English levels, college students' ability to use GSL1K is decreasing; that of using GSL2K and AWL is increasing; the use of Low-Frequency Words is increasing first and then decreasing slowly.

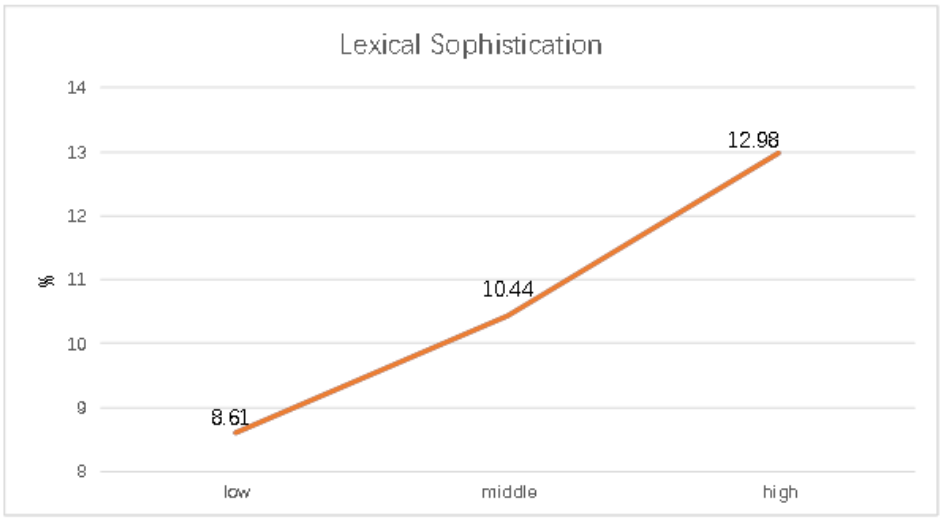

Figure 4 Lexical Sophistication

Figure 4 shows the average lexical sophistication of each level is 8.61%, 10.44% and 12.98%, respectively, and the sophistication increases with the level. The homogeneity test of variance and one-way ANOVA showed that there were significant differences among different grades ($F = 18.624$, $P < 0.010$). Further post-hoc analysis shows that there is a significant difference in lexical sophistication between the low and high grades ($p < 0.010$) but no significant difference between the middle and high grades ($p = 0.503 > 0.050$).

4.3 Lexical Density

The Lexical Density is the proportion of the notional words, and if the proportion of notional lexis in the total number of words is high, the lexical density of the text content will be high, and the information will be rich, too.

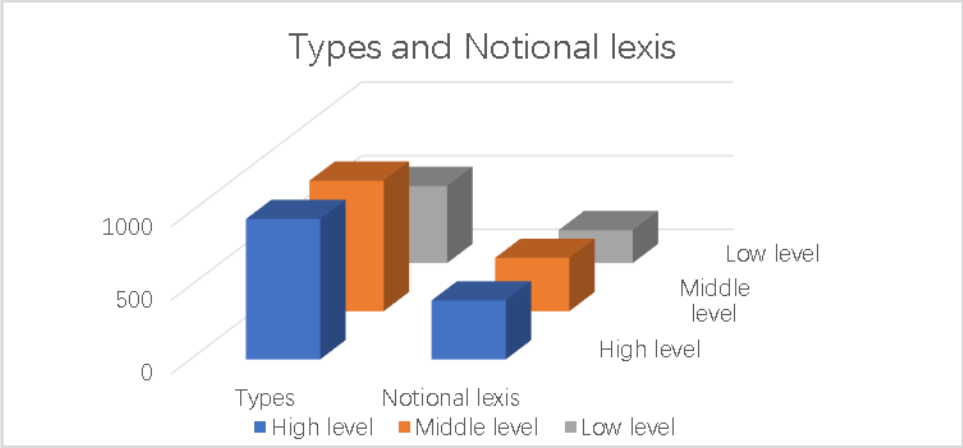

Figure 5 Types and Notional lexis

Figure 5 illustrates the types and notional lexis of each level. As shown in Figure 5, types and notional lexis in high, middle and low levels are 953, 400; 886, 362; 523, 223.

