LEARN Journal: Language Education and Acquisition Research Network
ISSN: 2630-0672 (Print) | ISSN: 2672-9431 (Online)
Volume: 18, No: 1, January – June 2025

Language Institute, Thammasat University
https://so04.tci-thaijo.org/index.php/LEARN/index

Putting in the Last Piece: A Comprehensive Profiling of Learners' Collocational Competence

Zhiliang Yue[a], Sugunya Ruangjaroon[b,*]

[a] zhiliang.yue@g.swu.ac.th, Faculty of Humanities, Srinakharinwirot University, Thailand
[b] sugunya@g.swu.ac.th, Faculty of Humanities, Srinakharinwirot University, Thailand
[*] Corresponding author, sugunya@g.swu.ac.th

APA Citation:
Yue, Z., & Ruangjaroon, S. (2025). Putting in the last piece: A comprehensive profiling of learners' collocational competence. *LEARN Journal: Language Education and Acquisition Research Network, 18*(1), 54-78. https://doi.org/10.70730/XCDR8218

Received
23/04/2024

Received in revised form
05/07/2024

Accepted
17/07/2024

ABSTRACT

This study aims to provide a comprehensive profile of collocational competence, a key component of one's overall linguistic competence. For the maximum of ecological validity, we elicited naturalistic oral/written production data from 84 Chinese intermediate EFL learners and performed a 2×2 fashion of analysis on their performance in each of the six aspects, namely, collocation accuracy rate, collocation associative strength, collocation density, collocation diversity, and two relevant lexical levels. The findings not only show learners' various inadequacies compared to native speakers, but also reveal the substantial discrepancies between their implicit and explicit collocational knowledge. Our result largely bears out Wray's Dual Model, and some pedagogical implications are suggested accordingly, including a learning mode shift from bottom-up to top-down to remedy the situation.

Keywords: collocational competence, naturalistic data, implicit linguistic knowledge, explicit linguistic knowledge, corpus

Introduction

"You shall know a word by the company it keeps" (Firth, 1969). This remarkable quote brings into light the inextricable relationship between a word and its textual surroundings. Indeed, every word has some specific "company" that it habitually co-occurs with, which cannot be fully explained from a mere grammatical perspective (Sinclair, 1991). The umbrella term for this enticing linguistic entity is formulaic sequence (Wray, 2002) which features a prefabricated nature (Ellis, 1994; Jiang & Nekrasova, 2007; Kim & Kim, 2012; Underwood et al., 2004). Accordingly, it covers anything that is readily available as a single unit rather than subject to compositional analysis and construction, including lexical bundles (Biber, 2009), idioms (Cowie, 2013), collocations (Nesselhauf, 2003), etc.

Of all its subordinate terms, collocation seems particularly magical and mysterious. On the one hand, the mastery of collocations is the recipe for success as it enables learners to sound more idiomatic with fluency (Siyanova-Chanturia, 2015) and with precision (McIntosh, 2009). Thus, it is of crucial importance for learners to further their interlanguage at this junctural level, that is, between word and clause. On the other hand, collocations are notoriously difficult for non-native speakers to acquire, and even the advanced learners often fail to develop a feel about which words naturally go together (Ferraro et al., 2014). As a result, "their language sounds stilted and awkward" (Sinclair, 1991, p. 79). Thus, it is no surprise that this linguistic aspect has been drawing increasing academic attention. As Harmer (2015, p. 28) stated, "word combinations (also known as collocations) have become the subject of intense interest in the recent past."

This study joins the trend and takes collocation as the focus, aiming to profile learners' collocational competence and diagnose where their inadequacies lie. This study contributes to the line of inquiry in three ways, (a) expanding the scope of profiling to a total of six aspects of collocational competence, (b) taking a minimal-control approach to examine both oral and written naturalistic production data which shed light on both learners' implicit and explicit collocational knowledge, and (c) focusing on the group of intermediate-level learners.

Literature Review

Definition of Collocation and Collocational Competence

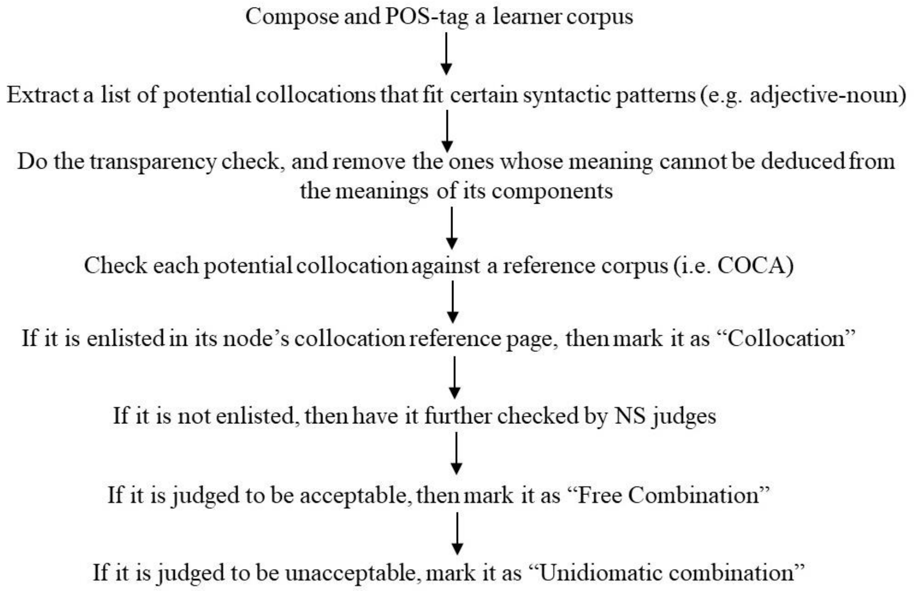
The very first step of studying anything is to define it properly. Many corpus linguists, with their expertise on large-scale data, tend to adopt a frequency-based definition (Biber, 2009; Gablasova et al., 2017; Sinclair, 1991). Here collocation is often understood as a statistically significant co-occurrence of words within a predetermined distance of each other (Walker, 2011). This frequency-based approach is prized for being reasonably objective, statistically reliable, and easily replicable. But it also suffers from its tendency to highlight some frequent yet irrelevant combinations. Thus, even some ardent proponents of this approach acknowledged the necessity of "a preliminary step … where the linguist chooses 'interesting' target words" (Biber, 2009, p. 276).

Another approach is phraseological in nature (Cowie, 2013; Howarth, 1998). Instead of dealing with statistical measurements, it puts emphasis on syntactic structure and semantic properties, both of which are rooted in one's intuitive judgement. Scholars who belong to this camp (e.g., Chen, 2019; Cowie, 2013; Nesselhauf, 2003) proposed syntactic patterns of their interest, such as Verb-Noun (e.g. grab lunch), Adjective-Noun (e.g., grand hotel), Adverb-Adjective (e.g., blissfully ignorant), Adverb-Verb (e.g., abruptly end), etc. The phraseological approach has its face validity, as it is in accordance with the traditional way language is comprehended. As a result, the collocation list extracted this way is likely to sound reasonable and relevant. However, it is only as good as the judge's intuition which, in turn, is based on her personal experience with the language, thus prone to appear idiosyncratic and unrepresentative.

