ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0042

A Corpus-Based Study on Collocation of Verb by China EFL Learners

——A Case Study of the Word Improve

Jiaxing Mu *

School of English Studies, Tianjin Foreign Studies University, 300204, Tianjin, China

* Corresponding author: (Email: y240104@stu.tjfsu.edu.cn)

Abstract

Lexical collocations play a crucial role in language acquisition. However, Chinese learners of English frequently produce inappropriate collocations, a phenomenon particularly evident in verb–object structures. Drawing on data from the British Academic Written English Corpus (BAWE) and the Ten-thousand Chinese Learners' English Composition Corpus (TECCL), this study investigates Chinese learners' use of the English verb improve in term of collocation. The results reveal that while Chinese learners are generally familiar with the core meaning of improve ("to make it better than before"), they fail to fully acquire the semantic prosody associated with its collocates beyond the typical combinations they already know. In the TECCL corpus, improve frequently occurs with a variety of abstract nominal objects compared with BAWE corpus, most of which exhibit features of Chinese English due to L1 transfer, limited vocabulary and lack of authentic and real expressions. The study provides pedagogical implications for English teaching by highlighting the importance of incorporating authentic corpora into classroom to prevent students from staying in their linguistic "comfort zone," which may restrict the diversity and naturalness of their expression.

Keywords

Collocation; Improve; Corpus; Verb–Object Structure.

1. Introduction

Corpus refers to a large collection of texts or language data systematically gathered and organized for linguistic research or natural language processing. In recent years, because of the authenticity and reliability of corpus data, corpus-based research has been increasingly applied in linguistics and education. Studies on collocations, in particular, have shown a steady growth.

In second language learning, mastering verbs plays a crucial role. A good command of verb–object structures helps learners construct complete English sentences and improve their overall language ability. However, due to the influence of Chinese verb–object patterns, Chinese learners of English often produce expressions that reflect features of "Chinglish."

This study adopts a data-driven corpus approach and focuses on the verb improve to explore how Chinese learners use verb–object structures. By comparing the collocations of improve in the writings of Chinese learners and native speakers, this research aims to identify the main differences in collocation and semantic patterning. The findings are expected to provide useful insights for English teaching, especially in helping learners use verb–object combinations more naturally and appropriately.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0042

2. Literature Review

In the field of corpus linguistics, natural language data serve as the main object of study, and researchers adopt a data-driven and empirical approach. From a macro perspective, corpus linguistics investigates large amounts of linguistic evidence and examines language use and learning behavior from multiple dimensions, aiming to summarize the underlying patterns of language use. Among these areas, Gui (2010) pointed that word collocation has long been a major focus of contemporary linguistics, especially within corpus-based studies.

2.1. Corpus-Based Studies on Collocations

During the 1950s and 1960s, research on word collocations was still in its early stage. Firth, (1957) known as the father of collocation studies in linguistics, proposed that "you shall know a word by the company it keeps". He emphasized that collocation is not a simple co-occurrence or just a position of words, but rather a relationship determined by the mutual expectancy and mutual attraction among lexical items, as well as the associative links between them.

With the development of corpora, research on collocations gradually moved from introspective approaches to the use of statistical methods to calculate frequency. Jones and Sinclair (1974) were the first to systematically investigate collocations based on corpus data. They established a set of principles and methods to identify word collocation. Sinclair (1995) then introduced the concept of semantic prosody to describe the attitudinal and evaluative meaning associated with collocation. Building on this, Stubbs (1995) categorized semantic prosody into positive, neutral and negative types. In the later decades, more researches were engaged in the field. Laufer and Waldman (2011) found through corpus analysis that second language learners use significantly fewer and less accurate than native speaker. Wang (2020) highlights similarities and differences in the use of I think by HKE and BrE speakers and also examines possible reasons that may lead to these.

Entering the 21st century, the growing availability of large-scale corpora has expanded collocation research into multilingual and cross-disciplinary contexts, including diachronic analyses of collocation change. For instance, Tang (2012), through a study of English translations of Chinese government white papers, identified typical errors and features in verb–object collocations. Niu and Shen (2022) distinguish the difference between "try to" and "try doing" by using the corpus BNC.

