Lecture Notes on Language and Literature (2025)
Clausius Scientific Press, Canada

DOI: 10.23977/langl.2025.080311
ISSN 2523-5869 Vol. 8 Num. 3

Research on Review and Future Prospects of Lexical Richness

Manzi Zhu

School of International Education, Shandong University, Jinan, China

Keywords: Lexical Richness; Measurement; Chinese as a Second Language; Writing Quality; Trends

Abstract: Vocabulary richness serves as a crucial indicator of the diversity and sophistication of learners' language output, playing an essential role in assessing their linguistic development. This paper provides a preliminary review and synthesis of lexical richness research conducted since the 1980s. It introduces and evaluates pertinent studies and major perspectives on various aspects, including the formulation of measurement indicators, methodologies for assessment, developmental patterns of second language vocabulary richness, and the correlation between lexical richness and quality in second language writing. Additionally, we address key issues within existing studies on vocabulary enrichment and outline potential directions for future research.

1. Introduction

Vocabulary richness reflects the diversity and maturity of a learner's language output and is considered one of the key indicators of language development. Since the emergence of lexical diversity research in the 1980s, scholars both domestically and internationally have gradually integrated this concept into the field of second language writing, thereby enriching research content from both theoretical and practical perspectives, resulting in significant findings. Lexical richness encompasses two primary meanings: one is general; for instance, Read (2000) presents a "lexical richness" framework that includes lexical diversity, lexical complexity, lexical errors, and lexical density. The other meaning serves as an alternative term for lexical diversity itself, which is calculated and observed using similar methodologies (Fiona & Baayen, 1998). This paper adopts the broader definition and aims to systematically review research outcomes related to lexical richness since the 1980s by introducing and evaluating them with respect to conceptual positioning, measurement indicators, and domestic applications. Building on this foundation, we will discuss key issues and emerging trends within the realm of lexical diversity studies to elucidate future directions for research on vocabulary richness.[1]

2. Research on the measurement of lexical richness

The measurement of lexical richness has been a focal point in early research within this domain and serves as a foundation for applications such as enhancing conceptual models, examining the linguistic features of learners, and investigating the relationship between lexical richness and

writing quality. As experience has accumulated alongside advancements in data analysis and processing techniques, the vocabulary richness index system has become increasingly comprehensive and well-established. Moreover, its measurement methods have been refined to achieve greater precision and scientific rigor through iterative exploration and experimentation.

2.1. Construction of measurement indicators for lexical richness

From the 1980s to the early 21st century, the development of measurement indicators for lexical richness remained in an exploratory phase, with significant research findings primarily concentrated outside of China. Arnaud (1984) proposed that measurement methods could be categorized into three distinct types: lexical diversity, lexical complexity, and error index. Linnarud (1986) employed four methodologies to assess lexical richness: lexical individuality, lexical density, lexical complexity, and lexical diversity. However, subsequent studies discarded the dimension of lexical individuality due to its susceptibility to variations arising from group changes (Laufer & Nation, 1995). Laufer and Nation (1995) contended that four dimensions—lexical variability, lexical density, lexical complexity, and lexical novelty—are essential for measuring lexical richness.

Engber (1995) centered his analysis on lexicon-related errors and delineated four aspects of lexical richness: incorrect lexicon variability, error-free lexicon variability, proportion of lexicon errors, and lexicon density. In the context of second language writing research, Read (2000) synthesized these concepts into four components: lexical diversity, lexical complexity, lexical density, and error count. This multi-dimensional model is recognized for its systematic approach to classification and has gained widespread acceptance.

Since then, much of the discourse surrounding the measurement of lexical richness has been grounded in Read's framework. In later investigations within this framework; however; researchers began categorizing error counts separately as "lexical correctness" alongside "lexical richness." Consequently, only the first three dimensions will be addressed in subsequent discussions regarding measurement methods.

