International Journal of English Linguistics; Vol. 14, No. 4; 2024
ISSN 1923-869X E-ISSN 1923-8703
Published by Canadian Center of Science and Education

A Comparative Study on the Lexical Collocations in Academic Discourse by International Scholars and Chinese EFL Learners

Nan Wang[1]

[1] School of Chinese Language and Literature, Beijing Foreign Studies University, Beijing, China

Correspondence: School of Chinese Language and Literature, Beijing Foreign Studies University, Beijing, China.

Received: May 3, 2024 Accepted: June 28, 2024 Online Published: July 16, 2024

doi:10.5539/ijel.v14n4p90 URL: https://doi.org/10.5539/ijel.v14n4p90

Abstract

Lexical collocations are essential in academic English and present considerable challenges for L2 learners. This study compiled three text corpora: one from academic written texts published in top international linguistics journals, and the others from MA theses and doctoral dissertations of Chinese college students majoring in Foreign Linguistics and Applied Linguistics. An Academic Collocation List (ACL) was initially created from the journal articles and then the collocation lists of the top 200 verbs were compared across the three corpora. The study identified potential collocates for over 2,400 content words in the ICAE (International Corpus of Academic English) journal articles, resulting in 375 potential combinations. Among these, noun combinations (adj+n, n+n) accounted for nearly 70% of the total entries, followed by verb+noun/adj combinations, adv+adj combinations, and verb+adv combinations. Statistical analysis revealed significant differences between Chinese EFL learners and international scholars, with a larger discrepancy observed between MA theses and journal articles than between doctoral dissertations and journal articles. Chinese EFL learners tended to overuse a limited set of collocations that, while grammatically correct, often sounded unnatural to native speakers. Additionally, the verb+noun combinations used by Chinese EFL learners were frequently physically and semantically different from those employed by international scholars. These findings underscore the need for targeted pedagogical interventions to improve the collocational competence of Chinese EFL learners.

Keywords: lexical collocation, Academic Collocation List (ACL), international scholars and Chinese EFL learners

1. Introduction

During the 1980s, scholars such as Pawley and Syder (1983) and Peters (1983) spearheaded the recognition of the importance of dissecting phraseological units and word combinations, shedding light on their prevalence in both written and spoken discourse. These preconstructed linguistic units, often termed “prefabricated units” (or “prefabs”) are deemed fundamental to the linguistic proficiency of native speakers (Sinclair, 1991). Simultaneously, the exploration of multi-word constructions, encompassing idioms, collocations, and chunks, has underscored their practical significance in language utilization and acquisition (Sinclair, 1987; Schmitt, 2004; Meunier & Granger, 2008; Simpson-Vlach & Ellis, 2010; Wright, 2019).

Despite the burgeoning interest in multi-word constructions, the utilization of corpus analysis to profile lists of such constructions did not become widespread until Durrant (2009) directed attention towards two-word collocations, sparking increased interest and research in the subsequent years. Collocation, as a substantial branch of multi-word constructions, has been delineated and categorized from various scholarly perspectives (Sinclair, 1991; Cowie, 1998; Hoey, 2005). Investigations into collocations among non-native learners have indicated that their grasp of conventional collocations may be comparatively weaker and somewhat misguided compared to that of native speakers (Granger, 1998; Cowie, 1998). Furthermore, research has consistently underscored the challenges second language learners face in mastering collocations, given their inherent grammatical and lexical unpredictability (Nation, 2001).

While proficiency in the use of collocations is crucial for academic writing, learners often grapple with restricted collocations, which constitute the bulk of phraseological material and pose formidable learning hurdles (Howarth, 1996; Cowie, 1998). The differentiation between restricted collocations and grammatical collocations, as suggested by Benson (1985), accentuates the intricacies of collocational knowledge and its significance in language acquisition.

This study aims to compare the utilization of lexical collocations by Chinese EFL learners with the usage patterns observed among international scholars. In this context, international scholars refer to experienced linguistic researchers who have published articles in top linguistic journals, while Chinese EFL learners include both postgraduate and doctoral students who are novice academic writers and publishers. Identifying commonalities and disparities between these cohorts will provide valuable insights for refining academic writing skills and informing pedagogical strategies.

2. Literature Review

2.1 Definitions and Classification of Collocations

Collocations, a term frequently employed in linguistic research, have been subject to diverse definitions reflecting various theoretical perspectives. Two primary viewpoints, the "frequency-based approach and the "phraseological approach", have emerged as prominent frameworks in understanding collocational phenomena (Sinclair, 1991; Cowie, 1998).

The frequency-based approach, also known as the statistically oriented approach, emphasizes the statistical tendency of words to co-occur within a certain proximity. This perspective, influenced by Firth's assertion that "you shall know a word by the company it keeps", underscores the intrinsic relationship between words' lexical meanings within a sentence (Firth, 1957). Scholars such as Halliday and Sinclair have further developed this approach, with Sinclair (1991) defining collocations as "the occurrence of two or more words within a short space of each other in a text." Sinclair also distinguishes between significant collocations and casual collocations, highlighting the former's interpretive value through examples such as the various collocates of the word *achieve* within a collection of texts.

Conversely, the phraseological approach, termed the significance-oriented approach, focuses on word combinations wherein one element deviates from its usual meaning or where restrictions exist on permissible combinations. Cowie, a key proponent of this perspective, categorizes collocations as a type of word combination, delineating between "composites" and "formulae". While formulae serve pragmatic functions, composites primarily serve syntactic roles. Cowie further distinguishes between free combinations, restricted combinations, figurative idioms, and pure idioms, all of which exhibit varying degrees of arbitrary restrictions (Cowie, 1994).

