

Lexical Chunk Use in Chinese College EFL Writing: A Corpus-Assisted Developmental Study

Xugui Gui

Zhejiang Yuexiu University, Shaoxing, China

Abstract: This study investigates the characteristics and influencing factors of chunk usage in English writing among Chinese college students, based on a learner corpus compiled from College English Test (CET) writing tasks. By comparing student essays with a reference corpus of model writings from CET Band-4 or Band-6 exams, the research examines chunk usage in terms of word class or parts of speech (POS), frequency, accuracy, and diversity. The analysis reveals that students tend to rely heavily on high-frequency chunks and demonstrate limited flexibility in applying chunks in more complex contexts. Cross-linguistic interference is particularly prominent in the research result. Furthermore, factors such as English language proficiency, negative transfer from their mother tongue (Chinese language), and the instructional environment significantly affect chunk acquisition. The findings offer targeted pedagogical implications and underscore the importance of optimising chunk-based teaching strategies in EFL writing instruction.

Keywords: Corpus-Assisted; College English; Writing; Lexical Chunks; Writing Pedagogy

1. Introduction

Lexical chunks - pre-constructed multi-word units - are ubiquitous in natural language and play a crucial role in enhancing both the fluency and accuracy of written expression. With the rapid pace of globalisation, English proficiency has become a central goal of foreign language education in China. However, despite years of systematic instruction, many students continue to struggle with unclear phrasing and unnatural expression in real-life communication. Common problems in college students' English writing include excessive reliance on basic vocabulary, mis-collocations, and a lack of idiomatic usage - all of which are closely tied to insufficient chunk

acquisition.

For instance, students may overuse simplistic expressions such as "I think" instead of the more natural "It seems to me that," or may produce incorrect collocations like "make" homework rather than the proper "do" it. These issues stem not only from negative transfer from the native language, but also from the insufficient emphasis on chunk input in current pedagogical practices. Thus, analysing college students' chunk use through a corpus-assisted approach is essential for uncovering cross-linguistic patterns and offering empirical evidence for pedagogical improvement.

This study, based on a self-built reflective corpus of College English Test (CET) writings and a reference corpus of corresponding model compositions, investigates Chinese college students' chunk use in English writing. Through examining chunk categories, frequency, accuracy, and diversity, the study aims to identify usage patterns and influencing factors, and to provide data-informed strategies for optimising writing instruction.

2. Literature Review

2.1 Chunk Theory and Second Language Acquisition Research

Chunks, also known as collocations or formulaic sequences, refer to multi-word expressions that frequently co-occur and exhibit varying degrees of fixedness, such as idioms, set phrases, and syntactic frames. Chunk theory challenges the traditional grammar-centric view of language acquisition by emphasising that language learning is not solely governed by rule-based composition but rather facilitated through the accumulation and retrieval of pre-assembled units that enable efficient communication.

In the context of second language acquisition (SLA), chunks play a vital role in the following three major areas.

First, chunks significantly enhance fluency: stored and retrieved as wholes, chunks reduce

the need for on-the-spot construction, thus easing cognitive load and enabling more natural writing flow.

Second, chunks improve accuracy by reducing errors resulting from negative mother tongue transfer or overgeneralised rules.

Third, chunks carry rich pragmatic functions, enabling college students to express themselves appropriately across different communicative contexts.

Nevertheless, acquiring lexical chunks poses significant challenges for college students, especially when they are learning L2. Common issues include overreliance on a limited set of high-frequency chunks, lack of creative usage, and frequent collocational errors due to negative transfer. These problems suggest that mastering chunks requires more than rote memorisation - it demands deeper cognitive processing and contextual adaptation. As such, promoting effective chunk acquisition has become a focal point in SLA research and higher education pedagogy.

2.2 College English Writing and Specially-built Corpora in Writing Research

student's corpora, especially those of College English writing, systematically document authentic language output from college students, provide a robust empirical foundation for chunk research. Through large-scale, quantitative analyses of real writing data, teachers and researchers can identify common difficulties and developmental patterns in Chinese college students' L2 acquisition. Specifically, self-built Corpora such as the Testing College English Writing Corpus (TCEWC), the Controlled College English Writing Corpus (CCEWC) and the Referential College English Sample-Writings Corpus (RCESC) offer a comparative and data-rich platform for investigating chunk usage among Chinese college students.