2.2. Methods for measuring lexical richness

2.2.1. Methods for measuring lexical diversity

Lexical diversity has evolved into the concept of lexical variability, which refers to the extent to which language users employ words in diverse manners with reduced repetition. High values may indicate a broad range of vocabulary [2]. In the progression of research on lexical diversity, one of the most influential metrics is the traditional Type-Token Ratio (TTR) calculation formula, originally proposed by Johnson (1944): TTR = Types / Tokens. This approach posits that lexical diversity can be quantified by the ratio of different types to total tokens within a text. However, this method is evidently influenced by text length; as text length increases, TTR values tend to decrease, making it challenging to accurately assess lexical complexity across texts of varying lengths [3].

Consequently, scholars have developed mathematical variants based on the TTR formula aimed at mitigating the impact of text length on assessments of lexical diversity. Notable examples include Carroll's modified TTR ($T/\sqrt{2N}$) [4] and Herdan's logarithmic TTR ($\log T/\log N$) [5]. Nonetheless, Vermeer found that these centralized measurement methods exhibit unsatisfactory reliability and validity [6]. Building upon this critique, Malvern et al. (2004) introduced a novel calculation method: RTTR = $V/\sqrt{N}$, which seeks to diminish the instability associated with measurement outcomes.

In addition to mathematical deformation, some scholars have transcended the inherent concepts of the Type-Token Ratio (TTR) formula and developed new measurement methods. For instance, Yule's

characteristic constant reflects lexical diversity through variations in word frequency distribution (Yule, 1944). Additionally, McKee (2000) proposed a D parameter formula aimed at determining the optimal D value by comparing actual and theoretical TTR curves. However, McCarthy (2005) demonstrated that neither of these calculation methods can eliminate the influence of discourse length. Building on this foundation, the Measure of Textual Lexical Diversity (MTLD) was introduced; it incorporates a segmentation point at 0.71 to divide discourse into several factor analyses, significantly enhancing reliability and validity. Furthermore, Scott (2008) presented the Standard Type-to-Type Ratio (STTR), which has been widely adopted in studies concerning Chinese as a second language.

Lu (2012) compared 20 measures of lexical diversity, and found that the total number of words and the square root TTR (total number of words /$\sqrt{Total\ number\ of\ words}$) are the most reliable means, while the Uber index is a commonly used measure of lexical diversity in English and Chinese second language vocabulary in China (Bao, 2008 ; wang Haihua, 2012 ; wu Jifeng, 2016).

2.2.2. Methods for measuring lexical complexity

Lexical complexity refers to the deliberate choice of utilizing low-frequency words that are more suited to the subject matter in writing, as opposed to opting for everyday common terms. This includes the use of technical terminology, jargon, and various other distinctive expressions to accurately convey the author's intended meaning (Read, 2000:200). The assessment of lexical complexity serves as an external measurement method for lexical richness, employing external standards such as word frequency tables [7]. Given that different languages possess unique characteristic systems, the subsequent discussion will explore these aspects from both international and domestic research perspectives.

Research on the measurement of lexical sophistication has been conducted internationally since the 1980s. Linnarud (1986) introduced the formula LS = Advanced Tokens/Tokens to assess lexical complexity, aiming to differentiate between native and non-native speakers. Laufer et al.'s theory of lexical frequency profile (LEP), in conjunction with Nation's (1984) seminal work "Vocabulary: Words, Affixes, and Roots," established a vocabulary framework comprising four categories: the most frequently used 1,000 words, the least frequently used 1,000 words, academic vocabulary, and additional words not included in these categories. This framework has proven reliable and effective in practical applications.

Daller et al. (2003) proposed two methodologies—Advanced TTR and Guiraud advanced—to investigate the diversity of advanced vocabulary within Turkish and German output texts; however, they did not examine their correlation with text length. Meara (2005) utilized the Lex30 method based on psycholinguistic associative approaches. He posited that "learners with a larger vocabulary are more likely to employ low-frequency words than those with a smaller vocabulary." By employing the P_Lex tool to compute metrics for vocabulary complexity, he significantly mitigated the influence of text length on measurement outcomes.

A common characteristic among these studies is their reliance on standardized word frequency tables. The most recognized methods for assessing lexical complexity are LEP and P_Lex[7].