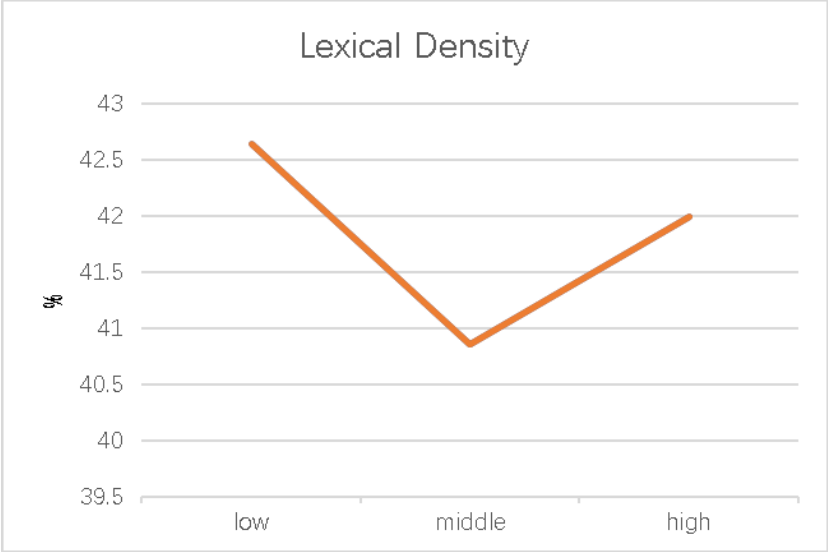

Figure 6 Lexical Density

From Figure 6, by comparing the proportion of notional lexis in different levels, it is found that the density of low, middle and high-level spoken words fluctuates zigzag, falling first and then rising (42.64% > 40.86%; 40.86% < 42.00%).

Further post-hoc analysis shows there are significant differences in lexical density between low and middle grades ($p < 0.010$) but no significant differences among other grades (for the low and high, $p = 0.354 > 0.010$; for the middle and high, $p = 0.089 > 0.010$).

5. Discussion

5.1 Lexical variability

(1) Lexical variability is increased with the levels. Figure 1 illustrates the types and tokes relatively. Zheng (2016) and Zhang & Daller's (2020) found that the types and tokens increase with the level of written and oral forms, and we confirm this law from the figure. However, the fastest type growth rate is from low level to middle level, while the fastest token growth rate is from middle level to high level, which shows the growth situation. Shi and Lei (2021) also compared the oral English of the lower, middle and upper classes with the STTR index.

(2) The lexical variability of oral English is relatively higher than that of written English. Zheng (2016) found that the variability of English Majors' writing vocabulary in one academic year ranged from 22.71 to 27.16. The variability degree of written vocabulary is lower than that of oral vocabulary, and this may be due to more professional vocabulary and more focused topics. In Zhu & Wang (2013), the lexical variability is always below 20%, which shows a normal level of L2 English learners. However, Figure 2 shows that the variability of vocabulary is relatively low, while Engber (1995) found that different levels of American native speakers' written lexical variability were all above 50% (ranging from 54% to 72%). This shows that the diversity of words needs to be improved. For example, in oral speaking, when it comes to analogy, Chinese students always use the words "like"; however, in writing, they often use words like "as if" and "seem", and some skills like parallelism, even metaphor.

(3) Different levels of lexical variability exist in significant differences, and lexical variability increases with the increase of the levels. One-way ANOVA and post-hoc analysis show that productive vocabulary increases with the improvement of grades. With the expansion of their vocabulary, their spoken words tend to be diversified, which shows that the degree of lexical variability can effectively distinguish different oral grades. Most of the previous studies were about middle and high English learners, and they were different in genre, topic and length (González 2017; Bao 2008; Wan 2010). They found that the learners' ability of lexical variability is significantly enhanced with the increase of the levels, which is consistent with this study. The results show that learners at different levels have a consistent track in terms of lexical variability. At the same time, second language beginners also show their own characteristics of the lexical acquisition stage; that is, the degree of lexical variability is generally low, and the productive lexical ability is not strong. This is consistent with the conclusion of (Fairclough & Belpoliti, 2016). This is in line with the common law of second language acquisition; in other words, they are still in the initial stage of second language learning, and it is difficult to internalize all the lexical learned through limited classroom learning and flexibly apply it to oral English.