As shown above, the frequency-based approach and the phraseological approach are like two ends of the same stick. One emphasizes evidence while the other values intuition, and the optimal solution appears to be somewhere in between. As Stubbs (2002, p. 217) suggested, "[t]he ideal would be to combine the best of both approaches, so as to make more precise quantitative generalizations about collocations across the whole of the vocabulary of a language." Accordingly, this study adopts a mixed approach to define, identify, and evaluate collocations. Any word pair must meet the following criteria to proceed to the subsequent analysis: 1) fits predetermined syntactic patterns (i.e. Adj-N, Adv-Adj, Adv-V); 2) is fairly transparent in meaning; 3) passes the threshold values of frequency and associative strength in a large corpus or the judgement of native speakers. The workflow is as below:

Figure 1

The Identification Workflow



Compose and POS-tag a learner corpus

Extract a list of potential collocations that fit certain syntactic patterns (e.g. adjective-noun)

Do the transparency check, and remove the ones whose meaning cannot be deduced from the meanings of its components

Check each potential collocation against a reference corpus (i.e. COCA)

If it is enlisted in its node's collocation reference page, then mark it as "Collocation"

If it is not enlisted, then have it further checked by NS judges

If it is judged to be acceptable, then mark it as "Free Combination"

If it is judged to be unacceptable, mark it as "Unidiomatic combination"

Furthermore, we shall operationalize the notion of "collocational competence," since it is the very subject we aim to examine. In the existing collocation studies to date, most do not explicitly mention this term (e.g., Biber, 2009; Chen, 2019; Vilkaite & Schmitt, 2019; Walker, 2011). And even for the few exceptions (e.g., Alangari, 2019; Ferraro et al., 2014; Peng, 2016) which do, none bothers to explain what it means, despite the fact that their studies are centered on this concept. This is no trivial matter, since only when collocational competence is clearly defined can we reliably measure it and effectively develop it.

Ellis (1994) presented two contrasting views of the term *competence*. One is represented by Chomsky (2014) which considers competence as the mental representations of linguistic rules which constitute one's internal grammar. In this case, it is altogether abstract and implicit. Others (Ellis, 1990; Hymes, 1972; Tarone, 1990; Taylor, 1988; Widdowson, 1983) saw competence as one's ability to use the knowledge in specific contexts. In this way, competence becomes more closely intertwined with one's actual performance, with an expanded scope of including explicit knowledge as well. As our ultimate goal is to help learners develop their actual ability to use collocation, we shall side with the second group and propose our working definition of *collocational competence* as follows: collocational competence is one's actual ability to use his or her collocational knowledge to establish effective communication in specific contexts.

Learner Level

Learner's proficiency level is an important factor in understanding the process of second language acquisition (Ellis, 1994). According to the Common European Framework of Reference for Languages (CEFR), learners' proficiency could be divided into six bands (A1/A2/B1/B2/C1/C2) which correspond to three levels: basic, intermediate, advanced (Council of Europe, 2020). When it comes to collocational competence, most scholars seem to particularly favor the advanced group (e.g., Alangari, 2019; Chang et al., 2008; Nesselhauf, 2003, among others). For example, Alangari (2019) examined the academic writings of advanced Saudi learners of English and found that they could use Adj-N type of collocation more accurately than V-N type. Nesselhauf (2003) picked the essays written by 3rd-year and 4th-year college students (which he explicitly equaled as advanced learners) for exploration of the error types and potential reasons.

While the advanced group warrants much investigation, any conclusion drawn based on this group may not apply to learners at lower levels (Öksüz et al., 2021). In fact, those less proficient ones are no less important, because any necessary remedial intervention would be best to be discovered and put in place earlier than later in one's development path. So far, only a handful of scholars have targeted at the intermediate-level learners (Chen, 2017; Khonamri & Roostaee, 2014; Saeedakhtar et al., 2020), and some of them are questionable regarding their criteria of proficiency. For example, Chen (2017, p. 231) described her participants as students who "had studied English for about seven years and could be defined as intermediate learners." It is not too hard to challenge a direct correspondence between years of study and the level of proficiency, as many teachers and students can testify with their own firsthand experience. Any conclusion based on such questionable assumptions could only be considered as tentative, at best. Therefore, the authors consider the intermediate learners still quite under-researched as far as their collocational competence goes and will devote our due attention to this group.

Collocation Types and Aspects

From the phraseological point of view, collocation can be categorized into a number of types, depending on the syntactic structure it fits in. The following are some common types with examples:

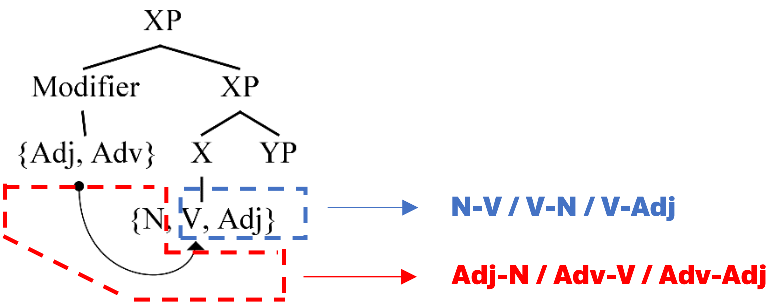
- Verb-Noun: *grab lunch, attend meeting, do homework*
- Noun-Verb: *election approach, task involve, train arrive*
- Verb-Adjective: *grow dark, stay safe, turn red*
- Adjective-Noun: *grand hotel, brilliant idea, desperate need*
- Adverb-Adjective: *blissfully ignorant, stunningly beautiful, clearly visible*
- Adverb-Verb: *fully understand, carefully choose, abruptly end*

Of all these above, the Verb-Noun type seems to be the most researched in the field of SLA. For example, Nesselhauf (2003) analyzed the use of Verb-Noun collocations made by advanced German-speaking EFL learners in writing and identified nine error types with the respective frequency of each type. Peng (2016) explored the use of Verb-Noun collocations by Chinese heritage learners, Chinese foreign learners and Chinese native speakers and identified several patterns for the first two groups, including their underuse of such collocations. Laufer and Waldman (2011) investigated the use of Verb-Noun collocations by learners whose L1 is Hebrew and discovered that learners produced far fewer collocations than their native counterparts and errors persisted even for the advanced level learners. Significantly fewer studies touched on the other types. Durrant and Schmitt (2009) investigated Adjective-Noun and Noun-Noun types in academic writing and found that learners tended to compose collocations common in composition and weak in associative strength. Eguchi and Kyle (2023) examined three types of collocation use, namely, Verb-Noun, Adjective-Noun and Adverb-Verb, of beginner-intermediate level learners, and concluded that associative strength of collocations produced rises as learners' proficiency progresses. However, to the best of our knowledge, a number of types remain unexplored, such as the Adverb-Adjective type. Furthermore, scholars seem to treat each type equally and offered no discussion with regard to their underlying similarity or difference. Is it a view that is too simplistic?