The research results of vocabulary complexity measurement in China mainly lie in linking the previous research results of foreign scholars on measurement formulas with Chinese as a second language writing and Chinese vocabulary Outlines in practical measurement and application. For example, Huang Li and Qian Xujing (2003) defined the range of complex vocabulary as level B (including level B) and above in the Outline when the subjects were beginner learners of Chinese; Wang Yixuan (2017) used language materials from learners taking the high-level Chinese proficiency test, so the range of complex vocabulary was defined as words at level C, level D and beyond the

syllabus. Wu Jifeng (2023) used the formula "lexical complexity = number of complex words/$\sqrt{N}$" to measure the language characteristics of academic Chinese writing by second speakers of Chinese, where the number of complex words was based on the vocabulary Outline of the New HSK, taking words at level 5, level 6 and beyond the outline as complex words. The above research practice can provide a reference for subsequent studies to select external norms for assessing language complexity.

2.2.3. Lexical density measurement methods

The study of lexical density originated from Ure (1971), which suggested that lexical density could be measured by calculating the proportion of lexical words in the total number of words. Halliday (1985) developed Ure's formula to calculate lexical density by calculating the ratio of the number of content words to the total number of small sentences in the text. Laufer (1991) and Laufer&Nation (1995) respectively placed variables in the total number of word digits and the total number of word species for study. However, the effectiveness of vocabulary density as a tool for measuring the quality of a learner's writing has been controversial [8]. Read (2000) argues that vocabulary density is more suitable for measuring oral quality than writing quality. Lu's (2012) study suggests that vocabulary density does not reflect a learner's language proficiency. Wu Jifeng (2016) also demonstrated that lexical density is of no help in observing the quality of Chinese writing. Therefore, vocabulary density does not play a significant role in measuring learners' vocabulary richness.

3. Applied research on vocabulary richness

From the above analysis, it can be seen that in the early stages of lexical richness research and theoretical foundation research, foreign scholars have made a relatively large proportion of academic contributions, while the achievements of domestic scholars have mainly focused on the applied research of lexical richness. Therefore, this section is dedicated to summarizing the achievements of domestic applied research on lexical richness.

3.1. Overall research status and characteristics

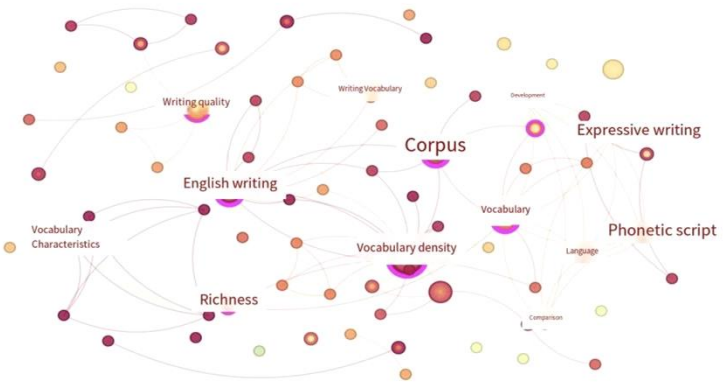

Figure 1: CNKI Vocabulary richness study Keyword Co-occurrence Knowledge graph

Using China National Knowledge Infrastructure (CNKI), a search was conducted under the theme

of "lexical richness", resulting in 74 literature results spanning from 1980 to 2024. A total of 40 papers were obtained after screening and removing conference minutes, irrelevant literature, etc. Keyword analysis was conducted using CiteSpace, and a co-occurrence knowledge graph of keywords was generated as shown in Figure 1.

By combining the keyword co-occurrence knowledge graph with secondary literature search, it can be seen that in the research direction of the application of vocabulary richness, domestic scholars mainly focus on the following aspects: English writing, composition quality, English proficiency, output tasks, etc. But the overall research is rather weak and scattered. Combining the summary of the three main lines on the application of lexical richness in existing review studies, namely the relationship between lexical richness and the quality of second language writing, the development characteristics of lexical richness in second speakers, and the similarities and differences between lexical richness in second speakers and native speakers [9], the first two aspects will be analyzed in detail below.