(4) In psycholinguistics views, characteristics of lexical variability can be proved by factors of Costa's (2005) model. According to Costa's model, the first stage is semantic activation and lexical selection. The increase of CTTR with the development of oral levels shows the learners have strong abilities towards the first stage. That is to say, this type of representation is high and easy to extract with more different words in the process of speech production, and the ability of the first stage is improved with the growth of the level. Moreover, lexical variability shows the degree of word diversity, while semantic priming shows that one word activates another related word, and this results in more different words in a text and makes lexical variability higher. High-level oral learners have high lexical variabilities, and they have the ability to make more different words. On the other hand, semantic priming needs time to produce different, but related words in oral performance, and this skill will be more proficient in highly skilled learners (Meyer & Schaneveldt, 1971). What's more, individual differences can also explain the characteristics of lexical variability. With the improvement of their proficiency, bilinguals gradually form a specific language selection mechanism, which can directly extract target words without the help of non-target words; that is, the activation of target words will not spread to non-target languages and will not be disturbed by non-target words (Costa & Sentesteban, 2004). Therefore, with the improvement of language level, the activation of target vocabulary will not spread to non-target language and will not be disturbed by non-target language vocabulary. The activation and extraction of language will be faster and more conducive to the output of spoken vocabulary.

To sum up, the Chinese college students' oral lexical variability is slightly low (around 30%). However, with the increase in level, the degree of lexical variability increases significantly, the types of spoken words increase largely, and the productive vocabulary ability also improves. The degree of lexical variability does not involve the use of low-frequency words. Learners' vocabulary is constantly expanding, and it is easier to use a variety of words in oral English. Therefore, the degree of lexical variability shows a significant growth trend in the three levels.

5.2 Lexical Sophistication

(1) Chinese students use the "avoidance" strategy and familiar high-frequency words to express their ideas roundly. Through literature review, Yang (2015) compared Chinese and American college students' compositions and found that the word frequency distribution (GSL1K-GSL2K-AWL-Low Frequency Words) of Chinese students' compositions was 79.02% - 7.19% - 5.51% - 8.28% and similar results are reported in Li and Zhang's (2021) study. Through comparison, low-level spoken English Chinese students tend to use more GSL1K words; however, high-level spoken English Chinese students tend to use less GSL1K than written English. Moreover, the distributions of Low-Frequency words are all lower than in written English. Chinese college students prefer to use the high-frequency words of GSL1K rather than the low-frequency words in oral English. This may be due to the low level of College

Students' acquisition of low-frequency words, which makes it difficult for them to use such words freely, so they use the "avoidance" strategy and use familiar high-frequency words to express their ideas roundly. This also reflects that teachers may not frequently instruct the students to use low-frequency words.

(2) The growth rate of lexical sophistication is first fast (167%) and then slow (85%). In other words, we can see that students improve their oral English by increasing lexical sophistication. However, Zhu & Wang (2013) and Zhang et al. (2021) found that the lexical sophistication of L2 English writing is always over 20%, and it can reflect that the lexical sophistication of oral English is relatively low. Post-hoc analysis shows that lexical sophistication increases with the increase of the levels. At the same time, the increase of lexical sophistication firstly goes fast and then slow, and the acquisition speed of high-level low-frequency words is the slowest, especially in Low-Frequency Words, because Schmitt (2000) claimed that the lexical ability is gradual. Low and middle-level learners acquire them faster, while high-level learners are difficult to improve because of their large accumulation. However, the slowing increase rate may be due to SLA theories of the ceiling effect.

(3) The Chinese students' Lexical Sophistication curve is unique, with a critical point of growth rate. Most of the previous studies were based on the lexical frequency distribution (e.g. Engber, 1995; Lu, 2012), and most of them involved middle and high-level learners(e.g. Kyle & Crossley, 2015; Kyle & Crossley, 2016). The oral ability of the beginners in this study is different from the previous study. The growth rate of Chinese learners' vocabulary complexity level from low level to middle level is faster than that from middle level to high level. The different growth rates lead to a critical point on the chart, which shows that the speed is decreasing with the improvement of level. Lexical sophistication has its own development uniqueness: beginners make progress at every level, but low-level students develop faster, and high-level students develop slowly. The results may be attributed to the stage of vocabulary teaching: on the one hand, low-level learners are in the high-frequency word teaching stage, and the low-frequency words reserve is less. From the low level to the middle level, the use of low-frequency words will rapidly improve; On the other hand, low-frequency Words are mostly used in the middle level, and the high-level students tend to paraphrase the obscure meaning rather than using the low-frequency words. In addition, the findings of this study are contrary to Laufer (1998). Based on the "lexical frequency distribution", Laufer compared the English words used in Grade 10 and grade 11 of Israeli students and found that the lexical sophistication had no change. Laufer interpreted it as a plateau phenomenon of vocabulary development in Grade 11. We think that the result may be related to the unreasonable choice of time. Therefore, it is difficult to effectively evaluate their vocabulary development by "lexical frequency distribution".