If we examine the matter from the perspective of a syntax tree (*Figure 2*), it becomes clear that these types of collocation occur on various levels, indicating that their corresponding mental representations could operate differently. Furthermore, all these types could be categorized into two kinds: one is "*head-complement*" (including Noun-Verb, Verb-Noun and Verb-Adj) and the other is "*modifier-head*" (including Adj-Noun, Adv-Adj and Adv-Verb). For the former, both the head and the complement are obligatory, while for the latter, only the head is obligatory but not the modifier. That is to say, while the omission of a complement would result in incorrect expression (e.g., *My mother prepared the meal. -> *My mother prepared.*), one could readily omit a modifier without compromising grammaticality of an utterance (e.g., *I live in a grand hotel. -> I live in a hotel.*). This difference in obligatoriness is important when it comes to interpretation of the measurement of certain aspects of collocational performance, such as collocation density. Some studies which focused on a "head-complement" type drew conclusions of overuse/underuse based on density measurements (e.g., Peng, 2016). We deem such interpretation to be too simplistic, because learners did not have full freedom in making as many V-N collocations as they wish but were restricted by the number of clauses as well as the transitivity of the verbs. In other words, the density here could well reflect other things rather than overuse or underuse.

Figure 2

A Syntax Tree Showing the Two Kinds of Collocation



🟦 the "head-complement" types of collocation 🟥 the "modifier-head" types of collocation

To avoid such confounding factors, in this study we shall focus on the "*modifier-head*" group which includes three types, namely, Adjective-Noun, Adverb-Adjective, and Adverb-Verb. On the one hand, they share the same structural relationship, making themselves comparable to each other. On the other, since a user could freely supply or omit the modifier without any grammatical constraint, we argue that any phenomenon of overuse or underuse in this case should be genuine and could be interpreted with confidence.

Aside from the types of collocation, the aspect of the collocational competence is another crucial matter. Besides the aforementioned collocation density, most existing collocational competence studies focused on associative strength (e.g., Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Eguchi & Kyle, 2023) and/or accuracy rate (e.g., Nesselhauf, 2003; Peng, 2016; Vu & Peters, 2023) of the collocations produced by learners. While these two aspects are informative, they do not suffice to construct the whole picture, and other aspects await our investigation. Table 1 below is a summary of all the six aspects we aim to investigate in the current study.

Table 1

Summary of Aspects of Collocational Competence

Aspect Term	Definition	Significance	Measurement	Previous Studies	Previous Findings

Accuracy Rate	the percentage of collocations which conform to the norm	indicate the quality of one's collocational knowledge and the source of errors	number of accurate collocation/total number of collocations * 100%	Nesselhauf (2003), Peng (2016), Vu & Peters (2023), etc.	Learners' accuracy rates vary across different types of collocation, and L1 influence seems to be a major cause
Associative Strength	the degree of exclusivity between the two component words of a collocation	indicate how keen one is in noticing, absorbing and making use of strongly-linked collocations	Mutual Information (MI)	Bestgen & Granger (2014), Durrant & Schmitt (2009), Eguchi & Kyle (2023), etc.	Learners tend to use more frequent yet weaker collocations
Collocation Density	how often certain types of collocation manifest in one's utterances	indicate whether one is overusing and/or underusing certain types of collocation	Number of collocations per 1000 words	Peng (2016)	Learners tend to underuse Verb-Noun collocations
Collocation Diversity	how diverse one's collocations are	indicate the size of one's collocation repertoire	Lemmatized collocation counts / raw collocation counts	none	/
General Lexical Level	the overall level of words used in one's utterances	indicate one's lexical proficiency	Distribution of percentages according to general service list (West, 1953)	none	/
Collocation Lexical Level	the level of words used in one's collocations	indicate one's ability of using known words in making collocations	Distribution of percentages according to general service list (West, 1953)	none	/

As the table shows, this study will expand the scope of profiling to a total of six aspects. The first two (i.e., accuracy rate and associative strength) are often covered in previous studies, yet whether their findings apply to intermediate learners and/or the "modifier-head" types of collocations remains to be confirmed. The third one, collocation density, was less explored previously and the findings were not without room for doubt, as discussed shortly before. We believe the interpretation of collocation density would be much more straightforward for the non-obligatory "modifier-head" types (i.e., Adj-N, Adv-Adj, Adv-V) than for the obligatory "head-complement" types (i.e., N-V, V-N, V-Adj). As for the last three aspects, namely, collocation diversity, general lexical level and collocation lexical level, it seems that no previous studies have ever touched on them, to the best of our knowledge. Therefore, any findings concerning these aspects should serve as a valuable piece to the overall picture. By encompassing all the aspects above, we expect the current study will offer fresh insights into learners' collocational competence and further the enterprise towards a comprehensive profile of such competence.

Two Kinds of Linguistic Knowledge

Another major contribution this study aims to make to collocational competence research is related to knowledge itself, which, in turn, serves as the foundation for competence (Ellis, 1994). There exists an established distinction between two kinds of linguistic knowledge: implicit and explicit. This notion is closely associated with Krashen's well-known Monitor Theory with its five hypotheses. In the elaboration of the acquisition-learning distinction hypothesis, Krashen (2009) stated that L2 learners have two contrasting ways (i.e., acquisition and learning) of developing their competence, each of which leads to a different kind of knowledge (i.e., implicit and explicit, respectively). While the significance of implicit knowledge is unquestionable, that of explicit knowledge arouses much dispute. Some scholars (e.g., Krashen, 2009) considered the role of explicit knowledge in second language acquisition to be very limited while others (Ellis, 1994; Gass, 1988) deemed that this type of knowledge deserves more credit than that. For example, it could help learners with noticing as well as comprehending input, thus facilitating the development of their implicit knowledge.

Therefore, to fully account for a learner's collocational competence, it is necessary to measure both his explicit and implicit knowledge. However, none of the existing collocation studies ever differentiated these two kinds, which is quite clear from their data. Many studies simply extracted a fraction from a learner's *written* corpus to analyze (Bestgen & Granger, 2014; Durrant & Schmitt, 2009; Eguchi & Kyle, 2023; Nesselhauf, 2003). As such writings are typically produced without instant pressure, they should reflect one's implicit and explicit knowledge combined. By only having such data, we will not be able to tell them apart. Fewer studies choose to focus on the *oral* production from learners. Peng (2016) examined the use of Verb-Noun type of collocation by advanced learners of Chinese by conducting 1-on-1 interviews to elicit their spoken data. Due to its online nature, such oral performance could only shed light on learners' implicit knowledge. An even less satisfactory situation lies with many other studies which relied on some indirect measurements instead of authentic production. For example, Chen (2017) designed gapped sentences for participants to fill in with the fifty-two target collocations. Vu and Peters (2023), in their study on the effect of different learning modes, designed a test of form recall. Webb and Chang (2022) implemented both word matching and meaning recall tests to measure the development of learners' collocational competence. As all these studies administered some kind of test rather than obtaining authentic utterances of the learners, not only do they fall short in differentiating implicit knowledge and explicit knowledge, but also seem questionable regarding the validity of their conclusions (Nesselhauf, 2003).