2.2. Corpus-Based Studies on Collocations among Chinese EFL Learners

With the continuous improvement of English learner corpora, an increasing number of studies have focused on English learning and teaching from a corpus-based perspective. For example, Zhang and Yang (2009) extracted all verb–noun collocations from the Chinese Learner English Corpus (CLEC) to examine collocation errors in Chinese students' English writing and their underlying causes. Chen and Lin (2010) compared the associative and collocation of the high-frequency adjective good between Chinese learners and native speakers. Zheng and Xiao (2015) employed the Chinese Students' Spoken English Corpus (SSEC) and conducted an error analysis of Chinese learners' spoken collocations, identifying native language transfer, synonym confusion, over generalization, and incorrect word choice as the major sources of collocation errors. Wang(2016), drawing on the CLEC, the Freiburg–LOB Corpus (FLOB), and the British National Corpus (BNC), analyzed the semantic prosody differences in noun collocates of "do-type" verbs between Chinese learners and native speakers. Liu and Wang (2015) examined the usage patterns of adjectives in the English writing of non-English-major postgraduates, aiming to provide pedagogical insights for English writing instruction. Fu and Chen (2022) investigate the acquisition of English delexical verb collocation by Chinese speaking students with intermediate and advanced English proficiency. The research finds that learners used restricted types of delexical verb collocations and seldom used the collocations that were most frequently

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0042

used by native English speakers. Based on the "Language Exposure Hypothesis" as the theoretical framework, Zhang and Zhang (2023) conducted a study using the Corpus of Spoken English by Chinese Beginners. Taking the high-frequency verbs "like", "go", and "make" as examples, they investigated the similarities and differences in collocation errors of these high-frequency verbs among Chinese English beginners across three stages. The purpose of this study was to explore the regular patterns characterizing the development of beginners' collocation competence in spoken English verbs.

In general, both domestic and international studies on lexical collocations can be categorized into theoretical and empirical research. Among them, empirical research in China has produced a relatively rich amount of findings. However, most studies have concentrated on adjective–noun and verb–noun combinations, often based on standardized test corpora. Research at the micro level, focusing on specific verbs, remains few. This study draws on a corpus of Chinese EFL learners' routine written assignments and compares it with authentic academic English corpora from native speakers. By analyzing the data, this study aims to more precisely uncover Chinese learners' habitual collocation and the underlying causes of their collocation errors.

3. Research Questions

This study, on the one hand, investigates the frequency and common collocation of the verb improve in the Chinese EFL learners' writing corpus (TECCL), and on the other hand, conducts a comparative analysis with a native-speaker corpus (BAWE) to explore the differences in the use of this verb between Chinese learners of English and native speakers. Specifically, it aims to address the following three questions:

1) What are the differences in the frequency distribution of the verb improve used by Chinese and British university-level English learners in the TECCL and BAWE corpora?

2) What are the differences in the collocations of the verb improve used by Chinese university-level English learners in the TECCL corpus compared with those in the BAWE corpus?

3) What are the possible causes of the differences in the collocation of the verb improve between the TECCL and BAWE corpora?

4. Research Design

The tool adopted in this study is WordSmith Tools 5.0, and the research method employed is the collocation analysis method. Using WordSmith Tools 5.0, the node word improve was retrieved, with the collocational span set at 0/+5. The frequency and Z-score of each collocate were calculated to identify the typical collocations of improve used by Chinese EFL learners. These were then compared with the typical collocations found in the British Academic Written English Corpus (BAWE), in order to summarize the frequency distributions and collocational differences between the two corpora.

The corpus used for Chinese learners is the Ten-thousand English Compositions of Chinese Learners (TECCL) corpus, established by Xue Xizhe and Xu Jiajin. The TECCL corpus contains approximately 10,000 compositions, totaling 1,817,335 words. The writing tasks included in the corpus cover in-class timed writing, after-class assignments, mid-term and final exam essays, speech drafts prepared for class presentations and group writing projects, all of which are academic tasks within the English curriculum. The corpus contains over a thousand essay topics, among which university-level writing is the majority. The ratio of samples from Project 985/211 universities and non-Project 985/211 institutions closely reflects the actual composition of Chinese higher education.