3.2. Development characteristics of vocabulary richness among second speakers

Early research on the development of vocabulary richness among second speakers, with domestic English second speakers as the subjects, has yielded abundant results. Tan Xiaochen (2006) divided the lexical enrichment development process of English major students in colleges and universities into three stages with significant differences in knowledge breadth, progress speed and comprehension depth. Bao GUI (2008) used the time-limited compositions on the same topic of students in three level groups: first-year graduate students of non-English majors, second-year students of non-English majors, and junior and senior students of English majors as corpora and found that there were significant differences in lexical complexity among different groups, and it developed in a linear manner; The other three dimensions of lexical richness showed no significant difference between adjacent groups and developed in a non-linear manner. Wan Lifang (2010) used the professional CET-4 and CET-8 compositions of college students in Shanghai as the corpus and found that the diversity and complexity of vocabulary in compositions increased significantly with the improvement of students' English proficiency. The study by Wang Haihua et al. (2012) further confirmed this view. Zhu Huimin et al. (2013) conducted a longitudinal study of 30 English argumentative essays on the same topic in a certain university and found that lexical variability in English writing showed a non-linear development trend, and lexical density and complexity generally increased, which was consistent with the conclusion of Xu Guoqin (2017). Zhang Huiping (2020) argues that English beginners in China have poor vocabulary richness, but it keeps improving with grade, and the development of various dimensions is uneven.

In recent years, the development characteristics of vocabulary richness among second speakers of Chinese have also become a hot topic. Zeng Dechun (2021) found that with the improvement of students' Chinese proficiency, the diversity and complexity of vocabulary in their writing have significantly increased, which is consistent with the development characteristics of English second speakers in China mentioned above. Wang Haoxue et al. (2022) examined the dynamic changes in the Chinese composition proficiency of Korean learners in China based on the "Chinese Interlanguage Corpus of Korean Learners in China" compositions and found that the number of word types and lexical complexity (the number of advanced word types) had the highest correlation with grade distribution. Wu Jifeng (2023), taking the academic Chinese writing texts of 225 Korean students of different levels as the research subjects, found that with the improvement of students' Chinese proficiency, the lexical diversity and complexity of their writing output increased significantly, while the lexical density failed to effectively distinguish all the different levels of second speakers. This is consistent with the conclusion drawn by Engber (1995), Wu Jifeng (2016), and Lee Chun-lin (2017).

3.3. The relationship between lexical richness and Second language writing quality

Early domestic research on the relationship between vocabulary richness and second language writing mainly focused on the field of English teaching. Liu Donghong (2003) found through a time-limited composition analysis of college students that vocabulary can improve the quality of writing by increasing the length of the composition, which constitutes an indirect effect. Qin Xiaoqing and Wen Qiufang (2007) found through their study of English compositions by college students that lexical variability and lexical complexity were positively correlated with composition scores. Bao GUI (2008) studied the second language vocabulary application ability of efficient English major students in China and found that vocabulary richness was moderately correlated with the quality of second language writing.

With the rise of the Chinese language craze and the increase of international students in China, research on writing Chinese as a second language has begun to emerge widely. Wu Jifeng (2016) found that the combination of four independent variables - lexical variability, lexical complexity, lexical density, and lexical error rate - could account for 46.2% of the total variation in the second language Chinese writing performance of native English speakers. Wang Yixuan (2017) found that three parameters, namely the number of word types, the proportion of lexical errors and the number of common words, could account for 92.8% of the variation in composition scores. Wu Jifeng (2019) found that lexical diversity and lexical complexity were significantly associated with writing scores and content quality scores based on 210 compositions of Korean students with different levels of Chinese proficiency. Among them, lexical complexity was the most significant predictor of content quality scores. Zhang Juanjuan (2019) found that lexical richness could account for 71.4% of the variation in Chinese as a second language composition scores. Xiao Huimin et al. (2021) studied the relationship between language features and writing quality using the picture-based writing texts of 132 Indonesian children as the corpus and found that the number of keyword types could significantly predict the total variation in writing scores.

