

Study on the Evolution of Lexical Richness in Chinese Language Textbooks for Primary and Secondary Schools Since the Founding of New China

Xiaoxuan Zheng^{1,a}, Yongquan Li^{1,b,*}

¹College of Teacher Education, Quzhou University, Quzhou, Zhejiang, China

^azzhengxiaoxuan@126.com, ^byongquanli@qzc.edu.cn

*Corresponding author

Abstract: The evolution of lexical richness in Chinese language textbooks is closely intertwined with national education and social-cultural development. This article is based on the Chinese primary and secondary school Chinese language textbook corpus from 1949 to 2025, using a combination of quantitative and qualitative methods to systematically examine the historical evolution characteristics and influencing factors of lexical richness in textbooks. Research has found that: (1) Lexical richness presents a phased development: in primary school, the number of semantic categories in 1955 was mostly single digit to double digit, with a significant increase after 1978 (mainly three digit), reaching a peak in 1987 (some exceeding 600), and fluctuating upward after 2000; In the early stage of junior high school, the values were relatively high (mostly in the three digits in 1960, with some exceeding 1000), and after 1978, they fluctuated and remained high. (2) There is a significant difference in grade levels, with the number of semantic categories generally higher in junior high school than in primary school, reflecting the progressive nature of educational content. (3) The evolution trajectory is driven by multiple factors such as social culture, education policies, and cognitive laws, manifested as a transition from political discourse dominance (1949-1977) to knowledge system diversification (1978-1999), and then to a balance between cultural connotations and cognitive adaptability (after 2000). The study also constructs a lexical richness evaluation system (quantity, type, semantic dimension), combined with Type-Token Ratio (TTR) and semantic category statistics, to reveal the nonlinear development law of the vocabulary ecology in textbooks, providing theoretical basis and practical reference for optimizing textbook writing and promoting language proficiency development.

Keywords: Chinese Language Textbooks; Lexical Richness; Evolution over Time; Semantic Category; Education Policy

1. Introduction

1.1 Research Background and Significance

As the core carrier of language inheritance and cultural dissemination in basic education, the evolution of vocabulary systems in Chinese language textbooks reflects the historical trajectory of national ideology, educational philosophy, and social-cultural changes. For example, during the “Education Revolution” in 1958, textbook vocabulary was dominated by political slogans, while after the 1988 reform, with the shift to economic construction, scientific and technological terms significantly increased. Since 1949, textbooks have undergone institutional changes from “national unified use” to “multiple versions under one syllabus” and then to “unified compilation” with lexical richness showing distinct diachronic characteristics in quantity, type, and semantic function. This evolution not only follows linguistic laws but also correlates closely with national educational strategies, cultural policies, and disciplinary goals.

Lexical Richness as a key indicator of textbook quality Chinese language textbooks for primary and secondary schools are vital for students’ language development, and lexical richness directly impacts their linguistic literacy. Diverse and appropriate vocabulary expands students’ word banks, enhances expressive precision, and cultivates thinking and aesthetic abilities through knowledge and cultural transmission. Existing research indicates that shared vocabulary in textbooks accounts for less than 40%, underscoring the need for dynamic update mechanisms [1]. Monotonous or obscure vocabulary may

hinder cognitive development or reduce learning efficiency, making the exploration of evolutionary patterns crucial for textbook optimization and educational quality improvement. Therefore, systematically analyzing the mechanisms of lexical richness evolution holds theoretical and practical value for deepening the understanding of Chinese education and optimizing textbook compilation.

1.2 Definition of Core Concepts

Lexical richness is a composite concept spanning multiple dimensions. Quantitatively, it refers to the total number of distinct words in textbooks, reflecting the breadth of knowledge and linguistic diversity. It involves the richness of word classes (nouns, verbs, adjectives, adverbs) and the proportion of categories like technical terms, daily expressions, literary vocabulary, and scientific terms. Semantically, it encompasses the roles of words in conveying knowledge, expressing emotions, shaping images, and reflecting cultural meanings. Additionally, it must align with students' cognitive levels—neither too simplistic to meet learning needs nor too abstruse to cause comprehension barriers.