To fill this gap, the current study aims at tapping into both implicit and explicit collocational knowledge of the learners for a more detailed profile of their relevant competence. One of the major distinctions between these two kinds of knowledge is about their level of accessibility. While one can easily and often unconsciously draw on her implicit knowledge in communication, she has to be given ample time to truly make use of the explicit knowledge (Krashen, 2009). To maximize the ecological validity (Eisenbeiss, 2010), we collected naturalistic production data with minimal intervention from the researchers. These data include spontaneous speech which is based on learners' implicit knowledge (Ellis, 1994), and untimed writing which accounts for their explicit knowledge together with their implicit knowledge. This way, we expect to learn whether learners' implicit collocational knowledge differs from their explicit one, and if yes, how they differ from each other.

In summary, this study has two research questions as follows:

1. What is the collocational competence profile of intermediate Chinese EFL learners?
2. What is the difference between their implicit collocational knowledge and explicit

collocational knowledge, if any?

Methodology

Participants

The target population for this study is Chinese intermediate EFL learners who grew up in their L1 context. We used purposeful (nonprobability) sampling as we recruited volunteers among English-major undergraduates from a university, and any students that were willing to participate took the placement test for their proficiency. Their backgrounds are as follows:

- Number of participants: 84
- Age range: 18-20
- Major of study: English
- Proficiency level: intermediate

Placement Test

We used the online Cambridge English Test to identify qualified participants (i.e. intermediate learners) for the experiment. This test, composed of 25 multiple-choice questions, seems to be the best choice due to its relevance, reliability and cost-effectiveness. It encompasses the collocational aspect, with questions containing the V-N/Adj-N/Adv-Adj types of collocations (e.g., *order a pair of shoes*, *unfasten one's seatbelt*, *retrace one's steps*, *highly reliable*). The test was administered to a total of 245 students who signed up for the test. It was carried out within the time limit of 15 minutes under the monitoring of the researchers to ensure reliable results. In the end, 84 students whose scores turned out to fall within the range of intermediate users (i.e., between 14 and 21, as Cambridge's official interpretation of the score claims) agreed to participate by signing the consent form.

Data Collection

As mentioned earlier, we collected both spontaneous oral data and untimed writing data from the participants to learn about their implicit and explicit collocational knowledge. For the oral part, each participant was interviewed individually and asked to give a 3-minute talk on a common topic, such as "*why do you choose this major,*" "*tell me something about your hometown,*" etc. Their talks were recorded, transcribed and proofread for further analysis. Since the topics were not made known to them beforehand, their impromptu online performance should largely be based on their implicit knowledge alone. For the written part, they were given four days to compose an essay on a common yet major-relevant topic ("*the significance of reading*" and "*how to learn a language*"). The essay should be composed of at least 200 words, and should be done without any external help, such as consulting a dictionary. Since the participants have ample time to conceive, compose and revise their works, their essays should reflect the overall condition of their implicit and explicit knowledge combined. We concede that, although both tasks were assigned common topics to facilitate production, the topics for interviews have to be different from, and possibly less formal than, those for essay writing. This difference could compromise the comparability of the two kinds of production to some extent.

We also built corresponding native speaker (NS) corpora for comparison from the Contemporary Corpus of American English (Davies, 2008). Of all the eight genres in COCA (i.e. SPOKEN, TV/Movies, FICTIOIN, MAGAZINES, NEWSPAPER, ACADEMIC, WEB-

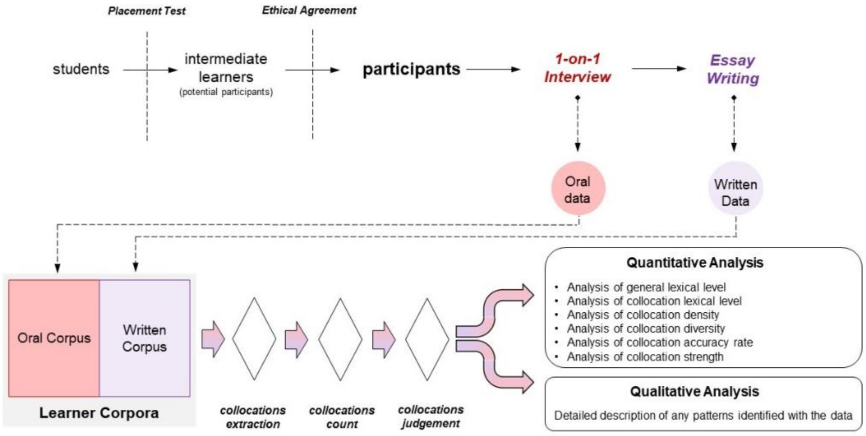
GENL, WEB-BLOG), the first two are oral in nature while the remaining six are written. The principle of balance and randomization was adhered to in this corpora-making process. To build the NS oral corpus, we randomly picked ten pieces of 300-word-long texts from the two oral genres (five from each), and thus the token count amounts to about 3,000. To build the NS written corpus, we randomly picked six pieces of 500-word-long texts from the six written genres (one from each), making the total count also about 3,000. This way, each of the learner (i.e. non-native speakers, or NNS) corpora has its NS counterpart to compare with.

Procedure

The following chart (*Figure 3*) illustrates the experiment design. Based on the results of the placement test, 84 intermediate learners were chosen to participate in the experiment. We first conducted a 1-on-1 interview with each participant and then gave them a topic for essay writing which was due four days later. All the interviews were transcribed and formed the oral corpus while all essays were collected and formed the written corpus. The size of the oral corpus and written corpus amounts to 36,046 and 40,968 tokens, respectively.

Figure 3

The Experiment Procedure



After the learner corpora were formed, we first manually checked through every piece of utterance to eliminate any spelling error (as long as the orthodox form is clear from the context). The reason for this is that, in this study, we are not concerned with any inadequacy in spelling but just want to tap into the underlying knowledge behind their usage of collocations. If the spelling errors remain uncorrected, software would not be able to properly POS-tag the tokens and the results would not faithfully reflect their true collocational competence. After getting rid of any spelling errors, we POS-tagged the corpora, and extracted all the "modifier-head" type of collocations with the help of several computer programs, including Laurence Anthony's AntWordProfiler, AntConc and TagAnt (Anthony, 2022a, 2022b, 2023).

Results and Discussion

Overall, we performed a 2×2 fashion of investigation of the data, encompassing two dimensions of comparison and two categories of analysis. The two dimensions include external comparison and internal comparison. The former means to compare NNS performance with that of NS to profile learners' collocational competence in general. The latter means to compare NNS oral performance with their written performance, to reveal any discrepancy between their implicit and explicit collocational knowledge. The two categories of analysis include quantitative analysis and qualitative analysis. In the quantitative part, we shall examine six aspects: general lexical proficiency, collocational lexical proficiency, collocation accuracy rate, collocation density, collocation diversity, and collocation associative strength. In the qualitative part, we will dive into the actual cases of learners' collocation use, aiming to identify any pattern and peculiarities of their performance.

Quantitative Analysis

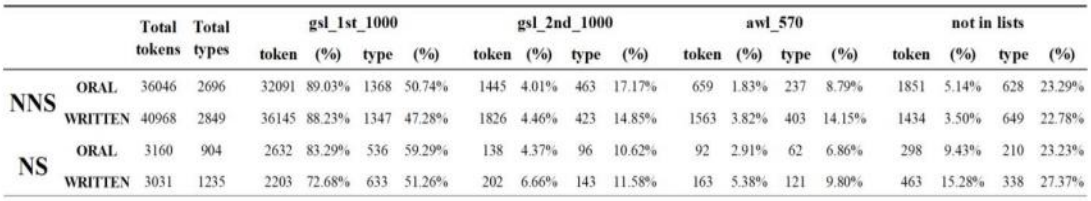
In this section, we shall examine the six aspects of NNS's collocational competence. As explained earlier, the two-dimension comparison will be made first between NNS and NS, and then within NNS, between their oral and written performances.