(4) From the psycholinguistic perspective, the features of lexical sophistication can be proved by Costa's model (2005). The non-specific language model is better to explain the unique curve of lexical sophistication than a specific language model. Costa's model indicates that the target language and non-target language are both activated at the same time. That is to say, the concepts of two languages are both activated in the process of choosing vocabulary and how to select them comes to the activation level (Green, 1998). For example, native speakers usually choose "good" to describe food; however, L2 learners have a word or phrase of their native languages (e.g. "hǎochī", "oishii", "schmeckt", "delicieuse") and translate them into "delicious". Unfortunately, there are differences in representation types and activation levels, and the activation level of L2 learners' native language is high, which means it is easy to extract. With the development of the L2 level, L2 learners consciously escape "delicious", and generally, the lexical sophistication growth rate becomes slower, and there is a critical point in the chart. In this study, we use the wordsmith 6.0 to conduct the concordance, and it shows that the high level contains 0.7 "delicious" per text, and as to the middle and low levels, they are 1.0 and 1.7, respectively. This shows that the growth rate is first quick and then slow, which confirms the non-specific language model. What's more, the factor of frequency can influence the activation time, and high-frequency words generally need a shorter time. GSL1K and GSL2K represent high-frequency words which means that they will not be activated for a long time. In the process of lexical access, L2 learners' ability is improved. They can respond in a shorter time, so they can choose low-frequency words on the basis of short response time (Whaley, 1978). What's more, Mackay (1978) found that the response time of speakers varies with the complexity of derivation. For example, in linguistics, - *ment* is simpler than - *ence*, and - *ence* is simpler than - *ion*. The more complex the affixes, the longer it takes for words to produce them. This indirectly shows that the factors of lexical complexity in lexical access will affect the response time of vocabulary and then affect the use of vocabulary. As in this study, by using Wordsmith 6.0 to query three groups of corpus data of high, middle and low degree, taking - *ion* as an example, the number of words with - *ion* suffix expected in each high, middle and low degree learner is 3.17, 2.56 and 1.38 respectively, which proves that high degree learners will use more sophisticated derivation, thus confirming Mackay's discovery.

To sum up, Chinese college students rely too much on high-frequency words and seldom use low-frequency words because of the "avoidance" strategy. With the increase of their levels, high-frequency words decrease, low-frequency words increase, and lexical sophistication increases, but the growth rate is uneven, first fast and then slow, causing the unique lexical sophistication curve.

5.3 Lexical Density

(1) Notional words and types increase with the development of the levels. It shows that the types and notional words increase with the levels. What's more, the types grow from low level to middle level, and the notional words grow similarly, which shows the

growth situation. Li & Zhang (2021) calculated the lexical word tokens from three stages, and the findings are similar to this study. Zhang et al. (2021) also measure them from three grade levels, and it grows with the development of the level. What's more, Shi & Lei (2021) took the oral corpora to measure different types of notional words with different social classes, and notional words grow with the development of classes. These studies confirm that oral and written notional words both grow with the development of level.

(2) Less notional words are used in Chinese college students' spoken English than in written English. Zheng (2015) found that the lexical density of English Majors' compositions fluctuates between 50% and 57% in one academic year. Fairclough & Belpoliti (2016) took Spanish beginners as subjects, and the average written lexical density was 46.4%. The results show that the density of spoken words is slightly lower than that of written words. This implies the differences between oral and writing: writing tends to choose the notional lexis. Data from this study show that first or second-person pronouns, imperative sentences and exclamations are frequently used in Chinese college students' spoken English, such as "I once studied in this primary school." "You are now listening to the radio." "Put on the coat, please." "How clever a boy he is!" so the lexical density is relatively low. Different levels of Lexical Density have significant differences. According to previous studies, the lexical density of English major freshmen surveyed by Zheng (2015) fluctuates between 50% and 57%. But in this study, the lexical density did not reach 45%, which is closely related to the L3 ethnic minority and L2 beginner learners' lexical density (Li & Zhang, 2021; Zhang et al., 2021). and it can be seen that L2 learners' oral lexical density is generally low, and the content of text information is insufficient.