2. Related Work

2.1 Exploration into the Study of Lexical Richness

Measurement Methods of Lexical Richness in Textbooks Early foreign studies on measuring lexical richness in textbooks relied on simple statistics like word count and frequency. With advancements, more sophisticated methods emerged, such as the Type-Token Ratio (TTR) for vocabulary diversity and corpus-based analysis to evaluate semantic richness through contextual variations and collocation patterns [2]. Domestic studies have adapted foreign methodologies to Chinese characteristics, incorporating factors like Chinese character stroke counts and structural types into difficulty and richness assessments [3]. They also emphasize practical teaching needs, using student feedback from textbook reading to indirectly evaluate the relationship between lexical richness and learning outcomes [4].

Research on the Evolution of Lexical Richness Foreign studies on textbook lexical evolution often focus on national educational developments. For example, after the U.S. education revival movement, professional and literary classic vocabulary increased, forming a trajectory from “practical life” to “academic advancement”. Domestic research on Chinese textbooks shows that from 1949 to 1978, vocabulary was politically dominated and limited; post-1978 reforms saw increased diversity; and since 2000, quality-oriented education has enriched semantic functions in knowledge transmission, cultural inheritance, and emotional education. However, domestic studies still lack systematic, in-depth analysis of complex interactive factors in evolution.

2.2 A brief Summary of the Historical Evolution of Lexical Richness in Chinese Language Textbooks

Foundational Period (1949–1977). Textbooks exhibited simple, politically dominated characteristics. Amid national reconstruction, education prioritized basic knowledge and political ideology, resulting in small vocabulary size, limited word types, and frequent political terms like “people” and “revolution”. Nouns and verbs dominated, with adjectives and adverbs used straightforwardly, serving primarily propagandistic functions.

Developmental Period (1978–1999). With reform and opening up, textbooks entered a phase of knowledge-systematic diversification. Total vocabulary increased significantly, with new scientific/technical and literary terms (e.g., microscope, single-cell in the 1988 Undersea World edition, up 17% from 1961). Nouns became more diverse, verbs emphasized abstraction, and semantic functions shifted toward multi-knowledge dissemination.

Deepening Period (2000–2025). Under quality-oriented education, lexical richness deepened in “cultural connotation” and “cognitive adaptation”. Textbooks adopted hierarchical vocabulary systems, increased traditional cultural terms, and incorporated neologisms. Semantic functions integrated multiple values (scientific cognition, literary aesthetics, ecological awareness), as seen in the 2011 curriculum standards promoting thematic vocabulary and local textbooks highlighting regional characteristics, evolving from political dominance to pluralistic value integration.

3. Construction of the Evaluation System for Lexical Richness

3.1 Theoretical Basis and Dimension Design

Quantity: Based on total vocabulary, analyze the hierarchical distribution of basic common words, advanced expanded words, and rare words to reflect learning difficulty and progression.

Type: Emphasize balanced word-class distribution and diversity of categories (e.g., technical terms, daily language) to ensure comprehensive expressive functions.

Semantics: Focus on polysemy, metaphorical usage, and contextual flexibility to cultivate comprehension and associative skills.

3.2 Construction of Chinese Language Corpus for Primary and Secondary Schools

We have currently collected 3413 Chinese language texts for primary and secondary schools by collecting information from libraries, databases, and other sources. Among them, there are 2318 primary school texts and 1095 junior high school texts. The time span of the textbook is from 1949 to 2025. We are still expanding the scale of our corpus. The specific corpus introduction is shown in Figure 1.

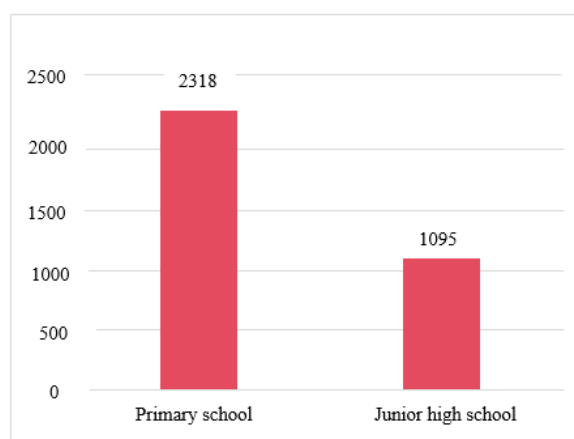


Figure 1: The number of textbooks included.

