# *A Comparative Study on Lexical and Syntactic Features of High-Scoring and Low-Scoring Writings in First Year High School English Application Writing*

**Yihan Zhang**

*Beijing Foreign Studies University, Beijing, China*
*18133819691@163.com*

***Abstract.*** This study investigates lexical and syntactic differences between high-scoring and low-scoring English practical writings of first-year high school students. Based on 40 exam writings selected from the top and bottom 20% of students at Shijiazhuang Foreign Language School, lexical diversity was measured using Type-Token Ratio (TTR), and syntactic complexity was assessed through three L2SCA indicators: Mean Length of T-unit (MLT), Dependent Clauses per Clause (DC/C), and Complex Nominals per T-unit (CN/T). Results show that high-scoring writings exhibited slightly higher lexical diversity, but the difference was not statistically significant. In contrast, significant differences emerged in MLT and CN/T, indicating that high-scoring writers use longer structural units and more complex noun phrases. DC/C showed no significant between-group difference. The study posits that in high school English practical writings, syntactic complexity, particularly sentence length and noun phrase complexity, plays a more critical role in differentiating writing ability, while lexical diversity contributes less. These findings provide empirical implications for teaching high school English practical writing.

***Keywords:*** Lexical diversity, Syntactic complexity, English practical writing, High school students

## 1. Introduction

In high school, fostering students' writing ability is a crucial aspect of teaching. The "Examination explanation for the National College Entrance Examination" requires candidates to use grammar and vocabulary accurately, employ a variety of sentence structures to express ideas clearly and coherently with transitional devices. However, in teaching practice, a marked disparity is observed in the English practical writing skills among first-year high school students. Consequently, it is crucial to systematically analyze the lexical and syntactic differences between high-scoring and low-scoring practical writings, thereby identifying these linguistic differences and implementing targeted instruction.

Existing research on the high school English writing characteristics mainly focuses on students preparing for the Gaokao or the general high school population, with specialized studies on first-year high school students being relatively scarce. For instance, Zhong explored the developmental

characteristics of syntactic complexity in the practical writing of high school students overall [1]. Liu concentrated on the Gaokao writings of third-year students, analyzing the nominal phrase syntactic complexity features across different score levels [2]. However, research indicates that the development of writing ability in the first year directly influences language progression throughout high school [3], and necessitates targeted pedagogical intervention [4]. Therefore, it is imperative to conduct an in-depth analysis of the linguistic features of first-year students' writings.

This study uses practical writing samples from the first-year final English examination at Shijiazhuang Foreign Language School. High-scoring writings (top 20%) and low-scoring writings (bottom 20%) are selected for comparative analysis. The research examines lexical diversity and syntactic complexity through the Type-Token Ratio (TTR) metric and core indicators of the L2SCA, respectively, to contrast the lexical and syntactic differences between the two groups. The study aims to identify the linguistic characteristics of high-scoring and low-scoring practical writings among first-year students, thereby providing targeted pedagogical recommendations.

By investigating the distinctions in lexical and syntactic features between high- and low-scoring writings, this research addresses the following questions: In terms of lexical diversity, what specific differences exist between high-scoring and low-scoring practical writings in the first-year English examination? Based on the L2SCA syntactic complexity analysis framework, what disparities are evident in syntactic complexity between these two groups?

## 2. Literature review

### 2.1. Lexical diversity and Type-Token Ratio (TTR)

Lexical diversity is a crucial metric for assessing second language writing proficiency, typically encompassing dimensions such as lexical diversity, lexical complexity, and lexical density. Among these, lexical diversity is considered a core aspect reflecting a learner's lexical usage ability, with the TTR being a commonly employed quantitative measure for assessing it.

In the TTR metric, types refer to each unique word form appearing for the first time in a text, while tokens refer to all word forms occurring in the text. The TTR is the ratio of types to tokens, calculated as TTR = (number of types/number of tokens) * 100%. Over the past few decades, research on TTR has continuously emerged. Research on TTR abroad started earlier, focusing on multiple research directions. For instance, in studies correlating TTR with writing quality, Engber examined 66 second language learners, calculating the TTR of their timed writings using two methods [5]. The results indicated that after excluding vocabulary errors, TTR showed a significant positive correlation with writing proficiency, a conclusion supported by subsequent research [6, 7]. Domestic research, on the other hand, often draws upon foreign frameworks, emphasizing localized validation and expansion. Wang compared argumentative writing written by Chinese and American university students and found that participants' lexical collocation ability, sentence patterns, and discourse structure were also closely related to differences in type-token frequency [8]. Chinese students tended to overuse frequently mastered vocabulary and keywords related to the writing topic, with token frequency being overly concentrated in specific semantic categories. Bao found that students with lower English proficiency showed significantly lower TTR than their mid-to-high proficiency peers, and advanced learners tended to use more unique vocabulary [9]. Wan, in a study of 100 English majors from four universities in Shanghai, found that TTR showed an upward trend with the progression of learning [10]. To sum up, existing research has sufficiently confirmed the effectiveness of TTR in evaluating lexical diversity in second language writing.