A trend favoring the integration of both frequency and lexical meaning in the study of collocations has emerged, aligning with research efforts from the late 1990s (Herbst, 1996). In this study, collocations are approached through the combination of frequency and lexical semantics.

Regarding the classifications of collocations, this section primarily addresses grammatical collocations and lexical collocations. Grammatical collocations involve the combination of a dominant word with a preposition or grammatical structure, categorized into sub-categories such as noun+preposition and adjective+to-infinitive (Benson & Ilson, 1986).

On the other hand, lexical collocations consist of two equal lexical components/words, classified into types such as adjective+noun and noun+noun (Hausmann, 1989; Aisenstadt, 1981). However, distinctions are made between restricted or lexical collocations and other combinations, considering factors such as semantic transparency and substitutability. In this study, collocations are analyzed based on their frequency and span, utilizing measures of association such as MI and t-score. Additionally, attention is given to grammatical patterns, with functional words added manually as needed.

2.2 Previous Research on Lexical Collocations

Collocation, as an integral component of phraseology, has garnered significant scholarly attention globally. While investigations into the phraseological obstacles encountered by learners are not uncommon, recent years have seen a surge in studies facilitated by the availability of native and learner corpora, along with the advancement of techniques for extracting phraseological patterns. Notably, corpus-based approaches have gained prominence in collocation extraction, displacing traditional methods. Given the emphasis of this study on lexical collocations, this section endeavors to provide a comprehensive overview of research on lexical collocations within the domains of academic writing, second language acquisition.

Research on lexical collocations has flourished, particularly in written registers such as academic writing. Investigations by Cowie (1991, 1992) into journalistic writing underscored the prevalence of restricted collocations, prompting subsequent studies. An analysis by Howarth (1996) of verb+noun collocations unveiled individual variations among non-native and native students, shedding light on phraseological tendencies. Similarly, studies by Nesselhauf (2003, 2004) and Laufer and Waldman (2011) on verb+noun combinations revealed learners' difficulties with collocations lacking word-for-word correspondence, highlighting the influence of learners' first

language (L1) on their second language (L2) usage.

In contrast to phraseological definitions, certain studies have adopted frequency-based perspectives. A comparative analysis by Siyanova and Schmitt (2008) of adjective+noun collocations unveiled a mixture of frequent, infrequent, and unattested collocations in non-native speakers' writing, suggesting comparable usage patterns with native speakers. These findings underscore the complexity of lexical collocation acquisition among non-native speakers.

Furthermore, the study of collocation usage is increasingly prominent in second language acquisition research (e.g., Boers & Webb, 2018; Granger, 2021; Siyanova-Chanturia & Spina, 2019), providing valuable insights into the development of collocational competence among EFL learners. Durrant and Schmitt (2009) analyzed learners' tendencies to overuse high-frequency collocations while underutilizing strongly associated collocations. Ander and Yıldırım (2010) categorized lexical errors among Turkish EFL learners, shedding light on common error types in written compositions. Granger and Bestgen (2014) examined differences in collocation usage between intermediate and advanced learners, further elucidating learners' struggles with collocations. These studies underscore the need for targeted pedagogical interventions to enhance learners' collocational proficiency. The importance of academic collocations in language learning and teaching has been widely recognized, with numerous studies highlighting their role in enhancing language proficiency. Academic vocabulary lists, such as the Academic Word List (AWL) by Coxhead (2000) and the Academic Vocabulary List (AVL) by Gardner and Davies (2013), have been developed to aid language learning. The academic collocation list by Ackermann and Chen (2013) is another useful resource, as it specifically addresses the frequent and pedagogically relevant collocations across various academic disciplines, providing a targeted approach to collocation learning. Additionally, studies have investigated the efficacy of collocation teaching in improving learners' language proficiency. Hsu and Chiu (2008) demonstrated the positive impact of collocation instruction on speaking comprehension and writing. Similarly, Rahimi and Momeni (2012) conducted a quasi-experimental study that revealed enhanced language proficiency among learners following collocation instruction. Such findings underscore the pedagogical significance of incorporating collocation teaching into language curricula.

The present study focuses primarily on lexical collocations, aiming to generate a comprehensive collocation list and investigate similarities and differences between Chinese EFL learners and international scholars. These insights have implications for second language learning and teaching practices.

3. Methodology

3.1 Research Questions

When comparing the use of lexical collocations between international scholars and Chinese MA and PhD EFL learners, we employed international scholars as a reference. First, we constructed the Academic Collocations List (ACL) from international journal articles. Due to space limitations, we then selected verb collocations for comparison among the three groups. Based on this, the study addresses the following two research questions.

This research aims to address the following two questions:

(1) As the reference corpora, what is the composition of Academic Collocations List in international journal articles?

(2) Are there any commonalities or differences in the use of lexical collocations between Chinese EFL learners and international scholars? If so, what are the specific commonalities and differences among MA theses, doctoral dissertations, and international journal articles?

3.2 Corpus of the Research

There are altogether three self-built corpora used in this study. The International Corpus of Academic English (ICAE) serving as a reference corpus, comprises over 15 million words of academic written texts published in three top international linguistics journals: English for Specific Purposes, Journal of Memory and Language and Language Learning. Learner corpora consist of the Corpus of MA Theses (CMT) and the Corpus of Doctoral Dissertations (CDD) from Chinese college students majoring in Foreign Linguistics and Applied Linguistics.

These three corpora share similar size. The source and capacity of the three corpora is provided in Table 1. In the process of compilation, all the diagrams, tables, footnotes, references, appendices were excluded to ensure the accuracy of data processing.