Figure 2 shows the number of texts available in each era. The horizontal axis represents the ID number of the text, and the vertical axis represents the era in which the text is located.

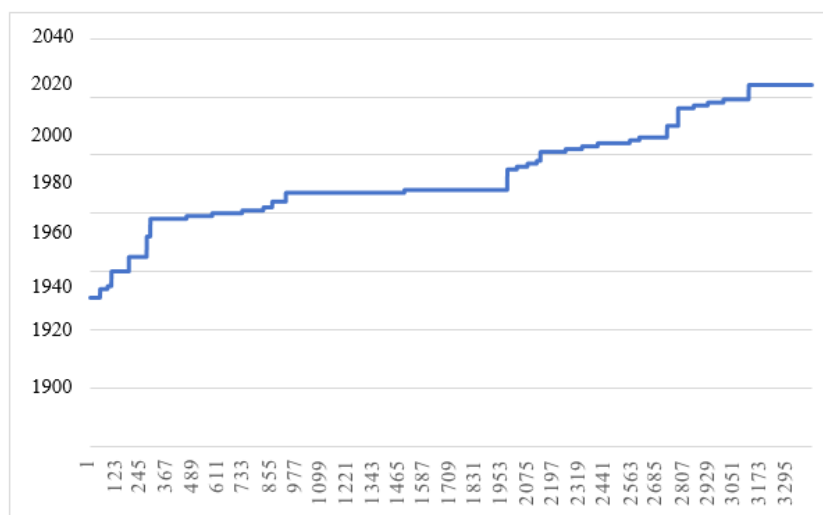


Figure 2: The year of textbooks included.

3.3 Analysis of Lexical Richness in Corpus Text

With the help of Stanford University's CoreNLP segmentation tool (version 4.5.9), we conducted a

detailed investigation into the number and types of words in the corpus text. The results are shown in Figures 3 and 4.

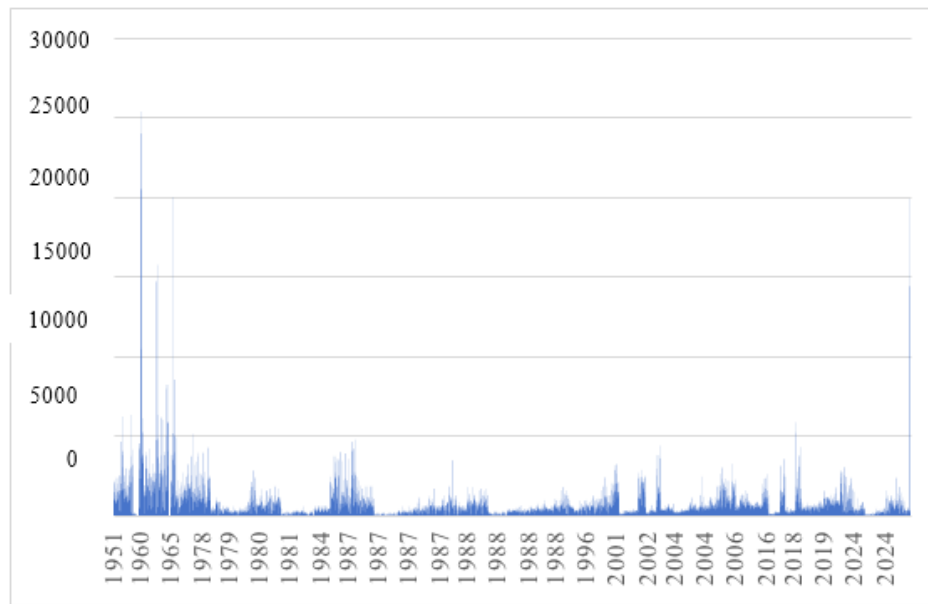


Figure 3: The statistics of the number of words in corpus.