These corpora support multi-dimensional analyses: frequency analysis of word classes and chunk types reveals usage tendencies; error analysis uncovers learning bottlenecks; semantic analysis highlights context-sensitive usage. The primary advantages of corpus-assisted methods lie in their large-scale, objective, and replicable nature, which enhances the reliability of findings and their pedagogical relevance.

However, corpus-based research is not without limitations. Static corpora cannot fully capture the dynamic progression of language

development; individual learner differences may be obscured by group averages; and corpus findings require theoretical interpretation to yield meaningful pedagogical insights. Therefore, integrating corpus methodologies with cognitive linguistics and SLA theories is essential for advancing research in this field.

2.3 Current Research Landscape: Global and Domestic Perspectives

Internationally, research on lexical chunks has developed into a mature field encompassing both theoretical and empirical approaches. Early descriptive studies of idiomatic expressions have evolved into in-depth explorations of chunk processing, pragmatic function, and acquisition mechanisms. Recent studies have increasingly focused on the dynamic relationship between chunk usage and language proficiency, showing that more advanced students demonstrate greater diversity, complexity, and contextual flexibility in their chunk use.

In contrast, domestic studies in China, though rapidly advancing particularly in corpus-assisted research, still face several key challenges. First, there is a lack of robust theoretical frameworks, especially concerning the cognitive mechanisms underlying chunk acquisition. Second, many studies focus on surface-level description without conducting systematic comparisons with model or reference corpora, limiting generalisability. Third, the predominance of cross-sectional designs makes it difficult to trace the developmental trajectory of chunk competence over time.

Future research should pursue methodological innovation. On the one hand, the integration of big data tools and natural language processing can enhance the efficiency and accuracy of chunk identification and annotation. On the other hand, diverse research designs, such as longitudinal tracking, cross-proficiency comparisons, and experimental interventions, can deepen our understanding of chunk acquisition mechanisms. Importantly, attention must also be paid to the pedagogical application of research findings, including the development of chunk-focused instructional models tailored to the needs of Chinese college students.

3. Research Methods

3.1 Corpus Building

This study adopts a comparative analysis

approach by selecting three representative corpora as the primary source for data collection: (1) a specially-designed corpus consisting of English compositions written by the target college student group, (2) a reference corpus of sample model essays, and (3) a control corpus composed of an equal number of compositions written by senior students who had passed the College English Test Band 6 (CET-6). The selection of these corpora adheres to three criteria: (1) topic relevance, to ensure comparability in writing genre and communicative context; (2) sample representativeness, achieved through stratified random sampling to reflect targeted student's characteristics; and (3) temporal consistency, to ensure all samples were collected within a similar timeframe, thereby minimising diachronic interference.

All corpora were anonymised and standardised into electronic text format and subsequently imported into the LancsBox X, one of the most updated corpus tools, for analysis. This multi-corpus comparative design enhances the representativeness, comparability, and comprehensiveness of the data, thereby providing a robust foundation for investigating the features of chunk usage in learner writing.

3.2 Analytical Tools and Annotation Methods

The study employs a suite of corpus analysis tools for data processing and annotation. LancsBox X is used for chunk extraction and frequency statistics; its advanced and diversified capabilities facilitate the identification of complex chunking patterns. Additionally, the software supports lexical co-occurrence analysis and collocation strength measurement. CLAWS, a part-of-speech (POS) tagging tool, is applied to automatically tag the corpora, laying the groundwork for syntactic-level analysis. And similarly, the annotation of semantic domains is also applicable to the corpus data by using LancsBox.