Table 1. Source and capacity of the three corpora

Corpus	Source	Number of texts	Tokens
International Corpus of Academic English (ICAE)	Published articles from *English for Specific Purposes, Journal of Memory and Language and Language Learning* during 2019–2022	176	1,558,608
Corpus of Doctoral Dissertations (CDD)	Doctoral dissertations from Chinese college students majoring in Foreign Linguistics and Applied Linguistics during 2019–2021	27	1,586,211
Corpus of MA Theses (CMT)	MA theses from Chinese college students majoring in Foreign Linguistics and Applied Linguistics during 2019–2021	104	1,553,639

3.3 Research Procedures

The development of the ACL involved two stages: computational analysis and human intervention. The computational stage extracted node words, computed mutual information (MI) scores and t-scores, and formed an initial collocation list. The human intervention stage then refined this list based on qualitative principles. Subsequently, a comparative analysis of academic collocations used by Chinese EFL learners and international scholars was conducted.

3.3.1 Computational Analysis

1) Content Word Extraction: Using AntConc 3.2.1w, content words from the ICAE corpus occurring at least five times per million words and in at least five different texts were extracted.

2) Filtering General Service Words: Words from General Service List (GSL) (West's 1953) were removed, leaving over 2400 content words.

3) Generating Initial Collocation List: Potential collocations were extracted using BFSU Collocator 1.0 and AntConc, ignoring function words, proper nouns, numbers, and non-words. The initial list included node words, collocates, their positions, raw frequency, text count, MI scores, t-scores, and examples, resulting in 797 entries.

3.3.2 Human Intervention

1) Initial Refinement: The list was refined to exclude:

① Incomplete phrases or parts of extended phrases

② Collocations closely related to linguistics

③ Frequency and time-related collocations

④ Fixed expressions with little replaceability

⑤ collocations with transparent adjectives

⑥ Hyphenated collocations

2) Systematic Classification: Part-of-speech tagging was used to extract target combinations (verb+noun, adjective/noun+noun, adverb+verb, adverb+adjective). Three linguistics experts independently classified these combinations to ensure consistency.

3) Further Refinement: Collocations were refined based on rules like adding articles and prepositions, retaining British English forms, and converting words to their base forms.

Table 2. A sample output from the original lexical collocation list

Pre-collocate	Academic word	Post-collocate	Position	No of texts	Raw freq	MI	t-score
well	established		-1	11	14	4.71	3.60
made	comment		-3	5	5	4. 18	2. 11
pose	challenge		-3	5	5	8.96	2.01
further	analyses achieve	goal	-1	17	25	3.39	4.52
			3	6	6	5.56	2.95
research	conducted differed	significantly	-3	18	26	3.05	4.48
			1	16	24	5.67	4.59

3.3.3 Comparative Analysis

1) Extracting High-Frequency Verbs: The top 200 verbs and their inflections from three corpora (CDD & CMT) were listed and selected.

2) Extracting Collocations: Using the top 200 verbs as node words, collocations were extracted as previously described.

3) Significance Calculation: The Log-likelihood ratio was used to calculate significant differences in overall frequency and shared combinations among the three corpora.

4) Detailed Analysis: Eight verbs were selected for analysis of verb+noun combinations regarding repetition rate and semantic preference.

4. Results and Discussion

4.1 Composition Analysis of Academic Collocations in ICAE

After multiple stages of computational analysis and manual proofreading, as detailed above, the ACL comprising 375 entries in the field of linguistics was completed. The number and proportion of various parts-of-speech (POS) collocations are presented in Table 3.

Table 3. The composition of final academic collocations

Combinations	Lexical Categories		No of Entries	Percentage
Noun combinations	adj	n	201	53.6%
	n	n	50	13.3%
	Sub-total		251	66.9%
Verb+noun/adj combinations	v	n	50	13.3%
	v	adj	5	1.4%
	Sub-total		55	14.7%
Verb+adv combinations	v	adv	7	1.9%
	adv	v	11	2.9%
	adv	vpp	13	3.5%
	Sub-total		31	8.3%
Adv+adj combinations	adv	adj	38	10.1%
	Sub-total		38	10.1%
Total			375	100%

Generally, noun combinations account for nearly 70% of the total entries (n=251). The second largest category is verb+noun or verb+adjective combinations (n=55), followed by adverb+adjective combinations (n=38). Verb+adverb combinations are comparatively fewer but share almost the same proportion as adverb+adjective combinations, including three subcategories: verb+adverb (n=7), adverb+verb (n=11), and adverb+verb past participle (adv+vpp) entries (n=13).

Compared to the development of an ACL from the written curricular component of the Pearson International Corpus of Academic English by Ackermann and Chen (2013), which covered 28 academic disciplines and identified 2,468 of the most frequent and pedagogically relevant entries, the results of the present study from the ICAE are largely consistent. They found that noun combinations constituted almost three-quarters of their total lexical collocations (74.3%), followed by verb combinations (13.8%), adverb+verb combinations (6.9%), and adverb+adjective combinations (5.0%). The results from this study align with their findings, except for the last two categories. These minor differences suggest that my research in the field of linguistics corroborates their study to a significant extent.