(3) Lacking conjunctions is one of the reasons for "falsely" high Lexical Density. In other words, with the increase in the English level, lexical density first decreases significantly and then increases slowly. Only from the numerical point of view do the middle and high levels seem to have regressed. But most of the research found that the lexical density is increasing year by year (Bao, 2008; Zhu & Wang, 2013), which is different from this study.

However, a further search of powerGREP found that low-level students use fewer conjunctions than middle and high-level students($2.36 < 5.68 < 6.25$).

Therefore, the higher level the learner is, the more conjunctions it contains, which leads to lower lexical density. This shows low level learners have limited language ability, low-level mechanical stacking of notional lexis, and they cannot use more conjunctions to connect the sentences and ignore the logic of language expression, which reduces the proportion of conjunctions. What's more, the average number is four, and it is less than that of high-level learners. Therefore, the low-level lexical density is falsely high. Relatively speaking, the middle and high-level's types of conjunctions are increasing. The increase in conjunctions leads to a decrease in notional lexis and the decrease in vocabulary density. But it doesn't mean that their lexical ability is declining. However, this characteristic may follow the famous u-shaped curve of second language acquisition, and in this study, we try to explain it from a psycholinguistic view.

(4) In psycholinguistics views, characteristics of lexical density can be explained by Costa's (2005) model. In the process of word selection, target language and non-target language are both activated to select words in the non-specific language model (Green, 1998). With the development of level, the higher percentage of ratio and the number of notional words means that more notional words related to semantics are activated at the same time. In order to deal with more semantic related notional words, learners, vocabulary selection ability becomes stronger and stronger, and more notional words will be produced. The first stage of Costa's model is improved with the development of level and non-specific language models that can better explain. Moreover, contextual factors may affect the use of conjunctions in lexical density. The degree of formality of discourse will regulate the activation and extraction of target vocabulary (Dewaele, 2001). As in this study, the oral corpus of high middle and low degree learners are recorded in CET-4 and CET-6, and their occasions are formal. This study finds that learners at different levels have different numbers of conjunctions output, which can prove that the degree of formality of discourse can regulate the activation and extraction of discourse.

To sum up, Chinese learners' use of notional lexis is relatively insufficient and lacks logic. With the increase in English ability, the lexical density first decreases significantly and then increases slowly. The reason is deficient of conjunctures, causing the falsely high lexical density, and this phenomenon shows the ability to produce complex sentences is relatively low.

6. Conclusion

This study examines the developmental characteristics of Chinese learners' spoken English lexical richness from three dimensions: lexical density, lexical sophistication and lexical diversity and finds the unique characteristic of Chinese students' oral performance.

The present study makes some contributions to the lexical richness in L2 learners' oral performance levels. First, as for lexical sophistication, this study shows a unique lexical sophistication curve with a critical point, and students use the "avoidance" strategy to escape the low-frequency words. Second, as for lexical density, there exists the virtual high lexical density, which is caused by the lack of conjunctions. The fewer conjunctions may probably cause the higher lexical density. Third, this study uses Costa's (2005)

psycholinguistic model to explain the characteristics of lexical richness and implies that a non-specific language model can be better applied than a specific language model.

Based on the above conclusions, some suggestions are put forward for oral English teaching. First of all, in order to escape the "avoidance" strategy, teachers should imply the accurate choice of words instead of the low frequency of words. Secondly, because of the virtually high lexical density, teachers should have the consciousness to emphasize the use of a conjunction.

However, it is important to note the limitation of this study. Firstly, the psycholinguistic model has not yet been empirically tested, and future research can conduct experiments to verify the validity of the model. Secondly, in this study, the characteristics were determined only by one specific measurement, and multiple measurements can be more convincing. It is suggested that future research could delve deeper into the three sub-dimensions of lexical richness, using various computational methods to comprehensively display the values of a certain sub-dimension and conducting empirical research to verify the effectiveness of the model.