General Lexical Level

We imported each corpus of interest into the software *AntWordProfiler* which would calculate the percentage of words falling in each category of the General Service List (West, 1953), including "gsl_1st_1000" (i.e., the first thousand headwords), "gsl_2nd_1000" (i.e., the second thousand headwords), "awl_570" (i.e. the 570 headwords of the Academic Word List by Coxhead, 2000), and "not_in_list" (i.e., any word not included the first three categories, usually of advanced/specialized vocabulary). To ensure accuracy of the results, any spelling errors were corrected before the import. This results in our confidence of "not_in_lists" words being actual advanced words.

Table 2

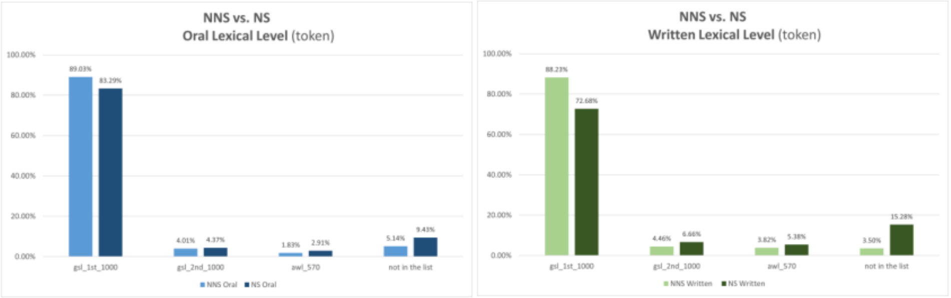
General Lexical Level Profiling



		Total tokens	Total types	gsl_1st_1000				gsl_2nd_1000				awl_570				not in lists			
				token	(%)	type	(%)	token	(%)	type	(%)	token	(%)	type	(%)	token	(%)	type	(%)
NNS	ORAL	36046	2696	32091	89.03%	1368	50.74%	1445	4.01%	463	17.17%	659	1.83%	237	8.79%	1851	5.14%	628	23.29%
	WRITTEN	40968	2849	36145	88.23%	1347	47.28%	1826	4.46%	423	14.85%	1563	3.82%	403	14.15%	1434	3.50%	649	22.78%
NS	ORAL	3160	904	2632	83.29%	536	59.29%	138	4.37%	96	10.62%	92	2.91%	62	6.86%	298	9.43%	210	23.23%
	WRITTEN	3031	1235	2203	72.68%	633	51.26%	202	6.66%	143	11.58%	163	5.38%	121	9.80%	463	15.28%	338	27.37%

Figure 4

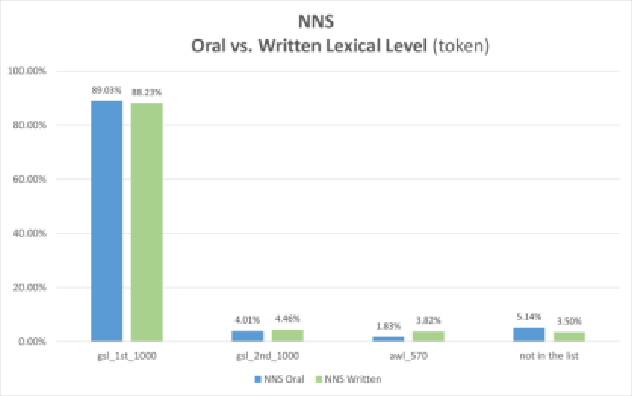
General Lexical Level Comparison (NNS vs. NS)



In the comparison between NNS and NS, it is clear that NS possess a more advanced vocabulary than NNS. The gap is already plainly visible in oral data, but even more pronounced in written data. For example, in writing, NS uses "not_in_lists" words four times more often than NNS (15.38% vs. 3.50%), and about one and half times more often with regards to "awl_570" words (5.38% vs. 3.82%). What this means is that NS's implicit lexical knowledge is superior to NNS's, and the margin between them is even wider in terms of explicit knowledge. This, in turn, serves as the basis for their collocational performance. As we explained earlier, individual word knowledge normally precedes the knowledge of how to use it in collocations.

Figure 5

General Lexical Level Comparison (NNS oral vs. NNS written)

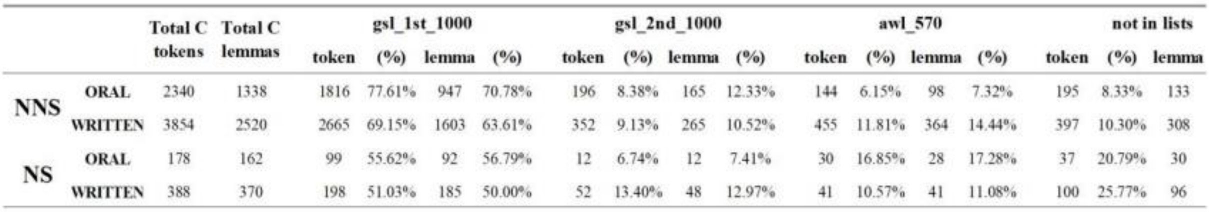


In the internal comparison within NNS (i.e., their oral vs. their written), the difference is surprisingly minimal. Therefore, it seems to indicate that, on the general lexical level, learners' implicit and explicit knowledge are quite the same. This is important to keep in mind because, as we shall see, when examining NNS's collocational performance, their oral and written performances differ in virtually every aspect. In other words, one's general lexical proficiency and his collocational proficiency are two related yet separate entities.

Collocation Lexical Level

Table 3

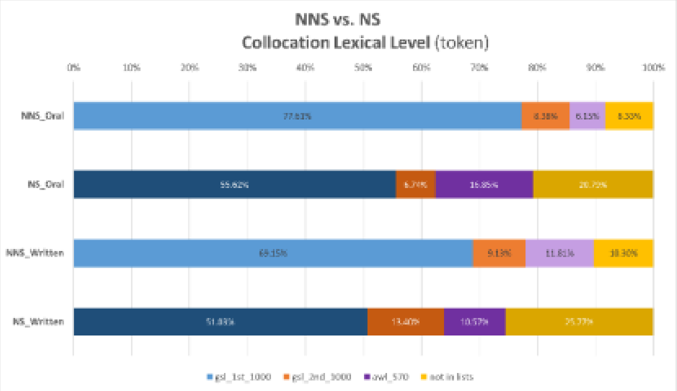
Collocation Lexical Level Profiling



		Total C tokens	Total C lemmas	gsl_1st_1000				gsl_2nd_1000				awl_570				not in lists		
				token	(%)	lemma	(%)	token	(%)	lemma	(%)	token	(%)	lemma	(%)	token	(%)	lemma
NNS	ORAL	2340	1338	1816	77.61%	947	70.78%	196	8.38%	165	12.33%	144	6.15%	98	7.32%	195	8.33%	133
	WRITTEN	3854	2520	2665	69.15%	1603	63.61%	352	9.13%	265	10.52%	455	11.81%	364	14.44%	397	10.30%	308
NS	ORAL	178	162	99	55.62%	92	56.79%	12	6.74%	12	7.41%	30	16.85%	28	17.28%	37	20.79%	30
	WRITTEN	388	370	198	51.03%	185	50.00%	52	13.40%	48	12.97%	41	10.57%	41	11.08%	100	25.77%	96