The annotation process for chunks follows a semi-automated workflow. First, an automatic match is performed using a predefined chunk list. Second, manual validation is conducted by the research team to ensure accurate tagging, especially in cases involving fuzzy boundaries or semantic ambiguity. In instances of disagreement, two experienced English writing instructors are consulted to reach a consensus. The annotation scheme includes four key

dimensions: (1) structural types (e.g., collocations, fixed expressions, sentence frames); (2) functional categories (e.g., discourse patterns, syntactic frames); (3) accuracy tagging (to identify erroneous usage); and (4) creativity tagging (to capture non-standard but communicatively effective usage). To ensure annotation reliability, two rounds of consistency checks and error evaluations were conducted, keeping the inter-annotator disagreement rate below 10%.

3.3 Analysis Dimensions

A multi-dimensional framework was applied to systematically examine the use of lexical chunks:

Frequency Analysis: The standardised frequency (occurrences per 100 words) and log-likelihood values were used to assess the significance of chunk use across different groups.

Accuracy Analysis: A categorised error system was established to include collocational, grammatical, and pragmatic errors. The chunk error rate was calculated to evaluate the target student's ability in using chunk patterns.

Diversity Analysis: Indicators such as the type-token ratio (TTR) and the number of unique chunk types were employed to assess the breadth of sampled student's chunk repertoire and to differentiate between simple and complex usage patterns.

Functional Distribution Analysis: The study compared the use of functional chunks related to expressing opinions, organising arguments, and achieving discourse cohesion across different tested and controlled groups, with a particular focus on their usage in academic versus personal contexts.

This mixed-methods approach, which combines both quantitative statistics and qualitative case analysis, enables a macro-level understanding of usage trends and a micro-level exploration of key usage features. It ensures that the findings are not only systematically grounded but also analytically explanatory.

4. Results and Discussion

4.1 General Characteristics of Chunk Use among College Students

The data reveal that the chunk use among the target college students displays distinct stage-specific characteristics. In terms of chunk

selection, students showed a strong preference for high-frequency and simple chunks, particularly those frequently found in textbooks and classroom instruction. This reliance led to a relatively limited chunk repertoire, as evidenced by a significantly lower type-token ratio compared to the reference corpus (0.32 vs. 0.51). Structurally, the students tended to use shorter and simpler chunks, with an average length of 3.2 words, whereas the chunks in the reference corpus averaged 4.7 words.

Analysis of distribution patterns showed that the sampled students most frequently used chunks to express personal opinions and attitudes (accounting for 38% of all chunk use), while their application in discourse organisation and interactional management was significantly underutilised. Notably, as students' language proficiency improved, their chunk use gradually shifted from a narrow focus on high-frequency forms to a more diversified distribution. However, even at advanced levels, their use of complex and varied chunks remained significantly less than that of the upper-year comparison group.

4.2 Comparative Analysis with Senior Students

A comparative analysis revealed a clear pattern of chunk overuse and underuse among Chinese university / college EFL students. The most pronounced overuse occurred in chunks with clear grammatical structures, with an average frequency 2.3 times that of the upper-year or senior group. In contrast, chunks requiring greater contextual flexibility were used far less frequently: only 40% of the frequency observed in the more proficient group.

More specifically, semantic analysis showed that the sampled students had skewed understandings of certain collocational meanings. For example, when using the verb "cause," students preferred its neutral collocates (e.g., "cause problems"), whereas advanced students more often chose negative collocates (e.g., "cause trouble"). Additionally, students frequently overlooked stylistic appropriateness, leading to frequent use of informal chunks in formal contexts.

4.3 Typical Error Types and the Influence of L1 Transfer

Error analysis identified three main categories of chunk-related errors among Chinese EFL students.

First, collocational errors (52%) were the most frequent, often stemming from improper verb-noun pairings (e.g., "make a discussion" instead of "have a discussion") and incorrect preposition use (e.g., "depend on" instead of "depend on"), most of which can be attributed to direct transfer from Chinese syntax.

Second, grammatical errors (28%) included collocational inconsistency (e.g., "a important problem") and tense errors, reflecting the student's lack of a comprehensive chunk awareness.

Third, pragmatic errors (20%) were commonly caused by inappropriate register use (e.g., informal pronouns like "you guys" in formal contexts) or the misapplication of functional expressions.