Table 4. A representative selection of adjective + noun combinations

	Adjective	Noun	No of texts	Raw freq	MI	t-score
1	academic	registers	5	5	5.56	2. 19
2	academic	writing	30	136	7.06	11.58
3	alternative	explanation	13	17	5.86	4.05
4	(a) considerable	amount (of)	8	10	6.17	3. 12
5	critical	factor	5	6	5.78	2.41
6	crucial	role	14	16	5.49	3.91
7	cultural	background	6	7	5.77	2.60
8	deep	understanding (of)	5	7	9.36	2.64
9	detailed	analysis	9	12	4.10	3.52
10	educational	background	5	7	5.87	2.78

Table 5. A representative selection of noun + noun combinations

	Noun	Noun	No of texts	Raw freq	MI	t-score
1	background	information	16	24	6.65	4.85
2	background	knowledge	18	28	6.95	5.25
3	data	set	11	54	5.29	7. 16
4	discussion	section	11	23	7.74	4.77
5	judgment	task	12	30	3.77	5.08
6	pilot	study	9	32	5.50	5.53
7	research	article	12	25	5.45	4.89
8	research	design	15	17	4.81	3.98
9	source	information	5	22	5.40	4.58

Tables 4 and 5 provide a representative selection of adjective+noun and noun+noun combinations in the ACL. It is evident that nominal collocations predominantly feature in the final ACL, potentially reflecting the characteristic academic style of the corpus. However, despite this prevalence, Nation (2001) have shown that it is the acquisition and utilization of verb collocations that pose the greatest challenge for L2 learners. Similarly, research by Laufer and Waldman (2011) highlights the considerable difficulties faced by L2 learners in mastering verb collocations, particularly in academic contexts. Therefore, the subsequent section will primarily delve into the intricacies of verb collocations across the three corpora.

4.2 Frequency and Distribution of Lexical Collocations in the Three Lists

The analyses of frequency employed the lexical collocations of the top 200 verbs from three corpora, along with the original collocation list derived from the reference corpus. These lists, ordered by frequency, are presented in Appendix A.

The overall frequency of the top 200 verbs showed an increasing trend across the three corpora: 24,509 instances in MA theses, 20,720 in doctoral dissertations, and 28,498 in journal articles. The computational process yielded 215 entries in MA theses, 120 in doctoral dissertations, and 133 in journal articles. Table 6 summarizes the frequency information for lexical collocations in each corpus. After human intervention, 98 entries remained in the final list for MA theses, which is more than twice the number in both doctoral dissertations and journal articles. This indicates that MA theses utilized the most lexical collocations among the three corpora, followed by journal articles, with doctoral dissertations using the least. The log-likelihood ratios for lexical collocations highlight significant differences between the three corpora. Doctoral dissertations ($LL=26.37$, $P<0.01$) and journal articles ($LL=19.38$, $P<0.01$) differ significantly from MA theses. More lexical collocations were used in MA theses and journal articles than in doctoral dissertations. However, the difference between doctoral dissertations and journal articles is minor, as indicated by the log-likelihood ratio ($LL=0.53$). This suggests that the proficiency in the use of lexical collocations by doctoral students approaches that of native speakers.

Table 6. Frequency information in the three lists

	Total freq of verbs	Total freq of original collocations	Total freq of final collocations
CMT	24,509	215	98
CDD	20,720	120	40
ICAE	28,498	133	46
LLR between CMT & CDD	401.27***	29.32***	26.37***
LLR between CMT & ICAE	287.87***	19.77***	19.38***
LLR between CDD & ICAE	1374.67***	0.92	0.53

Note. * P<0.5; ** P<0. 1; *** P<0.01; LLR=log-likelihood ratio.

After close scrutiny, some interesting similarities and differences emerged among the lexical collocations across the three corpora. Only five common collocations appeared simultaneously in all three lists: *(be) closely related (to)*, *mainly focus/es/ed (on)*, *analyz(s)e/d data*, *achieve (a) goal(s)*, and *convey meaning(s)*. Unlike the compilation of the ACL in journal articles, this analysis retained all different forms of these collocations, explaining the varied forms. Table 7 shows the frequency of occurrence of the five shared collocations in the three corpora and the log-likelihood ratios among them. The overall frequency of possible forms of a collocation was presented, despite being part of a considerable number of collocations in the three corpora.

Table 7. Frequency of 5 shared collocations in the three lists

Lexical collocations		CMT	CDD	ICAE	LLR between CMT & CDD	LLR between CMT & CDD	LLR between CMT & CDD
(be) closely	related (to)	106	68	31	9.17**	43.64***	13.53***
mainly	focus (on)	49	7	8	36.31***	32.91***	-0.09
analyze	data	14	6	18	3.46	-0.49	-6.49*
achieve	(a) goal	24	7	6	10.21**	11.62***	0.06
convey	meaning	6	7	6	-0.06	0.00	0.06

Note. * P<0.5; ** P<0.1; *** P<0.01; LLR=log-likelihood ratio.

Take the combination *(be) closely related (to)* as an example. Among these shared lexical collocations, there were striking differences in the usage of (be) closely related (to) between MA theses and doctoral dissertations (P<0.01), with this collocation appearing much more frequently in the learner corpora. This collocation indicates a close relationship between two or more objects. Consulting the ACL revealed that international scholars used a variety of collocations to express similar meanings, such as *(be) highly correlated*, *rely heavily (on)*, *directly/highly/particularly relevant*, *(be) strongly associated (with)*, and *primarily concerned (with)*. This suggests that international scholars are more adept at using a broader range of collocations instead of repeating the same ones. In contrast, learner corpora included collocations like *(be) significantly correlated (with)*, *(be) closely linked (to)*, rely more (on), and *(be) significantly related (to)* to partially replace *(be) closely related (to)*. This pattern may reflect an over-reliance on a limited set of collocations by Chinese EFL learners, leading to language that may sound unnatural or less varied compared to that of international scholars. This finding aligns with Cobb (2003), which observed that learners tend to overuse a few specific collocations even when used correctly. Similarly, Laufer and Waldman (2011) found that learners employed fewer different collocations than native speakers.