### 2.2. Syntactic complexity and L2SCA

Syntactic complexity is a vital dimension for measuring written output in second language acquisition. Scholars both domestically and internationally have conducted extensive research in this area, proposing approximately 40 specific indicators across various dimensions to measure syntactic complexity. As research deepens, measurement indicators have evolved from a singular focus on the length of language production to a more comprehensive, multi-dimensional, multi-indicator approach [1].

Early research primarily used the length of linguistic output as the core measurement indicators, such as the average length of sentences, clauses, and T-units, which became the most fundamental and widely applied indicators at the time [11]. With further research, Norris & Ortega conceptualized syntactic complexity as a multidimensional construct including overall sentence complexity, complexity through subordination, complexity through coordination, and complexity through phrasal elaboration [12]. Based on their research, Lu further established a five-dimensional framework including overall sentence complexity, subordination, coordination, phrasal elaboration, and sentence structure diversity, emphasizing the need to select targeted indicators based on specific research contexts [13].

With the advancement of computer technology, the measurement of syntactic complexity indicators has gradually shifted from manual annotation to automated measurement. Currently, widely used tools include Biber Tagger, Coh-Metrix, L2SCA, and TAASSC [1]. Among these, the L2 Syntactic Complexity Analyzer (L2SCA), developed by Lu Xiaofei, has gained widespread recognition among domestic and international scholars due to its strong adaptability and ease of operation. The L2SCA, with its five dimensions and 14 indicators, comprehensively captures the core facets of syntactic complexity. Previous scholars have also conducted numerous studies using L2SCA. For instance, Zhong used L2SCA to analyze a corpus of 995 argumentative writings by high school students, showing the developmental process of syntactic complexity across different grade levels [1]. Peng Jinhua examined argumentative writings by 11th-grade students, illustrating the relationship between syntactic complexity and writing quality [14]. The validity and reliability of the L2SCA framework have been sufficiently verified in existing research. And this study will use core dimensions of L2SCA to investigate the syntactic complexity contrast between high-scoring and low-scoring practical writings of first-year high school students.

## 3. Research design

### 3.1. Research setting

This research targets all first-year high school students at Shijiazhuang Foreign Language School, using a timed writing task from their final English examination as the research context. The first-year high school students of this school are in different classes under liberal arts and science based on their comprehensive abilities. The English teaching progress of students in each class is basically the same. When conducting sampling, the selection was also carried out evenly from each liberal arts and science class based on the composition scores to ensure the representativeness of the sample.

The writing task was an English practical writing task in the genre of an invitation letter. Students were required to write as Li Hua, inviting a foreign teacher to attend the closing ceremony of the English Culture Festival. In the writing, three key points, the time of the event, its main content, and its significance, should be involved. This writing task was scored out of 20 points, with assessment

dimensions including content completeness, lexical richness, syntactic variety and accuracy, discourse coherence, and formatting norms. All scripts were independently graded by two teachers.

### 3.2. Data collection

This study used stratified sampling for participants and data selection. The initial corpus comprised 257 final examination papers from all the classes. Prior to any analysis, informed consent was obtained from both students and the school, and all writings were anonymized. Subsequently, 40 valid writings were selected as the research sample and segmented into a high-scoring group (top 20%) and a low-scoring group (bottom 20%) based on their writing scores. To ensure data integrity, samples were screened and replaced post-sampling, excluding writings with word counts below 80 or exceeding 150 words, as well as those significantly deviating from the prompt. Following screening, all selected writings were converted into TXT form for analysis.

### 3.3. Scoring standards

The scoring standards used in this study align with the principles of the College Entrance Examination English writing assessment, adopting a five-band scoring system (0-20 points). The scoring primarily considers the following aspects: completeness of content points, richness and accuracy of lexical usage, diversity and correctness of grammatical structures, discourse coherence and cohesion, format adherence, and legibility of handwriting.

Based on pedagogical experience, high-scoring writings typically show rich vocabulary, diverse sentence structures, and coherent expression, whereas low-scoring writings tend to demonstrate more limited lexical and syntactic resources.