4. Conclusion

Looking back on the development of lexical richness research over the past 40-plus years, in general, some progress has been made in this field: the framework of lexical richness measurement indicators has been basically established, and the measurement methods in various dimensions have been repeatedly improved to achieve high reliability and validity; On this basis, scholars at home and abroad have produced rich results on the development characteristics of second speakers' vocabulary richness and the relationship between vocabulary richness and second language writing, which are of great guiding significance for the study of language acquisition and language output. The overall analysis also shows that lexical richness research is still relatively weak, especially in China, which is mainly manifested in three aspects: First, the overall number of papers related to lexical richness is not large, and there are many research gaps; Second, the measurement methods of lexical complexity still cannot fully overcome the influence of text individuality, are constrained by text length and overall coherence, and the complex measurement formulas simultaneously increase the difficulty of promotion; Third, there is an imbalance in vocabulary richness research at home and abroad. Most of the fundamental achievements and theoretical frameworks in the field have been constructed by foreign scholars, while the achievements of domestic scholars are concentrated at the application level, and the procedures for measuring indicators are more of inheritors and users rather than modifiers. These deficiencies constitute constraints on the development of lexical richness research and are issues that scholars in the field are striving to solve. Therefore, the following aspects may be the future development trends of lexical diversity research and also the entry points for the comprehensive improvement of the current state of the discipline.

Automatic extraction empowers feature measurement, and big data supports fine processing. With the advancement of information processing technology and the development of automation technology, the extraction of language features represented by lexical richness will be more efficient and accurate, providing technical support for the promotion and application of complex calculation formulas. At the same time, the independent development of lexical richness measurement programs by scholars at home and abroad will also become an important means for countries to gain a say in language research. Under this technical condition, not only is it possible to measure lexical richness free from the constraints of text length, but it also indicates the coexistence of multiple measurement methods and the formation of personalized research approaches suitable for different types of texts.

Chinese as a second language will yield abundant research results, and multilingual contrast will become an important paradigm. Although research on lexical richness in Chinese as a second language started relatively late, it has shown a trend of rapid development and diverse output against the backdrop of an increasing number of international students coming to China in recent years. The application of lexical richness in the quality and acquisition rules of Chinese as a second language writing is bound to become a hot research topic. Multilingualism, as an important feature of the language world in the present era, also influences the research content of lexical richness, making the comparison and experience learning among different languages a new direction.

The study of the language itself is becoming more profound, with interdisciplinary and multi-directional functional exploration. At present, scholars in China have been greatly influenced by structuralism in the study of lexical richness. In the future, more studies may start from a functionalist perspective, focusing on the usage function and communicative effect of vocabulary in real contexts. The introduction of a psychological perspective into the measurement of lexical richness indicators is a positive trend in interdisciplinary research, and in the context of the new liberal arts, research in this field will also be more combined with other disciplines to enrich its output.

Finally, due to space limitations, this paper fails to discuss the influencing factors of vocabulary richness in teaching activities, which is a certain limitation and an important direction for improvement in subsequent studies.

References

[1] Bao GUI. A multi-dimensional study on the development of vocabulary richness and vocabulary of second language learners [J]. Foreign language electronic teaching, 2008,(05) :38-44.

[2] Zhang Yan, Chen Jiliang. Quantitative methods for measuring lexical richness in speech output [J]. Foreign Language Testing and Teaching,2012,(03):34-40.

[3] RICHARDS B. Type /token ratios: what do they really tell us[J]. Journal of Child Language, 1987,14 (2) : 201-209.

[4] CARROLL J B. Language and thought[M]. Englewood Cliffs, NJ: Prentice-Hall, 1964.

[5] HERDAN G. Type-token mathematics: a textbook of mathematical linguistics[M]. The Hague: Mouton, 1960.

[6] VERMEER A. Coming to grips with lexical richness in spontaneous speech data[J]. Language Testing, 2000,17 (1) : 65-83.

[7] Lu Yun. Lexical richness measurement methods and computer program development: Review and Prospect [J]. Journal of Nanjing Tech University (Social Sciences Edition),2012,11(02):104-108.

[8] Wang Yixuan. Correlation between Lexical Richness and Writing performance of Second Speakers of Chinese - also on multiple Linear Regression models and equations for measuring writing quality [J]. Language application, 2017 (02) : 93-101. The DOI: 10.16499 / j.carol carroll nki. 1003-5397.2017.02.011.

[9] Zhu Huimin, Wang Junju. The development characteristics of lexical richness in English writing - A longitudinal study based on a self-built corpus [J]. Foreign Language World,2013,(06):77-86.