Figure 6

Collocation Lexical Level Comparison (NNS vs. NS)



Now we examine the aspect of collocation lexical level. We simply imported the extracted collocations, rather than entire texts, into the software to generate the percentages. The trend is again quite clear. In the NNS/NS comparison, it seems that NS use much more advanced vocabulary words (i.e., "awl_570" and "not_in_lists") to make up collocations than NNS do. Interestingly, the relative width of the gaps is reversed this time, larger for oral than for written. This not only confirms that depth-wise NS's lexical knowledge are superior than NNS's, but also indicates that learners face greater difficulty in developing implicit collocational knowledge. This exactly matches what Wray's dual model would predict (Wray, 2002). Learners are used to the bottom-up mode, rarely absorbing multi-word unit as a whole but routinely breaking them down and recomposing their own combinations consciously.

This speculation seems to be affirmed in the internal comparison between NNS oral and NNS written data. It shows that, unlike general lexical proficiency, learners' oral performance is clearly inferior to their written performance (14.48% vs. 22.11%, if we consider "awl_570" together with "not_in_list" to represent advanced vocabulary). This means two things: (a) despite one's use of an advanced word in a general way, he or she may not be able to use it in collocations. In other words, general lexical proficiency and collocational lexical proficiency are related yet separate, with the latter being harder to develop. (b) Learners tend to make explicit progress more easily than implicit progress on this multi-word entity, again testifying the validity of Wray's dual model.

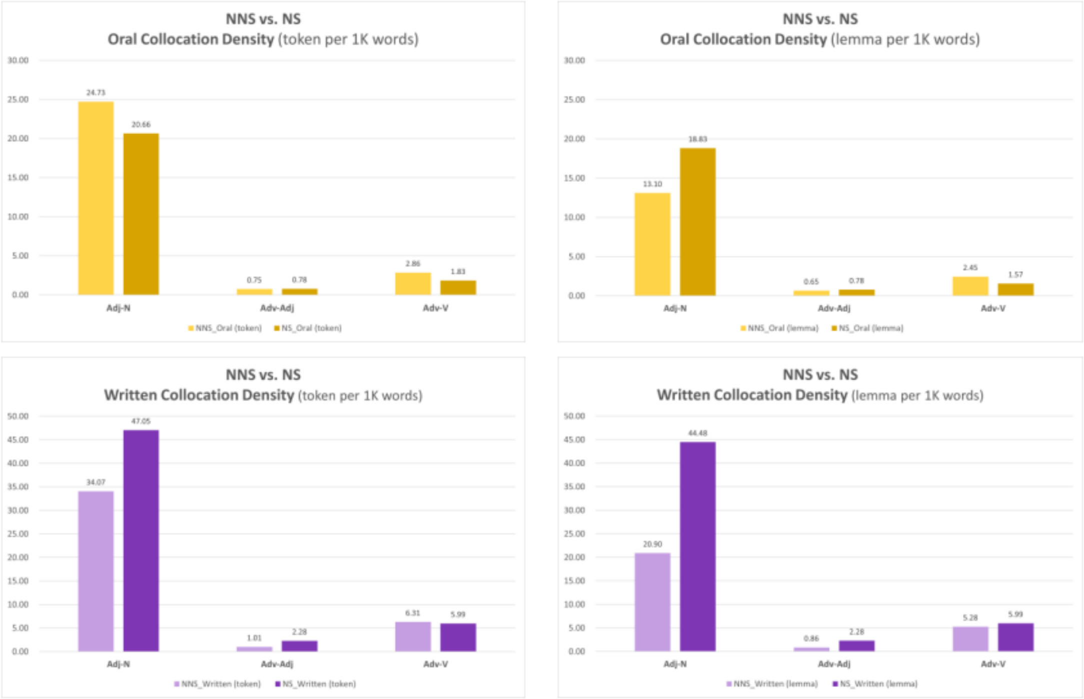
For pedagogical implications, this result calls for modification of learners' schema of linguistic knowledge, from bottom-up to top-down. Also, it proves that word knowledge has depth, with collocational aspect being more advanced than individual word knowledge, and worthy to develop.

Collocation Density

The third aspect for investigation is collocation density. It refers to the normalized collocation token (or lemma) counts per 1000 words, which reflects the degree of "pervasiveness" of collocations in a text. Taking NS as the reference, we shall know whether learners are underusing/overusing certain categories of collocations, and to what extent they are doing it.

Figure 7

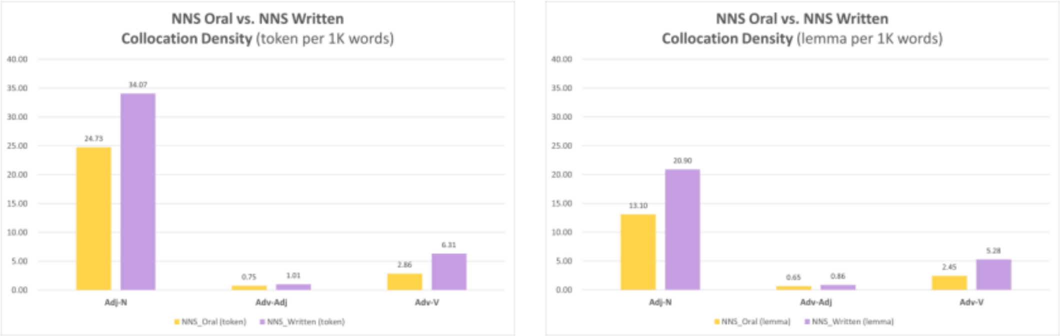
Collocation Density Comparison (NNS vs. NS)



We will first make the external comparison. Interestingly, in the oral part, NNS produced more collocation tokens than NS. Does it mean that NNS are more capable than NS here? The answer is likely negative. This is not a the-more-the-better story, because to truly proficient language users, other means (e.g., relative clause) to achieve the same desire effect are at their disposal. It is more plausible that NS are naturally striking an optimum balance between collocations and other competing linguistic means. Furthermore, in the charts for lemma, the trend is somewhat reversed, especially for the Adj-N category. This suggests that although NS did not produce as many collocation tokens in speaking, their production is showing a better diversity. In the written part, the picture is more straightforward. NS outperform NNS in virtually every category, and the gap is more profound for lemma than for token. We could draw a tentative conclusion here that, in general, learners are underusing collocation in their production, especially in writing.

Figure 8

Collocation Density Comparison (NNS oral vs. NNS written)



The internal comparison shows a clear and consistent trend. Learners use more collocations in writing than in speaking, both in terms of token and of lemma. This enables us to feel the impact of one's explicit knowledge: when learners have access to their explicit knowledge, they can produce better language. Therefore, the value of explicit knowledge is recognized. This contradicts with some scholars' view (e.g., Krashen, 2009) that explicit knowledge is marginal in L2 acquisition.