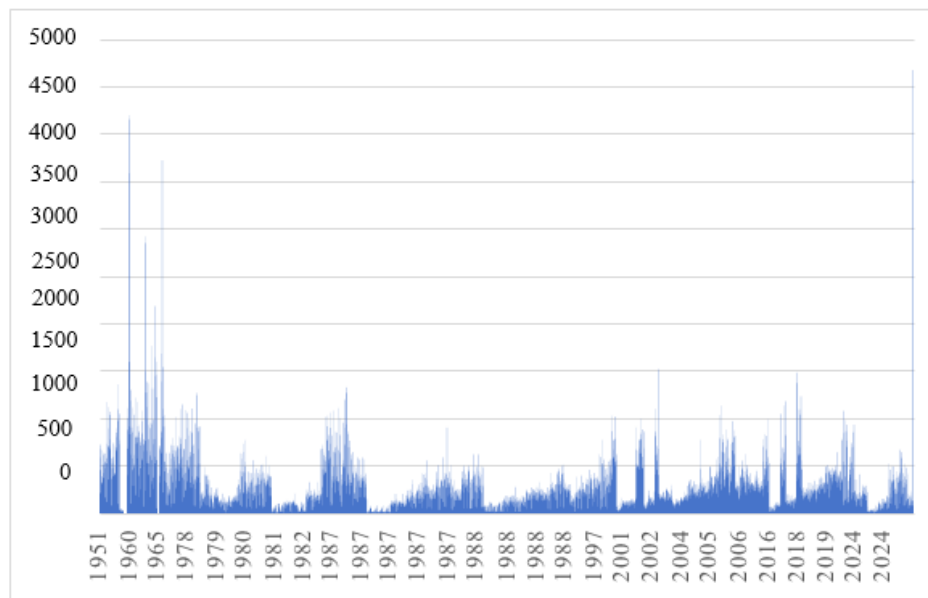


Figure 4: The statistics of the type of words in corpus.

In Figures 3 and 4, it is evident that Chinese language textbooks published between 1960 and 1965 had significantly higher types and numbers of words compared to other time periods. One is the uneven data density between years, such as the unusually dense data in 1951 and relatively sparse data in other years; The second is that the ratio between the number of words and the number of words shows a non-linear change over time, with extreme peaks in vocabulary diversity (number of words) occurring in both the early and recent periods. These characteristics indicate that in order to accurately interpret data trends, potential influencing factors such as historical background, text type, and data collection methods must be considered.

The type token ratio (TTR) is the ratio of all types to all tokens in a text, where tokens refer to the number of words in a text and tokens refer to the number of tokens that are not duplicated. This means that in a text, repeated tokens can only be considered as one type token. We calculated the TTR numbers of the corpus, as shown in Figure 5. Obviously, the TTR values in the text exhibit the characteristics of a compound normal distribution. But there are also many abnormal values.

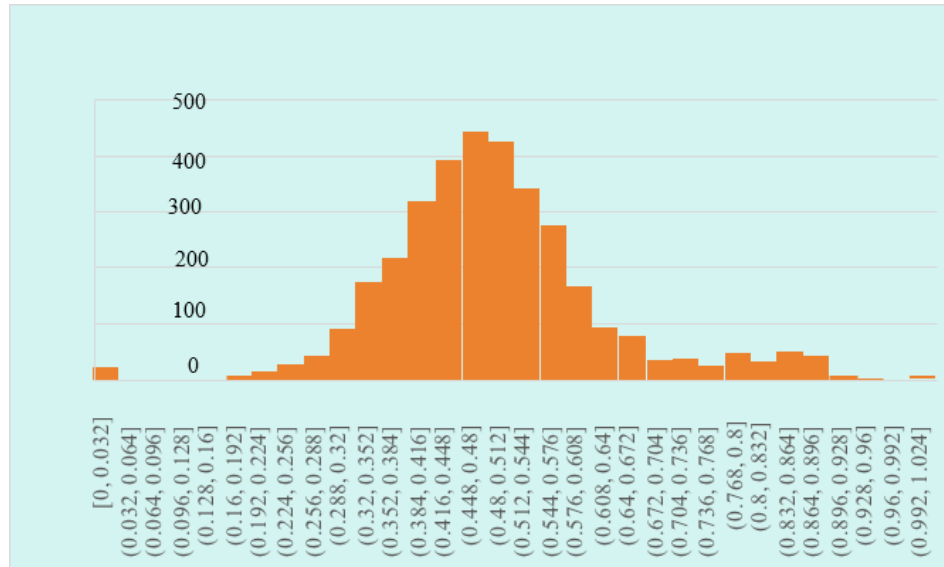


Figure 5: The TTR detail statistics.