### 3.4. Lexical measurement

This study adopts Type-Token Ratio (TTR) as the core measure of lexical diversity. In English writing, learners inevitably repeat certain words. Thus, a higher TTR indicates lower lexical repetition and greater lexical diversity. In this study, TTR is defined as the proportion of distinct words to the total number of words in a composition, and its formula is as follows:

TTR = (Number of distinct words / Total number of words) * 100%

During the statistical analysis, misspelled words and punctuation were excluded, word forms were normalized where necessary, and only content words (nouns, verbs, adjectives, and adverbs) were tallied. Lexical data analysis was conducted using Excel.

### 3.5. Syntactic measurement

This study employed the Second Language Syntactic Complexity Analyzer (L2SCA) to analyze syntactic complexity. The validity study report of the analyzer indicates that the accuracy rate for language unit identification ranges from 83% to 100%, with a correlation between automatic and manual measurements between 83.4% and 94.1%. Minor errors in sample writing, such as punctuation, capitalization, spelling, and singular/plural discrepancies, do not impact the analytical results [15]. Given the small sample size, and to ensure precision, this study analyzed 3 out of 14 selected indicators:

(1) Mean Length of T-unit (MLT): This metric measures the average number of words per T-unit, reflecting a student's capacity to convey information within a single sentence.

(2) Dependent Clauses Per Clause (DC/C): This measures the proportion of dependent clauses to total clauses, encompassing noun clauses, relative clauses, and adverbial clauses commonly used by 10th-grade students.

(3) Complex Nominals Per T-unit (CN/T): This quantifies the number of complex noun phrases per T-unit, indicating the extent to which students employ modification structures within noun phrases.

### 3.6. Data analysis methods

After obtaining lexical and syntactic indicator data, NeoSCA was employed for statistical analysis. In addition to comparative analyses of data between high-scoring and low-scoring groups, independent samples t-tests were used to determine the statistical significance of differences between the two groups.

## 4. Analysis results

By importing sample compositions into Excel and NeoSCA for analysis, specific differences in lexical diversity and syntactic complexity between the high-scoring and low-scoring groups were identified. Detailed data are presented in Tables 1-3.

Table 1. Descriptive statistics for lexical diversity

| Group | Mean TTR | Std. Deviation | T-value | P-value |
|---|---|---|---|---|
| High-score | 0.6615 | 0.0411 | 0.8043 | 0.4270 |
| Low-score | 0.6483 | 0.0610 | | |

Table 2. Descriptive statistics for syntactic complexity

| Group | MLT | MLT | DC/C | DC/C | CN/T | CN/T |
|---|---|---|---|---|---|---|
| High-score | 14.6666 | 1.8359 | 0.3093 | 0.0917 | 1.4590 | 0.3269 |
| Low-score | 11.4734 | 1.8215 | 0.2753 | 0.0916 | 1.0391 | 0.2218 |

Table 3. Independent-samples t-test results

| Indicator | T-value | P-value | Significance |
|---|---|---|---|
| TTR | 0.8043 | 0.427 | Not significant |
| MLT | 5.5217 | <0.001 | Significant |
| DC/C | 1.1740 | 0.248 | Not significant |
| CN/T | 4.7536 | <0.001 | Significant |

### 4.1. Lexical diversity analysis

Regarding lexical diversity, the high-scoring group had a total token count of 2,721, a type count of 1,793, a mean TTR of 0.6615, and a standard deviation of 0.0411. The low-scoring group had a total token count of 2,207, a type count of 1,424, a mean TTR of 0.6483, and a standard deviation of 0.0610. Descriptive statistics show that the high-scoring group was slightly higher than the low-scoring group in total tokens, types, and mean TTR, indicating a minor advantage in vocabulary

breadth. However, the independent samples t-test results revealed a p-value of 0.4270 for the difference in TTR between the groups, which did not reach statistical significance. This outcome implies that while high-scoring writings may exhibit a slight advantage in lexical diversity over low-scoring ones, this advantage is not statistically distinguishable. Potential reasons for the lack of a significant difference may be related to the learning environment and assessment mechanisms. On one hand, first-year high school students primarily acquire vocabulary from textbooks and classroom instruction. Within the context of practical writing, specific information and vocabulary reserves tend to be homogenized, thereby limiting the differentiation between students of different levels. On the other hand, in examination, scoring standards often prioritize accuracy and convention in expression over lexical diversity. Consequently, students tend to use familiar and safe expressions to mitigate the risk of errors and achieve higher scores, a mindset that may suppress the differentiating effect of lexical diversity, leading to a non-significant difference in mean TTR between high and low-scoring compositions.