Therefore, it seems that learners' explicit collocational knowledge plays a crucial role in their performance, and they generally underuse collocations than NS do, probably due to their underdeveloped repertoire of collocations and lack of automaticity of using them.

Collocation Diversity

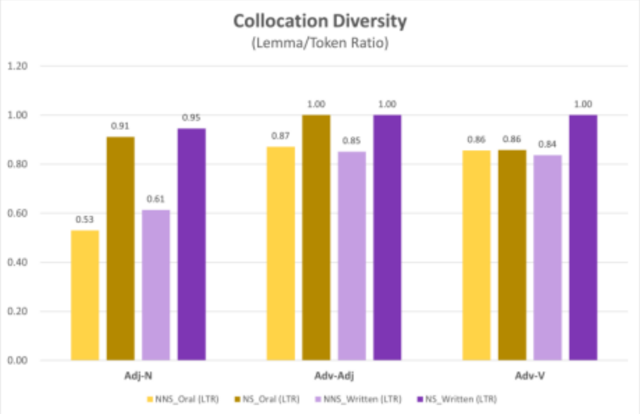
Table 4

Collocation Diversity Profiling

		Adj-N			Adv-Adj			Adv-V		
		Token	**lemma**	**TLR**	**Token**	**lemma**	**TLR**	**Token**	**lemma**	**TLR**
NNS	**Oral**	24.73	13.10	0.53	0.75	0.65	0.87	2.86	2.45	0.86
	Written	34.07	20.90	0.91	1.01	0.86	1.00	6.31	5.28	0.86
NS	**Oral**	20.66	18.83	0.61	0.78	0.78	0.85	1.83	1.57	0.84
	Written	47.05	44.48	0.95	2.28	2.28	1.00	5.99	5.99	1.00

Figure 9

Collocation Diversity Comparison



The fourth aspect to examine is collocation diversity, which is measured by lemma token ratio (LTR). In external comparison, as we can see in the chart above, NS outperform NNS in all types of collocation, both oral and written. This indicates that NNS resort to repetitive use of a quite limited repertoire of collocations, with regards to both implicit and explicit parts. This surely is an area in which they need to make improvement, and they seemingly need improvement in Adj-N type the most since the gap is widest here (0.53 vs. 0.91 for oral, and 0.61 vs. 0.95 for written).

In internal comparison, it might be fair to say that one's explicit knowledge is more diverse than implicit, because the most prevalent type (i.e., Adj-N) shows such a trend. However, in the other two types it is not as straightforward. So, more data and investigation could help to make the picture clearer here.

One last word of caution before we move on. Because the overall counts of collocation are much more for NNS than for NS, the comparability of them in terms of LTR are called into question to an extent. It is generally acknowledged that LTR drops as the number of token increases.

Collocation Accuracy Rate

This is one of the most high-stake aspects of data analysis: collocation accuracy rate. After all, both teachers and students share the goal of minimizing errors. As the previous Figure 1 illustrates, to get the values, we first identify all collocations in the NNS corpora. Then each case is investigated in COCA. If it is listed in the node's collocation page, then it is counted as a *collocation*. If not, it is further passed on to two NS judges. If they agree that it is acceptable, then it is counted as a *free combination*, otherwise as an *unidiomatic combination*. For the NS corpora, the process is simpler: a token is automatically a free combination is not listed in COCA, with the assumption that NS do not make unidiomatic combinations.

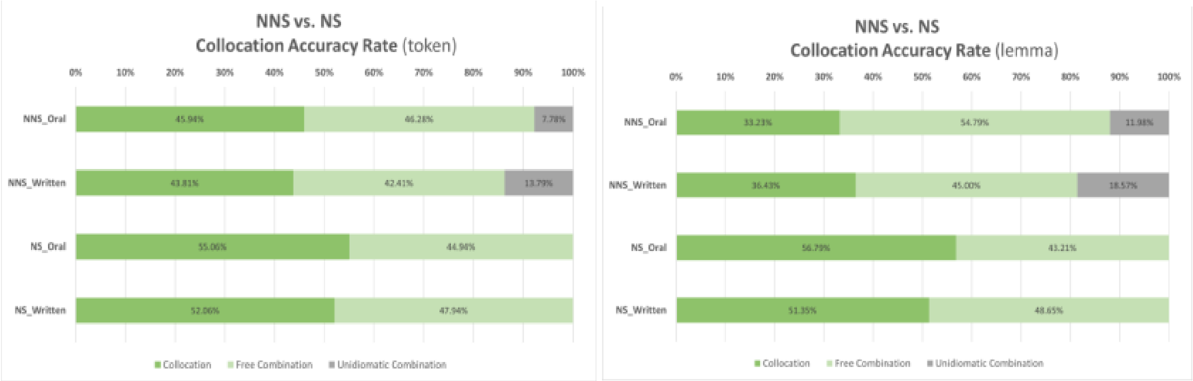
Table 5

Collocation Accuracy Rate Profiling

		Total C Token	Total C Lemma	Collocation				Free Combination				Unidiomatic Combination			
				Token	(%)	Lemma	(%)	Token	(%)	Lemma	(%)	Token	(%)	Lemma	(%)
NNS	Oral	1169	668	537	45.94%	222	33.23%	541	46.28%	366	54.79%	91	7.78%	80	11.98%
	Written	1929	1260	845	43.81%	459	36.43%	818	42.41%	567	45.00%	266	13.79%	234	18.57%
NS	Oral	89	81	49	55.06%	46	56.79%	40	44.94%	35	43.21%	0	0.00%	0	0.00%
	Written	194	185	101	52.06%	95	51.35%	93	47.94%	90	48.65%	0	0.00%	0	0.00%

Figure 10

Collocation Accuracy Rate Comparison



The trend from external comparison is crystal clear. NS outperform NNS by producing higher percentages of collocations (55.06% vs. 45.94% for oral, and 52.06% vs. 43.81% for written, respectively). This indicates that, unsurprisingly, NS possess a better feel about which words should habitually go together, and NNS, being somewhat blind on this matter, produce more free combinations (and, of course, unidiomatic ones). This again seemingly justifies Wray's position: NNS primarily rely on bottom-up mode which grants them too much freedom in composing multi-word expressions. Therefore, a mode shift could be beneficial.

Shifting our focus to internal comparison, we observe that NNS produce more unidiomatic combinations in writing than in speaking, and the margin is quite observable (13.79% vs. 7.78%). This suggests that, in writing, as learners have access to their explicit knowledge, they tend to make more mistakes. This calls into question the quality of their explicit collocational knowledge. Another thing worthy of notice is that the trend for collocation is reversed in terms of lemma. That is, lemmawise, learners produce more collocation in writing than in speaking, but tokenwise, it is the opposite. This indicates that they make more repetitions of the "safe ones" in speaking, while in writing, they are more willing to take a risk to diversify their collocational performance (which could also partly explain why they make more erroneous ones in writing).