Funding: This research received no external funding.
Conflicts of Interest: The author declares no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1] Anthony, L. (2014). *AntWordProfiler (Version 1.4. 1)[Computer Software]. Tokyo, Japan: Waseda University*.
[2] Bao, G. (2008). A survey of the development of lexical richness in L2 compositions from a multidimensional perspective. *Technology Enhanced Foreign Language Education*, *05*, 38–44.
[3] Carroll, D. W. (1986). *Psychology of language* (pp. xvii, 467). Thomson Brooks/Cole Publishing Co.
[4] Chen, M. (2015). A study on the complexity, accuracy and fluency of natural spoken Chinese as a second language. *Language Teaching and Research*, *03*, 1–10.
[5] Costa, A. (2005). *Lexical access in bilingual production*.
[6] Costa, A., Caramazza, A., & Sebastian-Galles, N. (2000). The cognate facilitation effect: Implications for models of lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(5), 1283–1296. https://doi.org/10.1037/0278-7393.26.5.1283
[7] Costa, A., & Santesteban, M. (2004). Lexical access in bilingual speech production: Evidence from language switching in highly proficient bilinguals and L2 learners. *Journal of Memory and Language*, *50*(4), 491–511.
[8] De Angelis, G., & Selinker, L. (2001). Interlanguage transfer and competing linguistic systems in the multilingual mind. *Bilingual Education and Bilingualism*, 42–58.
[9] De Bot, K. (2020). A bilingual production model: Levelt's' speaking'model adapted. In *The bilingualism reader* (pp. 384–404). Routledge.
[10] Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283.
[11] Dewaele, J.-M. (2001). Activation or inhibition? The interaction of L1, L2 and L3 on the language mode continuum. *Bilingual Education and Bilingualism*, 69–89.
[12] Eguchi, M., & Kyle, K. (2020). Continuing to explore the multidimensional nature of lexical sophistication: The case of oral proficiency interviews. *The Modern Language Journal*, *104*(2), 381–400.
[13] Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions. *Journal of Second Language Writing*, *4*(2), 139–155.
[14] Fairclough, M., & Belpoliti, F. (2016). Emerging literacy in Spanish among Hispanic heritage language university students in the USA: A pilot study. *International Journal of Bilingual Education and Bilingualism*, *19*(2), 185–201.
[15] Forster, K. I., & Davis, C. (1984). Repetition priming and frequency attenuation in lexical access. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(4), 680.
[16] Foss, D. J. (1969). Decision processes during sentence comprehension: Effects of lexical item difficulty and position upon decision times. *Journal of Verbal Learning and Verbal Behavior*, *8*(4), 457–462.
[17] González, M. C. (2017). The contribution of lexical diversity to college-level writing. *TESOL Journal*, *8*(4), 899–919.
[18] Green, D. W. (1998). Mental control of the bilingual lexico-semantic system. *Bilingualism: Language and Cognition*, *1*(2), 67–81.
[19] Gregori-Signes, C., & Clavel-Arroitia, B. (2015). Analysing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences*, *198*, 546–556.
[20] Hermans, D., Bongaerts, T., De Bot, K., & Schreuder, R. (1998). Producing words in a foreign language: Can speakers prevent interference from their first language? *Bilingualism: Language and Cognition*, *1*(3), 213–229.
[21] Higginbotham, G., & Reid, J. (2019). The lexical sophistication of second language learners' academic essays. *Journal of English for Academic Purposes*, *37*, 127–140.
[22] Jarvis, S. (2002). Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing*, *19*(1), 57–84.
[23] Kim, M., Crossley, S. A., & Kyle, K. (2018). Lexical sophistication as a multidimensional phenomenon: Relations to second language lexical proficiency, development, and writing quality. *The Modern Language Journal*, *102*(1), 120–141.
[24] Kroll, J. F., & Tokowicz, N. (2005). *Models of bilingual representation and processing: Looking back and to the future.* Oxford University Press.
[25] Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. *Journal of Second Language Writing*, *34*, 12–24.
[26] Kyle, K., & Crossley, S. A. (2015). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *Tesol Quarterly*, *49*(4),