Importantly, the study highlights the dynamic nature of L1 transfer. At lower proficiency levels, transfer was predominantly negative. However, as students became more advanced, they began to apply Chinese conceptual frameworks creatively. While these expressions may not fully conform to standard English norms, they effectively conveyed meaning in context, demonstrating the student's emerging flexibility in chunk use. This transition from negative to creative transfer illustrates a dynamic developmental trajectory across linguistic systems.

4.4 The Impact of Proficiency Levels on Chunk Competence

The relationship between language proficiency and discourse chunk use appears to be nonlinear. Beginner students, normally the freshmen and sophomores who are below CET-4 standards, typically demonstrated "low quantity, low quality" usage, with chunk density reaching only 60% of that found in the upper-year group and an error rate as high as 42%. Intermediate students (CET-4 to CET-6) entered a stage characterized by "increased quantity but unstable quality," with chunk frequency approaching that of the comparison group but error rates remaining around 35%. Advanced students (CET-6 and above) achieved the stage of "stable quantity and improved quality," marked by a consistent chunk usage rate and a significantly lower error rate of 15%. These students also began to employ more complex chunks and pragmatically functional expressions.

Different proficiency levels were also associated with different chunk acquisition pathways.

Beginners acquired chunks mainly through explicit classroom instruction; intermediate students began to engage with naturally occurring chunks via extensive reading and listening; advanced students demonstrated strong chunk awareness, actively observing and imitating proficient usage patterns.

These findings underscore several pedagogical implications: (1) at the beginner stage, instruction should focus on accurate input of basic chunks; (2) at the intermediate stage, training should emphasise contextualised chunk use; and (3) at the advanced stage, students should be guided toward chunk awareness and creative deployment. In addition, teaching strategies should be tailored to specific error types, particularly those arising from negative L1 transfer.

5. Teaching Implications

5.1 The Importance of Collocation Instruction in Writing Courses

The findings of this study confirm that systematic instruction in collocations should constitute a core component of English writing courses. Traditional pedagogy that focuses primarily on vocabulary and grammar is insufficient to meet the demands for fluent, accurate, and natural written expression. Based on the data, we propose three fundamental aspects of collocation teaching: First, collocations, as prefabricated language units, help reduce cognitive load during writing and support students in overcoming the challenge of thinking while composing. Second, standardised collocational input can prevent errors caused by negative transfer from the first language, thus enhancing linguistic accuracy. Third, a diverse range of collocations enables students to respond to various communicative situations appropriately, improving pragmatic competence.

In terms of curriculum design, this study recommends integrating chunk-based instruction throughout all phases of writing education. At the beginner level, emphasis should be placed on developing awareness of chunks and building foundational expressive skills through frequent exposure to high-frequency core chunks. At the intermediate level, the focus should shift to expanding and accurately applying these chunks, with particular attention to collocation training. At the advanced level, instruction should aim to

foster creative chunk usage and pragmatic sensitivity. This progressive chunk-based teaching framework can effectively support the comprehensive development of students' writing proficiency.

5.2 Corpus-Based Strategies for Teaching Collocations

Based on the empirical findings, this study proposes a structured strategy for teaching discourse chunks:

First, differentiated input strategies should be implemented to address our college students' overreliance on certain collocational patterns in writing.

Second, error-prevention instruction should be incorporated, including contrastive exercises that highlight common errors (e.g., misuse of patterns like "make + noun").

Third, contextualised practice should be promoted through the use of authentic materials and interactive activities (e.g., the short essay writing practice) to help students use discourse chunks accurately in varied communication scenarios.

Finally, data-driven learning (DDL) should be introduced to guide students in using corpus tools to independently explore patterns of chunk usage.

This strategy framework forms a closed-loop process from input adjustment and error prevention to contextualised practice and autonomous exploration. It emphasises both the guiding role of teachers and the cultivation of students' chunk awareness and autonomy, thereby enhancing the effectiveness and relevance of collocation instruction.

5.3 Recommendations for Textbook Design and Teacher Role Optimisation

Drawing on the study's findings, we advocate systematic reforms in both textbook design and teacher roles. For textbooks, vocabulary lists should be reconstructed to systematically integrate high-frequency authentic phrases, moving away from isolated word lists. In addition, targeted practice tasks, such as collocation cloze tests and phrase rephrasing, should be incorporated.