The British Academic Written English (BAWE) corpus, established around 2000 through collaboration among the University of Oxford, the University of Warwick, and the University of

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0042

Reading, comprises 30,000 academic texts totaling 6,727,486 words. It includes academic assignments written by British undergraduate and postgraduate students across various disciplines, and aims to investigate the academic writing proficiency of both domestic and international students in UK higher education.

To ensure comparability and data validity, texts from primary and secondary school students were removed from the TECCL corpus. The final database retained 6,700 university-level texts, totaling 1,333,600 words, which enabled a balanced comparison with the BAWE corpus.

5. Research Results

5.1. Frequency Analysis of the Verb Improve

Using WordSmith Tools 5.0, the frequency of the verb improve was calculated in both the TECCL and BAWE corpora. It should be noted that, due to the large difference in corpus size, frequency normalization was necessary. Specifically, the observed frequency of the target word was divided by the total number of words in the corpus and then multiplied by one million to obtain its normalized frequency per million words.

In the TECCL corpus, the normalized frequency of the verb improve, obtained through the Concord function of WordSmith 5.0, was 790.34. In contrast, in the BAWE corpus, the normalized frequency of improve was 266.96. The comparison of the normalized frequencies in the two corpora indicates that Chinese learners of English tend to overuse the verb improve.

To further strengthen the reliability of this finding, the log-likelihood value was also calculated. The log-likelihood ratio of improve between the TECCL and BAWE corpora was +690.11, confirming that the difference is statistically significant. Therefore, the study concludes that Chinese English learners show a clear tendency to overuse the verb improve, as shown in Table 1-3.

Table 1. Normalized Frequency of Use by Chinese Learners

Word item	Observed Frequency	Normalized Frequency (per million words)
Improve	1054	790.34

Table 2. Normalized Frequency of Use by British Learners

Word item	Observed Frequency	Normalized Frequency (per million words)
Improve	1796	266.94

Table 3. Log-likelihood value of improve

Item	O1	%1	O2	%2	LL	%DIFF	Bayes	ELL	RRisk	LogRatio	OddsRatio
Word	1054	0.08	1790	0.03 +	690.11	197.04	674.20	0.00001	2.97	1.57	2.97

5.2. Collocation Features of the Verb Improve

In the TECCL corpus, the concord function of WordSmith 5.0 was used to retrieve instances of improve as the node word. Within a right span of five words, verb–object structures were extracted and filtered according to two criteria: a minimum of 15 co-occurrences and a Z-score above 3. Fourteen noun collocates met these criteria, as shown in Table 4, ranked by descending Z-score: ability, quality, level, safety, English, standard, learning, life, skill, food, communication, living, environment and awareness. Among them, ability shows the strongest collocation strength (128 co-occurrences). In these cases, improve generally conveys the meaning of "enhance" or "strengthen," and typically co-occurs with abstract nouns such as ability and level, indicating a preference for expressing improvement in such abstract aspect instead of certain things.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0042

Using the same retrieval parameters, eight high-frequency noun collocates were identified in the BAWE corpus: quality, efficiency, accuracy, skill, situation, health, service and life. The most frequent collocate is quality (559 co-occurrences), where improve means"to make better" or "to refine," often appearing in expressions like improve product quality or improve service quality.

From the features of the word they choose, While Chinese learners in the TECCL corpus show some similarities with native speakers in using improve with words like quality and skill, a closer look at collocation reveals a notable difference in noun types. In TECCL, improve frequently co-occurs with abstract nouns such as ability and level, whereas in BAWE, it more often combines with concrete and measurable nouns like efficiency and accuracy. This suggests a clear semantic contrast between the two corpora, indicating that learner language tends to favor generalized and abstract expressions, while native academic writing emphasizes specificity.

In the TECCL corpus, improve predominantly collocates with abstract nouns such as ability, level, English, learning and skill. These nouns are closely associated with individual capability and personal effort, reflecting a tendency among Chinese learners to use improve to describe internal progress or the enhancement of one's own potential. Such collocations convey a sense of self-directed development and are typically situated in educational or personal achievement contexts.

In contrast, the BAWE corpus reveals a different collocation pattern, where improve is frequently linked with nouns such as quality, efficiency, accuracy, health and service. These items refer to measurable, institutional improvements rather than internal abilities. The usage reflects an impersonal and objective orientation, showing that native writers use improve mainly to indicate functional refinement or optimization of external conditions.