Further analysis revealed that besides the five common items, MA theses and doctoral dissertations shared another 15 lexical collocations, such as *achieve purpose*, *investigate(d) effects*, *summarize(s) findings*, and *exert (an) influence (on)*, all italicized in Appendix A. Among these items, only two collocations in MA theses occurred slightly less frequently than in doctoral dissertations: *implied meaning* (n=20 < n=21) and *(be) closely linked (to)* (n=5 < n=7). This suggests significant similarities in the use of lexical collocations by postgraduates and doctoral students. Additionally, MA theses shared four more common lexical collocations with journal articles, while doctoral dissertations shared two. These six items are both italicized and bold-faced in Appendix A. Up to now, there were 20 common collocations in learner corpora, while 34 lexical collocations in journal articles did not appear in learner corpora.

The statistical analyses reveal several key points regarding the use of lexical collocations: First, the similarities between the two learner corpora (MA theses and doctoral dissertations) are greater than those between the learner corpora and the reference corpus. This is evident in the shared items and the selection of a small number of collocations. Second, there are significant differences between postgraduates and doctoral students in terms of the

number and use of lexical collocations. In addition, the discrepancy between MA theses and journal articles is larger than the discrepancy between doctoral dissertations and journal articles. Furthermore, Chinese EFL learners tend to over-rely on a small number of collocations. Consequently, even if the use of these collocations is correct, non-native speakers may still sound unnatural to native speakers of the target language.

4.3 Investigation of Deviant Lexical Collocations in Learner Corpora

In addition to the general commonalities and differences identified above, 115 diverse lexical collocations were found across the three lists. Among these, 81 (65 in CMT and 16 in CDD) lexical collocations in the learner corpora will be investigated here. Additionally, the 15 commonly used collocations in MA theses and doctoral dissertations, which did not appear in the collocation list of the top 200 verbs in journal articles, will also be examined.

For a comprehensive analysis, we referred to the ACL generated from the journal articles and the concordance lines. The initial filtration revealed that nine items (e.g., *(be) widely acknowledged*) in MA theses appeared in the ACL, while five items (e.g.,*adopt (an) approach*) appeared in doctoral dissertations. Thus, the analyses between Chinese EFL learners and international scholars were conducted from two aspects as below.

One aspect focused on the 15 shared lexical collocations occurring only in the two lists of learner corpora, where verb+noun collocations occurred comparatively frequently, amounting to nine. Since the collocations a*chieve purpose* and *achieve effect* share the same verb, the other eight verbs were selected for the analysis of verb+noun lexical collocations between Chinese EFL learners and international scholars.

Table 8. A comparative analysis of verb+noun lexical collocations

Verbs	Nouns in ICAE	Nouns in CMT	Nouns in CDD
achieve	goal, *goals,purposes, level(s)*	goal(s), purpose(s), effect(s) *needs, understanding, level, result, target, progress, fluency*	Goal, purpose, effect(s) *understanding, comprehension*
investigate	role, extent, issue, *effect(s), relationship, possibility, question(s), differences, characteristics, ways, contribution, impact*	effects,*features, effect, relationship, attitudes, role, effectiveness, behaviors, differences,frequency, nature,factors*	role, effects *relationship(s), effect, roles, differences, features, process, relation*
summarize	*points, research, results*	findings, differences	findings
exert		influence, *impact*	influence, *impact*
perform	*task(s), actions, functions, work, analysis*	functions, *function, task, action*	functions, *function, action(s), task(s), analysis*
acquire	*language, skills*	knowledge, *language, information, meanings, competence*	knowledge, *information, meaning, skills, language*
illustrate	*point, categories*	point	point, *distinction, process*
emphasize	*importance*	importance, *claim, meaning*	importance

Note. Italic=collocates occur more than three times but did not belong to the targeted collocates of this study.

From Table 8, if the singular and plural forms of a noun are considered separate collocates, the total number of collocates in ICAE was 33. Fourteen collocates in CMT occurred in ICAE, representing 42.4%, while 15 in CDD occurred in ICAE, representing 45.5%. This suggests that doctoral students might be more proficient in the use of lexical collocations than postgraduates. It also highlights a considerable amount of similar collocates in the learner corpora. Generally, three categories can be classified concerning repetition and semantic preference:

(1) Complete Overlap with Similar Semantic Meanings:

Collocates used by Chinese EFL learners also occurred in ICAE, and their semantic meanings were similar. For instance, collocates serving as objects of the verb *perform* in ICAE were task(s), actions, functions, work, and analysis—except for *work*, all others did not appear in learner corpora.

(2) Partial Overlap with Similar Semantic Meanings:

Part of the collocates used by Chinese EFL learners occurred in ICAE, with similar semantic meanings. For example, collocates of *achieve* in ICAE were goals, purposes, and levels; four of these were used in MA theses,

while two were used in doctoral dissertations. Although Chinese learners used diverse collocates like fluency, progress, needs, and comprehension, these were less frequently used by international scholars.

(3) Limited Overlap with Different Semantic Meanings:

Some collocates used by Chinese EFL learners did not occur in ICAE, and their meanings differed significantly. For example, *acquire meanings* seemed inappropriate as *acquire* emphasizes something gained through effort, ability, or behavior, lacking correlation between these two words. Similarly, collocates of *exert* like influence and impact were less popular among international scholars who preferred expressions like *affect* combined with modifiers (e.g., *(be) negatively affected*).