757–786.
[27] Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*(2), 255–271.
[28] Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics*, *16*(3), 307–322.
[29] 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.
[30] Levelt, W. J. (1993). *Speaking: From intention to articulation*. MIT press.
[31] Li, X., & Zhang, H. (2021). Developmental Features of Lexical Richness in English Writings by Chinese L3 Beginner Learners. *Frontiers in Psychology*, *12*.
[32] Linnarud, M. (1986). *Lexis in composition: A performance analysis of Swedish learners' written English*.
[33] Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190–208.
[34] MacKay, D. G. (1978). Derivational rules and the internal lexicon. *Journal of Verbal Learning and Verbal Behavior*, *17*(1), 61–71.
[35] MacKay, D. G. (2012). *The organization of perception and action: A theory for language and other cognitive skills*. Springer Science & Business Media.
[36] Meyer, D. E., & Schvaneveldt, R. W. (1971). Facilitation in recognizing pairs of words: Evidence of a dependence between retrieval operations. *Journal of Experimental Psychology*, *90*(2), 227.
[37] Olson, D. J. (2013). Bilingual language switching and selection at the phonetic level: Asymmetrical transfer in VOT production. *Journal of Phonetics*, *41*(6), 407–420.
[38] Poplack, S. (1980). *Sometimes i'll start a sentence in spanish y termino en espanol: Toward a typology of code-switching1*.
[39] Read, J. (2000). *Assessing vocabulary*. Cambridge university press.
[40] Schmitt, N. (2020). *Vocabulary in language teaching*. Cambridge university press.
[41] Shi, Y., & Lei, L. (2021). Lexical use and social class: A study on lexical richness, word length, and word class in spoken English. *Lingua*, *262*, 103155.
[42] Shi, Y., & Lei, L. (2022). Lexical Richness and Text Length: An Entropy-based Perspective. *Journal of Quantitative Linguistics*, *29*(1), 62–79.
[43] 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*, 102770.
[44] Thordardottir, T., Susan, E., & Weismer, E. (2001). High-frequency verbs and verb diversity in the spontaneous speech of school-age children with specific language impairment. *International Journal of Language & Communication Disorders*, *36*(2), 221–244.
[45] Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing*, *17*(1), 65–83.
[46] Wan, L. (2010). A study of the development of lexical diversity in Chinese English majors writings. *Foreign Language World* (1), 40-46.
[47] Wang, H. (2021). A multidimensional study of lexical complexity in oral interaction of Chinese EFL Learners. *Foreign Language Teaching and Research*, *53*(05), 745-756+800-801. https://doi.org/10.19923/j.cnki.fltr.2021.05.010
[48] Wang, H., & Zhou, X. (2012). A historical study of lexical richness in non- English Majors' writing. *Language Teaching and Research*, *02*, 40–44. https://doi.org/10.13458/j.cnki.flatt.000482
[49] Wei, N. (2007). A study on the phrasal features of Chinese students' oral English—A lexical chunk evidence analysis of COLSEC corpus. *Modern foreign languages*, *03*, 280-291+329-330.
[50] Whaley, C. P. (1978). Word—Nonword classification time. *Journal of Verbal Learning and Verbal Behavior*, *17*(2), 143–154.
[51] Yang, Y. (2015). A comparative study of lexical and lexical chunks in Chinese and American College Students' compositions on the same topic. *Foreign Languages*, *03*, 51-58+75.
[52] Zhang, H., Chen, M., & Li, X. (2021). Developmental Features of Lexical Richness in English Writings by Chinese Beginner Learners. *Frontiers in Psychology*, *12*.
[53] Zhang, J., & Daller, M. (2020). Lexical richness of Chinese candidates in the graded oral English examinations. *Applied Linguistics Review*, *11*(3), 511–533.
[54] Zheng, Y. (2015). Diachronic development of free production vocabulary based on dynamic system theory. *Language Teaching and Research*, *47*(02), 276-288+321.
[55] Zheng, Y. (2016). The complex, dynamic development of L2 lexical use: A longitudinal study on Chinese learners of English. *System*, *56*, 40–53.
[56] Zhu, H., & Wang, J. (2013). The development of lexical richness in English Writing: A longitudinal study based on self built corpus. *Foreign Languages*, *06*, 77–86.