Concurrently, the standard deviation in the low-scoring group was slightly higher than in the high-scoring group, indicating that some students in the low-scoring group also possess a degree of lexical diversity but struggle with its appropriate application in sentences. Conversely, the high-scoring group exhibited a greater degree of consistency and stability. This finding suggests that improvements in writing proficiency are reflected not only in an increased quantity of vocabulary but also in the stabilization of vocabulary usage ability.

Based on the data, lexical diversity in this study demonstrated a non-significant difference. This suggests that while lexical diversity contributes to writing quality to some extent, it is not a definitive factor.

### 4.2. Syntactic complexity

In contrast, syntactic complexity proved to be a more significant differentiator between high and low-scoring writings. This study conducted a multi-dimensional examination of syntactic features using three indicators: Mean Length of T-unit (MLT), Dependent Clauses Per Clause (DC/C), and Complex Nominals Per T-unit (CN/T).

#### 4.2.1. Mean Length of T-unit (MLT)

In terms of Mean Length of T-units (MLT), the high-scoring group exhibited a mean of 14.6666, significantly exceeding the low-scoring group's mean of 11.4734. This difference of 3.193145 reached a highly significant level ($p < 0.001$), indicating a marked advantage in sentence length and information capacity for high-scoring students. Longer T-units suggest students' ability to integrate more information within a single sentence, thereby achieving more efficient expression. This also reflects a higher degree of syntactic organization in high-scoring writings. This finding suggests that students in the high-scoring group have developed strong syntactic expansion skills during their learning process, and the output sentence structure is gradually transitioning from simple sentences to complex structures. Furthermore, the proximity of standard deviations between the two groups also underscores the strong group consistency of this difference, implying that the high-scoring group's overall superiority stems from collective performance rather than individual outliers. Consequently, MLT can be considered a critical indicator of writing quality disparities.

### 4.2.2. Dependent Clauses Per Clause (DC/C)

Regarding Dependent Clauses Per Clause (DC/C), the high-scoring group's mean was 0.3093, compared to the low-scoring group's 0.2753. While the high-scoring group was slightly higher, the difference was not statistically significant ($p = 0.248$). This suggests that the frequency of subordinate clause usage has limited efficacy in differentiating writing proficiency levels. This phenomenon might be attributed to the nascent stage of subordinate clause acquisition among first-year high school students, where pronounced differences have not yet emerged. Moreover, an increase in subordinate clauses does not invariably enhance expression quality. Excessive reliance on clauses may even lead to lengthy sentences and disordered structures, thereby affecting the clarity of expression. Consequently, in exam writing tasks, students may prioritize clarity and coherence over the mere increase of subordinate clauses.

From the perspective of data distribution, the minimal difference in this metric between the two groups indicates that, in the first year of high school, students of different levels have not yet formed a significant differentiation in the use of clauses when it comes to practical writing. Therefore, DC/C is not considered a critical indicator for distinguishing between high and low-scoring writings.

### 4.2.3. Complex Nominals Per T-unit (CN/T)

Complex Nominals Per T-unit (CN/T) demonstrated a significant difference in this study. The high-scoring group's mean was 1.4590, substantially higher than the low-scoring group's mean of 1.0391, with the difference achieving a highly significant level ($p < 0.001$). This result indicates that high-scoring writings are more complex in the use of nominal structures. They can enhance information density by using postposition modifications of prepositional phrases and non-finite verb structures within a single sentence, making the expression more compact. Concurrently, the standard deviation for the high-scoring group (0.3269) was also higher than that of the low-scoring group (0.2218). This suggests that when using complex noun phrases, the high-scoring group exhibits greater diversity in phrase construction compared to the low-scoring group, leading to more flexible expression. This finding supports CN/T as a significant indicator for differentiating writing proficiency levels.

## 4.3. Comprehensive analysis

Based on the analysis results from both lexical and syntactic dimensions, it can be seen that the differences between high-scoring and low-scoring writings are primarily rooted in syntax rather than lexicon. Specifically, high-scoring writings are characterized by longer sentences, more complex structures, and more concise expression, whereas low-scoring writings tend to rely on simpler sentence structures and their overall expression is relatively monotonous. Data show that MLT exhibits the most obvious differentiating effect, followed by CN/T, while DC/C has a relatively weak effect. In contrast, although there are certain differences in lexical diversity, no significant discriminatory ability has been formed. This outcome suggests that in the context of English practical writing at the first-year high school level, syntactic complexity, particularly metrics related to information integration and phrase usage ability, is a key factor in differentiating writing proficiency. However, the diversity in vocabulary is not clearly distinguished.