The second aspect examined the 67 unique lexical collocations (56 in CMT and 11 in CDD). The results confirmed previous findings. Some collocations seemed like casual blends by Chinese EFL learners, such as *perceive world*, *creating meaning*, *processing effort*, *possess meanings*, *serve functions*. Certain collocations used frequently by Chinese EFL learners did not occur in the ACL of journal articles, e.g., *arouse attention/interest*, *attach importance (to)*, *ensure validity/reliability*. Chinese EFL learners tended to overuse a small number of lexical collocations, often changing the modifier in ways that were correct but less common in journal articles, e.g., *(be) highly/mainly/significantly related (to)*, *related studies/terms/concepts/theories, main/specific/multiple/primary function(s)*.

In summary, The analysis of verb+noun lexical collocations reveals distinct differences between Chinese EFL learners and international scholars. Collocates unique to Chinese EFL learners often share similar semantic meanings with those used by international scholars, yet there are also notable semantic divergences in some cases. Additionally, many collocates frequently employed by Chinese EFL learners are rarely or never used by international scholars, indicating significant variations in collocational usage patterns between these groups.

One significant factor is the influence of L1 knowledge on L2 learning and application, which can be both positive and negative. For instance, expressions like *mainly/negatively/highly related (to)* in doctoral dissertations are grammatically correct but rarely found in journal articles, likely due to direct translations from their mother tongue. EFL learners also create unique blends of acceptable words. As Pawley and Syder (1983) noted, native speakers can focus on other tasks since memorized sequences require little encoding, whereas EFL learners struggle to use collocations naturally due to limited knowledge and negative L1 influence. Another factor influencing L2 learning is the prototype effect, which can both facilitate and hinder the acquisition process. Salient examples or exemplars within a category aid individuals in gradually understanding other members. This concept applies to second language acquisition, where the ACL from the reference corpus acts as typical members of lexical collocations. These typical examples help EFL learners grasp various collocations in different contexts, explaining why shared collocations across the three corpora are more easily understood. In verb+noun collocations, some collocates used by Chinese L2 learners were semantically similar to those used by international scholars, indicating a positive impact of understanding central members in a category. However, misunderstandings of these salient examples can lead to divergent outcomes, as evidenced by unique collocates used by L2 learners to convey diverse meanings. Thus, a thorough comprehension of core lexical collocations can significantly enhance L2 learning and application, while misinterpretations can result in less natural and accurate language use.

5. Conclusion

The present study aims to compare the utilization of lexical collocations by Chinese EFL learners with the usage patterns observed among international scholars. Using three self-built corpora, this research first creates an Academic Collocation List (ACL) from journal articles and then compares the collocation lists of the top 200 verbs across the three corpora. The study identified over 2,400 content words in the ICAE journal articles, resulting in 375 potential combinations, with noun combinations (adj+n, n+n) accounting for nearly 70% of the total entries. This finding underscores the prevalence of nominal collocations in scholarly writing, reflecting the characteristic academic style of the corpus.

The analysis reveals both similarities and differences in the use of lexical collocations by international scholars and Chinese EFL learners. The number of verbal lexical collocations used in MA theses was more than twice that in doctoral dissertations and journal articles. While both doctoral dissertations and journal articles differed significantly from MA theses, the difference between doctoral dissertations and journal articles was minimal. Regarding the five shared lexical collocations in the three corpora, it was observed that Chinese EFL learners over-rely on a small number of collocations, such as (be) closely related (to). Additionally, significant differences in the use of lexical collocations were noted between postgraduates and doctoral students, with MA theses and doctoral dissertations sharing another 15 lexical collocations, indicating substantial similarities in usage patterns between these two groups. The results also suggested that doctoral students are more proficient in using lexical

collocations than master's students.

As for verb+noun combinations, the differences between Chinese EFL learners and international scholars were evident. While some collocates used by Chinese EFL learners had similar semantic meanings to those used by international scholars, others differed significantly. Additionally, many collocates used by Chinese EFL learners were rarely or never used by international scholars.

Despite these findings, there are some limitations to the present study. The extraction of the ACL was based on journal articles by international scholars in applied linguistics, without considering disciplinary variation. Additionally, while different part-of-speech combinations were listed in the ACL, the analyses between Chinese EFL learners focused only on verb lexical collocations. Future studies could extend the analyses to include all other categories of collocations, providing a more comprehensive understanding of lexical collocation usage across different disciplines and parts of speech.

Acknowledgments

Not applicable.

Authors' contributions

Not applicable.

Funding

Not applicable.

Competing interests

Not applicable.

Informed consent

Obtained.

Ethics approval

The Publication Ethics Committee of the Canadian Center of Science and Education.

The journal's policies adhere to the Core Practices established by the Committee on Publication Ethics (COPE).

Provenance and peer review

Not commissioned; externally double-blind peer reviewed.

Data availability statement

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

Data sharing statement

No additional data are available.

References

Ackermann, K., & Chen, Y.-H. (2013). Developing the Academic Collocation List (ACL) – A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*. https://doi.org/10.1016/j.jeap.2013.08.002

Aisenstadt, E. (1981). *Restricted Collocations in English Lexicology and Lexicography*. ITL Review of Applied Linguistics. https://doi.org/10.1075/itl.53.04ais

Ander, S., & Yıldırım, Ö. (2010). Lexical Errors in Elementary Learners' Compositions. *Procedia-Social and Behavioral Sciences*. https://doi.org/10.1016/j.sbspro.2010.03.864

Benson, M. (1985). Collocations and Idioms. In R. Ilson (Ed.), *Dictionaries, Lexicography and Language Learning* (ELT Documents 120). Oxford: Pergamon Press.