From the longitudinal comparison of semantic richness, the data shows a non-linear development trajectory of “peak stable new high”. The drastic fluctuations in early semantic systems gradually gave way to a relatively stable vocabulary ecosystem, but the recent high values indicate that semantic systems may be undergoing a new round of expansion. This pattern of change implies that the semantic capacity of language systems does not increase linearly, but rather exhibits a transitional development triggered by specific social, cultural, and technological changes. Meanwhile, the persistent existence of zero value records also indicates significant differences in semantic richness across different domains or text types, and this heterogeneity constitutes an important feature of language ecology.

Referring to the Yiwu Jiaoyu Changyong Cibiao (List of Common Words in Compulsory Education) [5] and Xiandai Hanyu Cidian (Modern Chinese Dictionary) [6], we conducted a statistical analysis of the number of semantic categories in the texts in the corpus. The calculation results are shown in Figure 6.

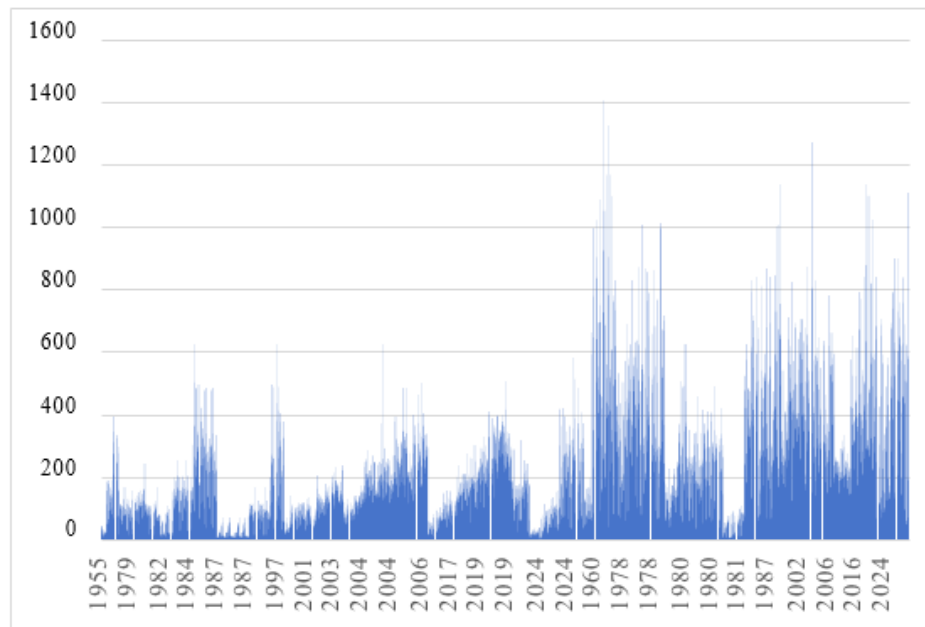


Figure 6: The semantic categories detail statistics.

Based on the characteristics of corpus text and semantic category data, it can be found that in terms of time span, the number of semantic categories in primary school was relatively low in 1955 (mostly single or double digits), significantly increased after 1978 (mostly three digits), reached its peak in 1987 (some exceeding 600), and fluctuated greatly after 2000 but overall higher than in the early stages; In the

early stage of junior high school, the values were relatively high (mostly in the three digits in 1960, with some exceeding 1000), and after 1978, there were significant fluctuations but still maintained a high level. Overall, the number of semantic categories is showing an increasing trend, which may be related to the deepening of educational content and textbook reform.

From the perspective of grade differences, the number of semantic categories in junior high school is generally higher than that in primary school, reflecting the progressive nature of educational content. Early primary school datasets (such as 10-40 in 1955), with increased dispersion in later periods (such as from single digits to 623 in 1987); The fluctuation range of middle school values is larger (such as from 90 to 1404 in 1965). The time distribution of data is uneven, with more records in some years (such as 1978 and 1987) and less data in years such as 1960. However, the overall trend is clear, reflecting the gradual enrichment of semantic categories in Chinese language education.