## 5. Conclusion

This research analyzed 20 high-scoring and 20 low-scoring practical writing samples from the final English examination of first-year high school students at Shijiazhuang Foreign Language School. The study employed a comparative analysis of writing characteristics across different proficiency levels from both lexical diversity and syntactic complexity perspectives. Through the statistical analysis of Type-Token Ratio (TTR) and L2SCA syntactic complexity indicators (MLT, DC/C, CN/T), this paper aims to explore the specific differences in the practical English writing by first-year high school students.

The research findings indicate that in terms of lexical diversity, writings from the high-scoring group exhibited slightly higher Type-Token Ratio (TTR), type counts, and token counts compared to the low-scoring group. However, these differences did not reach statistical significance. This suggests that while lexical diversity shows some variation within the context of this study, it is not a stable differentiator of writing proficiency and thus not a decisive factor influencing writing scores. In contrast, the high-scoring group significantly surpassed the low-scoring group in mean length of T-unit (MLT) and complex nominals per T-unit (CN/T). Although dependent clauses per clause (DC/C) were marginally higher in the high-scoring group, the difference was not significant, suggesting that the frequency of clause use has a limited impact in distinguishing writing proficiency levels.

Overall, this study indicates that the differences in the English practical writing proficiency of first-year high school students mainly lie in the syntactic level rather than the lexical level. This finding implies that the enhancement of writing ability is not solely dependent on increasing vocabulary or stacking complex clauses, but rather on the capacity to organize syntactic structures and optimize methods of information expression. Therefore, in the teaching of high school practical writing, teachers should emphasize the accuracy of vocabulary use and foster students' awareness of syntactic structures, enhancing the complexity and coherence of their expressions through sentence pattern instruction.

Although this study has to some extent revealed the differences in language features among students of different levels in writing, there are still some limitations. Firstly, the sample size of the research is relatively limited, with only 40 essays selected, which may affect the representativeness of the research results. Secondly, the language indicators selected in this paper are rather limited and fail to comprehensively cover all dimensions of writing complexity. Future research could expand the sample size and introduce more language indicators to more comprehensively depict the writing characteristics of high school students' practical writings.

In conclusion, this study highlights the critical role of syntactic complexity in differentiating writing proficiency. It offers valuable insights into the developmental characteristics of English writing ability at the high school stage and provides empirical support for writing instruction practices.

## References


[1] Zhong, M. W. (2024). Developmental characteristics of syntactic complexity in senior high school students' English practical writing and teaching strategies. Zhejiang Normal University, Jinhua, China. https://doi.org/10.27464/d.cnki.gzsfu.2024.001095

[2] Liu, L. S. (2022). A comparative study of the syntactic complexity features of noun phrases in college entrance examination compositions of different score levels. Journal of Foreign Language Teaching in Middle Schools, 9, 26–31.

[3] Yang, X. W. (2025). A study on English writing teaching for the transition from junior to senior high school. Campus English, 33, 175–177.
[4] Chen, M. F. (2010). A study on cohesion between junior and senior high school English teaching under the new curriculum. East China Normal University, Shanghai, China.
[5] Engber, C. A. (1995). The relationship of lexical proficiency to the quality of ESL compositions.Journal of Second Language Writing, 4(2), 139-155.
[6] Jarvis, S. (2000). Methodological rigor in the study of transfer: Identifying L1 influence in the interlanguage lexicon. Language Learning, 50(2), 245–309.
[7] Mellor, A. (2011). Essay length, lexical diversity and automatic essay scoring. Memoirs of the Osaka Institute of Technology, 55(2), 1–14.
[8] Wang, J. J. (2019). A comparative study of type-token ratio in Chinese and American college students' compositions on the same topic. Nanjing University of Aeronautics and Astronautics, Nanjing, China.
[9] Bao, G. (2008). A multidimensional study of lexical richness development in L2 learners' writing. Journal of Foreign Language Education and Technology, 5, 38–44.
[10] Wan, L. F. (2010). A study of lexical richness in Chinese English majors' L2 writing. Foreign Language World, 1, 40–46.
[11] Larsen-Freeman, D. (1978). An ESL index of development. TESOL Quarterly, 12, 439–448.
[12] Norris & Ortega (2009). Towards an organic approach to investigating CAF in instructed LA: The case of complexity. Applied Linguistics, 30, 555–578.
[13] Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. International Journal of Corpus Linguistics, 15(4), 474–496.
[14] Peng Jinhua (2022). On the features of syntactic complexity of senior high school students' English practical writings. Suzhou University of Science and Technology.
[15] Lu, X. (2014). Computational methods for corpus annotation and analysis. Springer.