Benson, M., & Ilson, R. (1986). *The BBI Combinatory Dictionary of English: A Guide to Word Combinations*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.bbi1(1st)

Boers, F., & Webb, S. (2018). *Teaching and learning collocation in adult second and foreign language learning*. Language Teaching. https://doi.org/10.1017/S0261444817000301

Cobb, T. (2003). *Analyzing Late Interlanguage with Learner Corpora: Quebec Replications of Three European Studies*. Canadian Modern Language Review. https://doi.org/10.3138/cmlr.59.3.393

Cowie, A. P. (1991). Multiword Units in Newspaper Language. In S. Granger (Ed.), P*erspectives on the English Lexicon: A Tribute to Jacques van Roey* (Louvain-la-Neuve: Cahiers de l'Institut de Linguistique de Louvain). https://doi.org/10.2143/CILL.17.1.2016699

Cowie, A. P. (1992). Multiword Lexical Units and Communicative Language Teaching. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and Applied Linguistics*. London: Palgrave Macmillan. https://doi.org/10.1007/978-1-349-12396-4_1

Cowie, A. P. (1994). Phraseology. In R. E. Asher & J. Simpson (Eds.), *The Encyclopedia of Language and Linguistics*. Oxford: Pergamon Press.

Cowie, A. P. (1998). *Phraseology: Theory, Analysis, and Applications*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198294252.001.0001

Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*. https://doi.org/10.2307/3587951

Durrant, P. (2009). Investigating the Viability of a Collocation List for Students of English for Academic purposes. *English for Specific Purposes*. https://doi.org/10.1016/j.esp.2009.02.002

Durrant, P., & Schmitt, N. (2009). To What Extent do Native and Non-native writers make use of collocations? *International Journal of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral.2009.007

Firth, J. R. (1957). A Synopsis of Linguistic Theory, 1930–1955. In *Studies in Linguistic Analysis*. Oxford: Basis Blackwell.

Granger, S. (1998). Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. In A. Cowie (Ed.), *Phraseology: Theory, Analysis and Applications*. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198294252.003.007

Granger, S. (2021). Phraseology, corpora and L2 research. In S. Granger (Ed.), *Perspectives on the second language phrasicon: The view from learner corpora*. Multiligual Matters. https://doi.org/10.21832/9781788924863

Granger, S., & Bestgen, Y. (2014). The Use of Collocations by Intermediate Vs. Advanced Non-native writers: A Bigram-based Study. *International Review of Applied Linguistics in Language Teaching*. https://doi.org/10.1515/iral-2014-0011

Hausmann, F. J. (1989). Le dictionnaire de collocations. In J. H. Franz, E. W. Herbert & Z. Ladislav (Eds.), *Wörterbücher, Dictionnaries*. Ein internationale Handbuch zur Lexikographie. Berlin: de Gruyter. https://doi.org/10.1515/9783110095852.1

Herbst, T. (1996). *What are Collocations: Sandy Beaches or False Teeth*? English Studies. https://doi.org/10.1080/00138389608599038

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. London: Routledge.

Howarth, P. (1996). *Phraseology in English Academic Writing: Some Implications for Language Learning and Dictionary Making*. Tübingen: Max Niemeyer. https://doi.org/10.1515/9783110937923

Hsu, J., & Chiu, C. (2008). Lexical Collocations and Their Relation to Speaking Proficiency of College EFL Learners in Taiwan. *Asian EFL Journal*.

Laufer, B., & Waldman, T. (2011). Verb-noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning*. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Meunier, F., & Granger, S. (2008). *Phraseology in Foreign Language Learning and Teaching*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.138

Nation, P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*. https://doi.org/10.1093/applin/24.2.223

Nesselhauf, N. (2004). *Collocations in a Learner Corpus*. Amsterdam: John Benjamins Publishing Company. https://doi.org/10.1075/scl.14

Pawley, A., & Syder, F. H. (1983). Two Puzzles for Linguistic Theory: Nativelike Selection and Nativelike Fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and Communication*. New York: Longman.

Peters, A. (1983). *The Units of Language Acquisition*. Cambridge: Cambridge University Press.

Rahimi, M., & Momeni, G. (2012). The Effect of Teaching Collocations on English Language Proficiency. *Procedia-Social and Behavioral Sciences*. https://doi.org/10.1016/j.sbspro.2011.12.013

Rogers et al. (2021). The Creation and Application of a Large-scale Corpus-based Academic Multi-word Unit List. *English for Specific Purposes*. https://doi.org/10.1016/j.esp.2021.01.001

Schmitt, N. (2004). *Formulaic Expressions an Idioms in English: A Corpus-based approach*. Amsterdam: John Benjamins.

Simpson-Vlach, R., & Ellis, N, C. (2010). *An Academic Formulas List: New Methods in Phraseology Research*. Applied Linguistics. https://doi.org/10.1093/applin/amp058

Sinclair, J. (1987). Collocation: A Progress Report. In R. Steele & T. Thomas (Eds.), *Language Topics: Essays in Honor of Michael Halliday*. Amsterdam: John Benjamins. https://doi.org/10.1075/z.lt2.68sin

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Siyanova-Chanturia, A., & Schmitt, N. (2008). L2 Learner Production and Processing of Collocation: A Multi-study Perspective. *Canadian Modern Language Review*. https://doi.org/10.3138/cmlr.64.3.429

Siyanova-Chanturia, A., & Spina, S. (2019). Multi-word expressions in second language writing: A large-scale longitudinal learner corpus study. *Language Learning*. https://doi.org/10.1111/lang.12383

West, M. (1953). *A General Service List of English Words*. London: Longman.