For teachers, the role should shift from "grammar instructor" to "collocation modeler." This involves demonstrating natural collocation usage in class, deliberately selecting key collocations for input, accurately diagnosing

student's collocational errors, and cultivating their ability to acquire collocations independently. To ensure the effectiveness of these reforms, this study recommends establishing comprehensive teacher training programs covering collocation knowledge, corpus tools, and pedagogical strategies. Classroom-based action research should also be encouraged.

Together, these measures form a dual-drive model supported by textbooks and guided by teachers, enhancing the systematicity and scientific grounding of collocation instruction. This integrated "chunk-instructor-method" approach not only addresses current pedagogical needs but also lays the foundation for sustainable development, offering a practical pathway for improving the writing performance of Chinese college students.

6. Conclusion

Drawing on self-built college English writing corpora, this study investigated the features of collocation usage in English writing by Chinese college students and yielded several key findings. Results reveal a clear developmental trajectory in students' use of collocations—from early-stage reliance on simple, high-frequency chunks to more diversified use at advanced stages. However, issues such as limited collocational repertoire and structurally simple usage persist.

Comparative analyses show that students tend to overuse certain collocations while underutilising others. Error analysis identified three major error types: collocational errors, grammatical errors, and pragmatic errors. Notably, the influence of L1 transfer is complex, exhibiting a shift from negative transfer in the early stages to creative L1-based usage at higher proficiency levels. Although such expressions may deviate from standard norms, they often convey meaning effectively, reflecting students' growing flexibility in chunk use.

Despite its important findings, the study has several limitations. The corpora were collected under specific exam conditions, and the cross-sectional design makes it difficult to trace individual developmental trajectories. Additionally, the study lacks consideration of non-linguistic factors that may influence chunk acquisition.

Future research could focus on four main areas:

(1) adopting updated designs to track the

development of collocational competence over time;

(2) integrating experimental techniques to explore the cognitive mechanisms of chunk processing;

(3) conducting instructional intervention studies to evaluate the effectiveness of various teaching methods; and

(4) exploring the application of emerging technologies in chunk research and instruction.

These avenues of inquiry will contribute to the construction of a more comprehensive theory of chunk acquisition and offer valuable insights for improving the quality of college English writing.

References

- [1] Zhao Shudong. A Study on the Use of Adverbs by College English Students from the Perspective of Corpus Linguistics [J]. *Journal of Science and Education*, 2023, (30):82-84.
- [2] Sun Huangke. A Study on the Comparison of Tense between Chinese English Students and Native Speakers from the Perspective of Corpus [J]. *Popular Literature and Art*, 2024, (15):84-86.
- [3] Zhang Zhengkun and Alia Eini. The Diversified Turn in Research on Chinese Students 'Corpora-Review of Chinese Students 'English Characteristics and Applications Based on Corpora' [J]. *Journal of Shanghai University of Technology (Social Sciences Edition)*, 2023,45(03):258-263.
- [4] Gao Jianwu, Zhang Mei, and Chen Wenting. Re-examining the Development of Reporting Verbs Driven by Intervention Theory: An Interdisciplinary and Hierarchical Study Based on a Corpus of High-Level Native Speakers [J]. *Foreign Language Studies*, 2022,39(05):52-56+82.
- [5] Zhao Huan-gai. A Study on the Construction of a Demand-oriented Corpus for Chinese Inherited Language Students [J]. *International Chinese Education (Chinese-English)*, 2022,7(02):15-22.
- [6] Cao Xianwen. Construction of a Corpus for Chinese Students from the 'Demand Side' Perspective in Second Language Acquisition Research [J]. *Chinese Teaching and Research*, 2020, (01):38-46.
- [7] Sun Zhenjuan. The Acquisition of Analytical Causative Construction [D]. *Hunan University*, 2019.

[8] Dou Wen-na, Chen Pin. Construction and
Application of a Specialized Learner Corpus

[J]. Journal of Philology,
2019,39(01):116-120.