4. Conclusions

This study systematically examines the lexical richness of Chinese language textbooks for primary and secondary schools from 1949 to 2025, revealing its historical evolution patterns and influencing factors. Research has found that the richness of vocabulary in textbooks exhibits clear phased characteristics: from the dominance of political discourse in the early days of the founding of the People's Republic of China, to the diversification of knowledge systems after the reform and opening up, and to the equal development of cultural connotations and cognitive adaptability since the new century. The study used a combination of quantitative and qualitative methods to construct a three-dimensional evaluation system for vocabulary richness, providing a new analytical perspective for textbook development and language education research.

However, there are still several shortcomings in this study: firstly, due to limitations in obtaining historical textbooks, the sample size for some years is insufficient, which may affect the representativeness of the data; Secondly, the exploration of the matching relationship between vocabulary difficulty and students' cognitive level is not deep enough, and it can be verified through psychological experiments in the future. Future research can continue to deepen in the following areas: firstly, expanding the corpus size, especially supplementing local versions of textbooks, to enhance the comprehensiveness of research; The second is to introduce cognitive science methods to explore the acceptance of vocabulary richness among students of different age groups; The third is to combine artificial intelligence technology to develop a dynamic vocabulary evaluation system, providing real-time feedback for textbook revision. These improvements will help establish a more scientific vocabulary evaluation system for textbooks and promote the innovative development of Chinese language education.

Acknowledgements

This research was supported by the following projects:

- 2024 Quzhou University the construction project of the “Mentor+Project+Team” Innovation and Entrepreneurship Workshop, “Smart Book Boat” (SH20241101)
- 2024 the Education Teaching Reform Project of the Education Teaching Guidance Committee of the Ministry of Education in China, “Design and Practice Research on Personalized Reading Material Matching System for Compulsory Education Chinese under the Drive of Large Language Model” (2024JGYB026)
- 2023 Quzhou University Doctoral Research Launch Fund Project, “Study on the Readability of Chinese Educational Reading Materials for Graded Reading” (BSYJ202306)
- 2024 Zhejiang Provincial Philosophy and Social Sciences Planning Project, “Study on the Readability of International Chinese Education Reading Materials for Graded Reading” (24NDQN034YB)
- 2024 Yunnan Provincial Department of Education Science Research Fund Project, “Research on Rhetorical Application of Chinese Reading Materials for Young Children in Lancang Mekong Coast Countries” (2024J0915)
- 2024 the Research Project on Education and Teaching Reform of Education Majors in Vocational Colleges of the Ministry of Education in China, “Research on Intelligent Chinese Education in Catering under the Background of International Chinese Education” (JZW2024-ZWB-7)

References

- [1] Lai Huaqiang. *Quantitative research on Chinese vocabulary: An urgent project*. *Language Planning*, 2006, (7), 59-61. DOI: 10.16412/j.cnki.1001-8476.2006.07.034
- [2] Coady J, Huckin T. *Second language vocabulary acquisition: A rationale for pedagogy*[J]. *Annual Review of Applied Lingus*, 1996, 9(9):92-106. DOI:10.1017/S0267190500000829.
- [3] Zhu Chaochao. *Research on Textbook Language Based on Word Frequency Statistics-Taking Chinese language in primary and secondary schools as an example* [D]. Hangzhou: Zhejiang University of Finance and Economics, 2019.
- [4] Yu Lixiang. *Research on Vocabulary Statistics and Rank Examination of Primary School Chinese Textbooks by People's Education Press* [D]. Wuhan: Central China Normal University, 2019. DOI: 10.27159/d.cnki.ghzsu.2019.000117
- [5] Language and Text Information Management Department of the Chinese Ministry of Education. *Common Vocabulary List for Compulsory Education* [M]. Beijing: The Commercial Press, 2019.
- [6] Institute of Linguistics in Chinese Academy of Social Sciences. *Modern Chinese Dictionary (7th edition)* [M]. Beijing: The Commercial Press, 2016.