Wright, H. (2019). Lexical bundles in stand-alone literature reviews: Sections, frequencies, and functions. *English for Specific Purposes*. https://doi.org/10.1016/j.esp.2018.09.001

Appendix A

A Representation of Lexical Collocations of the Top 200 Verbs in Three Corpora

	MA theses			Doctoral dissertations			International journal articles		
	Lexical	collocations	Freq	Lexical	collocations	Freq	Lexical	collocations	Freq
1	**(be) closely**	**related (to)**	106	**(be) closely**	**related (to)**	68	***(be) randomly***	***assigned***	45
2	*naturally*	*occurring*	35	processing	effort	31	**(be) closely**	**related (to)**	31
3	**mainly**	**focuses (on)**	25	*implied*	*meaning*	21	***differed***	***significantly***	24
4	*research*	*focus*	22	***(be) randomly***	***assigned***	19	facilitate	learning	21
5	focus	Attention (on)	19	*naturally*	*occurring*	19	significantly	predicted	17
6	attach	importance (to)	19	learning	process	15	(be) well	established	14
7	significantly	affect	17	rely	more (on)	14	**enhance**	**learning**	**13**
8	***enhance***	***learning***	17	*(be) directly*	*related (to)*	12	**analyzed**	**data**	12
9	*summarize*	*findings*	17	(be) significantly	related (to)	11	further	examine	10
10	convey	information	16	*exert*	*(an) influence (on)*	10	focused	primarily (on)	9
11	**convey**	**meaning**	16	*investigated*	*effects*	9	analyzed	separately	9
12	**achieve**	**(a) goal**	15	*perform*	*function*	9	affect	performance	9
13	*(be) randomly*	*selected*	15	involves	use	8	**mainly**	**focused (on)**	8
14	*illustrate*	*point*	15	underlying	factors	8	interpret	results	8
15	*achieve*	*purpose*	14	*achieve*	*purpose*	7	promote	learning	8
16	**analyze**	**data**	14	*achieve*	*effects*	7	focus	xclusively (on)	7
17	**mainly**	**focus (on)**	14	**achieve**	**(a) goal**	7	***investigate***	***role***	7
18	*investigate*	*effects*	14	**convey**	**meaning**	7	affect	ability	7
19	conduct	analysis	14	*research*	*focus*	7	participating	students	7
20	*implied*	*meaning*	14	**mainly**	**focus (on)**	7	facilitates	processing	7
21	*summarizes*	*findings*	14	*(be) closely*	*linked (to)*	7	(be) further	discussed	7
22	*exert*	*(an) influence (on)*	14	*emphasizes*	*importance*	7			
23	minimize	cost	13	adopt	(an) approach	7			
24	*perform*	*functions*	13	adopting	approach	7			
25	traditional	approach	13	(be) mainly	related (to)	6			
26	related	terms	12	***investigated***	***role***	6			
27	*(be) directly*	*related (to)*	12	*achieve*	*effect*	6			
28	*achieve*	*effect*	12	*illustrate*	*point*	6			
29	conduct	study	12	discuss	issues	6			

30	maximize	benefit	12	**achieve**	**(a) goal**	6
31	conducted	(a) research	11	**analysed**	**data**	6
32	conduct	research	11	***communicate***	***effectively***	6
33	minimize	dispraise	11	**convey**	**meaning**	6
34	minimize	praise	11	necessarily	imply	6
35	central	focus	10	examine	impact	6
36	**mainly**	**focused (on)**	10	investigate	extent	5
37	(be) highly	motivated	10	investigate	issue	5
38	specific	functions	10	(be) negatively	affected	5
39	**achieve**	**goals**	9	rely	heavily (on)	5
40	*achieve*	*purposes*	9	actively	participate (in)	5
41	mainly	occurs	9	facilitates	acquisition	5
42	*acquired*	*knowledge*	9			5
43	ensure	validity	9			
44	(be) widely	acknowledged	9			
45	differ	greatly	8			
46	*emphasize*	*importance*	8			
47	reach	(an) agreement	8			
48	infer	meaning	8			
49	conducted	(an) experiment	7			
50	(be) mainly	conducted	7			
51	*achieve*	*effects*	7			
52	***differ***	***significantly***	7			
53	establish	relationship	7			
54	(be) generally	assumed	7			
55	*emphasizes*	*importance*	7			
56	maximize	cost	7			
57	serve	functions	7			
58	(be) positively	correlated (with)	7			
59	(be) actively	involved (in)	6			
60	clearly	indicates	6			
61	mainly	involves	6			
62	shift	focus	6			
63	**convey**	**meanings**	6			
64	(be) carefully	designed	6			
65	creating	meaning	6			
66	minimize	disagreement	6			
67	(be) intrinsically	motivated	6			
68	obtain	information	6			
69	maximize	dispraise	6			
70	maximize	praise	6			
71	*implied*	*meanings*	6			
72	arouse	interest	6			
73	(be) generally	acknowledged	6			
74	processing	resources	6			
75	(be) significantly	correlated (with)	6			
76	(be) easily	identified	5			
77	analyze	results	5			
78	further	investigate	5			
79	(be) further	classified (into)	5			
80	***communicate***	***effectively***	5			
81	reach	level	5			
82	maximize	sympathy	5			
83	possess	meanings	5			
84	arouse	attention	5			
85	perceive	world	5			
86	*(be) closely*	*linked (to)*	5			
87	**analyze**	**data**	6			

88	*summarizes*	*findings*	6
89	adopts	approach	6
90	(be) negatively	related (to)	5
91	(be) highly	related (to)	5
92	*(be) randomly*	*selected*	5
93	*acquire*	*knowledge*	5
94	ensure	reliability	5
95	*summarize*	*findings*	5

Note. Bold indicates lexical collocations that occur in all three corpora. Italic signifies lexical collocations found in both MA theses and doctoral dissertations. Bold & Italic highlights lexical collocations present in both MA theses and journal articles and in both doctoral dissertations and journal articles. Grey indicates lexical collocations that appear in ACL.