From the perspective of semantic prosody, improve in TECCL carries a clearly positive and subjective tone, emphasizing aspiration and personal advancement. In BAWE, however, its prosody is neutral and pragmatic, highlighting problem-solving and effectiveness. This contrast suggests that Chinese learners may overgeneralize improve to self-related contexts, whereas native writers employ it with greater semantic precision and contextual restraint, as shown in Table 4-5.

Table 4. Z-scores of collocates for 'improve' in TECCL

Node word	Collocate	Z-score
improve	ability	47.18
improve	quality	40.87
improve	level	34.53
improve	safety	27.87
improve	English	14.14
improve	standard	13.56
improve	learning	12.36
improve	life	6.06
improve	skill	5.86
improve	food	5.14
improve	communication	5.07
improve	living	3.74

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0042

Table 5. Z-scores of collocates for 'improve' in BAME

Node word	Collocate	Z-score
improve	quality	52.24
improve	efficiency	20.27
improve	accuracy	14.21
improve	skill	10.35
improve	situation	9.98
improve	health	7.96
improve	service	5.31
improve	life	3.95

6. Discussion

From the perspective of L1 transfer, the distinctive use of improve among Chinese learners can be largely attributed to the influence of the Chinese verb “tǐ gāo.” In Chinese, this verb carries strong semantic versatility and can co-occur with both abstract and concrete objects, such as“tǐ gāo shui ping” (raise the level) and“ tǐ gāo zhi liang” (improve quality). This high degree of transitivity leads learners to treat improve as a universally applicable verb in English, thereby combining it with a wide range of abstract or generalized nouns. As a result, while the overall semantic prosody remains positive, the collocation precision and contextual appropriateness are weakened.

Similarly, the uses like improve the level or improve the standard are frequent. However, the combination “improve level” rarely appears in native-speaker corpora. This difference mainly results from native speakers’ pragmatic choices and learners’ conceptual transfer from their first language. In English, level is a label-like or container noun, referring to an abstract notion of degree or standard rather than a concrete entity. Native speakers tend to talk about the contents of the container instead of the container itself. Therefore, instead of saying improve your English level, they naturally say improve your English or use more precise expressions such as improve your fluency, enhance your proficiency, or expand your vocabulary. Moreover, the verb improve inherently contains the idea of “raising the level,” so adding level becomes redundant and violates the linguistic principle of economy.

In addition, native speakers possess a richer and more nuanced vocabulary, enabling them to select more specific and context-appropriate collocates: enhance skills, boost performance, raise standards or deepen understanding, instead of the vague improve level. By contrast, Chinese learners often produce this collocation due to conceptual transfer, since in Chinese expressions like “improve the English level” are extremely common. When unsure about idiomatic alternatives such as enhance or advance, learners rely on the safe combination improve level, which, though grammatically correct, sounds unnatural. In short, native speakers favor concise, semantically precise expressions while Chinese learners tend to overuse of improve level in learner corpora. According to the "Language Exposure Hypothesis", learners can only successfully acquire the rules of English verb collocations when their exposure to such collocations reaches a "critical threshold". In practice, the overuse of improve English by learners precisely indicates the singularity of their collocation exposure. Since their exposure to other types of collocations like “improve accuracy” fails to meet the "critical threshold", learners cannot effectively acquire these collocation rules, ultimately leading to the repeated use of only one familiar collocation in language production.

To sum up, Chinese learners’ problematic collocations with "improve" mainly derive from two key factors: L1 conceptual transfer and inadequate lexical resources plus limited authentic

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0042

language exposure. Influenced by the versatile Chinese equivalent "tǐ gāo," learners overgeneralize "improve" in English, producing non-native collocations like "improve the level" that is redundant in native expression and lack collocation precision. In contrast, native speakers use richer, context-specific vocabulary for accuracy. Due to limited lexical range and insufficient language input, learners rely on "improve" as a "safe" choice to avoid errors, further compromising collocation appropriateness. These issues collectively reflect the impact of L1 transfer and inadequate English linguistic development on learners' use of "improve."

7. Conclusion

A comparative analysis of the verb "improve" in the TECCL and BAWE corpora shows Chinese EFL learners significantly overuse "improve" compared to native English speakers. In terms of collocation, Chinese learners prefer abstract words when using "improve," whereas natives favor concrete terms. Additionally, Chinese learners rely on limited fixed collocations, lacking the diversity of native speakers. Two main factors contribute to these differences. first, L1 transfer allows flexible verb-noun collocations, leading learners to adopt a "linguistic safety" strategy; second, insufficient exposure to authentic and authoritative English input, which restricts their collocation possibilities and expression diversity.

Based on the problems identified in Chinese learners' use of improve, English vocabulary teaching should incorporate authentic corpora, enabling students to explore improve in real English contexts. Learners can gradually internalize frequent patterns and summarize the features of improve through repeated exposure to typical collocations. In classroom teaching, improve often appears with a single fixed meaning in textbooks. Teachers should therefore extend beyond the textbook, guiding students to observe how native speakers use improve with different meanings across contexts and emphasizing its semantic variation in authentic discourse. Teachers also need to closely monitor students' writing and speaking, since learners are most vulnerable to L1 influence when first encountering new words. Timely correction of redundant expressions helps learners recognize the source of their errors and acquire more natural usage.

In short, correct collocations cannot be taught merely through explaining word meanings. Effective vocabulary teaching must integrate lexical, semantic, and grammatical dimensions in a systematic way. Language learning is never achieved overnight, it requires continuous reflection and innovation. By combining vocabulary instruction with other supporting resources, teachers can foster learners' linguistic awareness, help them develop self-correction and monitoring skills, and ultimately enable them to achieve continuous improvement in their language proficiency.

References

[1] Chen, J., & Lin, T. (2010). An Analysis on the Colligation and Collocation of the High -frequency Word Good in Learner Corpus [J]. Journal of Tianjin Foreign Studies University.

[2] Firth, JR. Papers in Linguistics [M]. London:Oxford University Press, 1957.

[3] Fu, S., & Chen, Y. (2022) A Corpus-based Study of the Acquisition of English Collocations with Delexical Verbs by Chinese-speaking Students [J]. Language Education, 10(01):52-65.

[4] Gui, S., Feng, Z., Yang, H., et al. (2010). Corpus linguistics and foreign language teaching in China [J]. Modern Foreign Languages.

[5] Laufer, B., & Waldman, T. (2011). Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. Language Learning, 61(2), 647–672.

ISSN: 2637-6067 DOI: 10.6918/IJOSSER.202512_8(12).0042

[6] Liu, B., & Wang, Y. (2015). A Study on Use of Adjectives in English Writing by Non-English Major Postgraduates [J]. Chinese Foreign Languages, 12(4), 45–53.

[7] Niu, B., & Shen, S. (2022). Collostructional analysis of try to V and try V-ing constructions in English [J]. Foreign Language Teaching and Research, 54(5), 656–667, 798.

[8] Sinclair, J, Jones S. English Lexical Collocations: A Study in Computational Linguistics [J]. English Studies, 1974.

[9] Sinclair, J. Corpus, Concordance, Collocation [M]. Oxford University Press, 1991.

[10] Stubbs, M. Collocations and Semantic Profiles: On the Cause of the Trouble with Quantitative Studies [J]. Functions of Language, 1995.

[11] Tang, Y. (2012). Verb–object collocations and translation strategies in English versions of Chinese white papers: A corpus-based study [J]. Shanghai Journal of Translators.

[12] Wang, Q. (2020). A corpus-based contrastive analysis of I think in spoken Hong Kong English: Research from the International Corpus of English (ICE) [J]. Australian Journal of Linguistics, 40(3), 319–345.

[13] Wang, R. (2016). A Comparative Study on the Semantic Prosody Difference of CLE and NES: Using Verbs with "DO" Meaning [J]. Foreign Language Research.

[14] Zhang, H., & Zhang, S. (2023). A Study on the Developmental Characteristics of High-Frequency Verb Collocations in Spoken English of Chinese Beginner EFL Learners [J]. Foreign Language Research, 40(06): 52-59.

[15] Zhang, W., & Yang, S. (2009). An Analysis of V-N Collocation Errors in CLEC [J]. Foreign Languages Bimonthly, 32(2), 39–44.

[16] Zheng, L., & Xiao, Z. (2015). A Corpus-based Study of Collocational Use in Oral Production by Chinese EFL Learners [J]. Foreign Language Learning Theory and Practice.