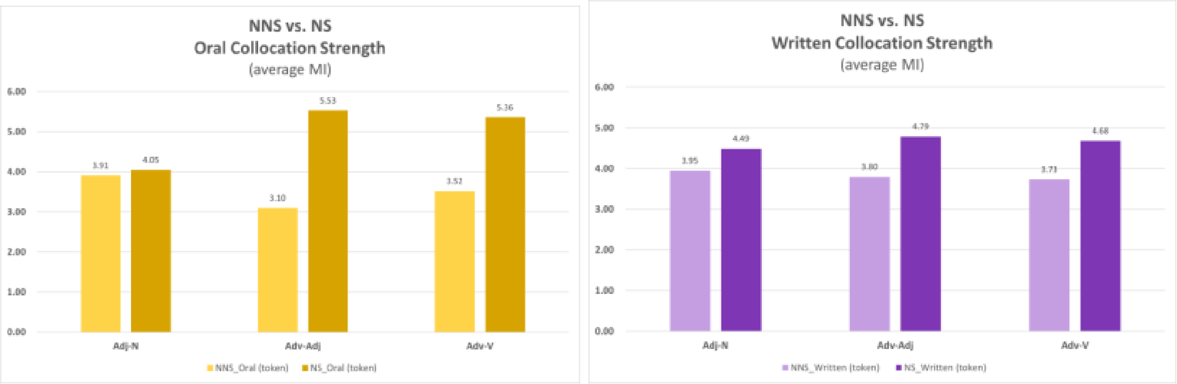
A last word before we move on. It is wise to keep in mind that, in reality, there is no clear dividing line between collocation and free combination, or even between free combination and unidiomatic combination. The difference is gradual rather than categorical. However, for the sake of presentation and discussion, we still have to adopt certain cut-off lines to categorize them.

Collocation Associative Strength

This is the last aspect of quantitative investigation, collocation strength. In all previous aspects, any word combination that fits the target syntactic patterns is included in the analysis; but here, only the true collocations proven by COCA (thus excluding free combinations and unidiomatic combinations) are involved. The parameter adopted is mutual information (MI), which indicates how strong the collocations are. This, in turn, should indicate how good the user feel about collocability among words (which is a part of his or her collocational competence).

Figure 11

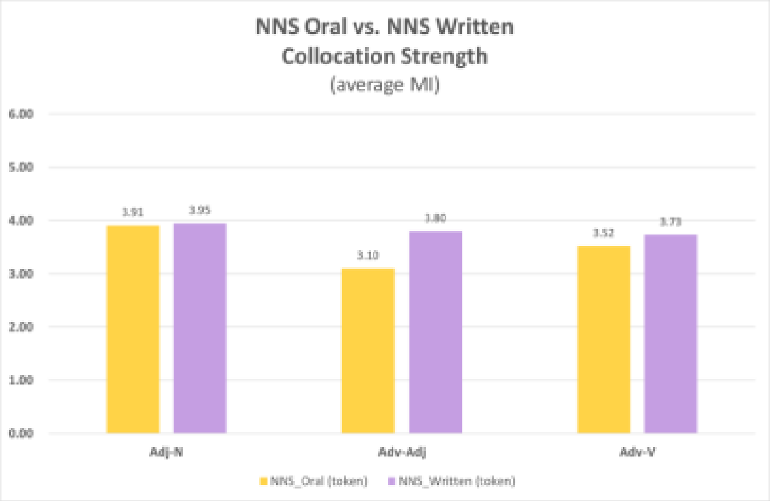
Collocation Associative Strength Comparison (NNS vs. NS)



We will follow suit and do the external comparison first. Clearly, NS outperform NNS in every single type, and the gap is considerable (e.g., 5.36 vs. 3.52 for Adv-V). This demonstrates that NS has a keener sense in perceiving and producing stronger collocations while NNS, probably due to lack of top-down mode, produce more free-combination-like collocations. This bears out the position of Sinclair and Wray (Sinclair, 1991; Wray, 2002) and echoes the conclusion of a number of studies (e.g., Eguchi & Kyle, 2023). Taking a closer look, we can see that the gap is larger for Adv-related types. This indicates while all types deserve attention, learners should devote more time to the ones involving adverbs.

Figure 12

Collocation Associative Strength Comparison (NNS oral vs. NNS written)



The internal comparison demonstrates a less pronounced trend. Learners produce slightly stronger collocations in writing than in speaking. This indicates that, when having time to reflect, learners could make a better choice in composing collocations. The contribution of their explicit knowledge is again felt and appreciated.

Qualitative Analysis

Besides quantitative analysis, a qualitative one would be also valuable, as it allows us to go beyond what mere numbers can reveal. We read through the subjects' speeches and essays to learn the patterns of their collocation making, aiming to account for the underlying reasons as well as ways of improvement. We identified several patterns that are worth mentioning, first from the learners' best exemplars, then from the errors they made.

Top Ten Lists

Due to space limitations, we would not be able to present and discuss every collocation the participants produced. Nevertheless, we deem it still enlightening to present the NNS' top ten collocations (i.e., the ten strongest ones) from each type and of each genre, as shown in the two tables below.

Table 6

The Ten Strongest Collocations Lists (Oral)

Adj-N			Adv-Adj			Adv-V		
Form	**Freq.**	**MI**	**Form**	**Freq.**	**MI**	**Form**	**Freq.**	**MI**
fluent English	363	8.58	*pretty good*	24448	3.89	*speak fluently*	330	7.71
deep breath	9431	7.90	*hard enough*	4167	3.83	*greatly improve*	925	5.93

cross-cultural communication	151	6.88	*deeply upset*	94	3.76	*read fluently*	108	5.71
fierce competition	598	6.85	*fully convinced*	97	3.59	*significantly improve*	1168	5.04
daily routine	1464	6.77	*increasingly aware*	260	3.50	*live happily*	1410	4.96
spoken English	114	6.70	*especially true*	2165	2.62	*listen carefully*	2170	4.91
interpersonal communication	309	6.65	*good enough*	14896	2.57	*work hard*	25386	4.83
heavy rain	2445	6.27	*greatly encouraged*	60	2.35	*quickly soured*	28	4.13
rural area	6183	6.11	*big enough*	6281	2.21	*read carefully*	1201	3.71
final exam	849	5.97	*way hotter*	86	2.09	*travel domestically*	18	3.68

Table 7

The Ten Strongest Collocations Lists (Written)

Adj-N			**Adv-Adj**			**Adv-V**		
Form	**Freq.**	**MI**	**Form**	**Freq.**	**MI**	**Form**	**Freq.**	**MI**
rote memorization	114	14.25	*commonly used*	1162	7.4	*speak fluently*	330	7.71
literal translation	540	9.13	*firmly convinced*	119	6.08	*communicate effectively*	820	6.71
childlike innocence	56	8.74	*extremely difficult*	2344	5.36	*greatly enhance*	449	6.52
fluent English	363	8.58	*seemingly impossible*	270	4.94	*exercise regularly*	504	6.31
idiomatic expressions	45	8.53	*extremely important*	2468	4.16	*sincerely hope*	758	6.26
English subtitles	613	8.42	*moderately difficult*	51	3.56	*greatly expand*	537	5.99
intercultural communication	153	8.26	*increasingly aware*	260	3.5	*greatly improve*	925	5.93
human being	31166	8.08	*spiritually empty*	13	3.39	*improve greatly*	925	5.93
rote learning	117	7.72	*increasingly important*	1197	3.32	*closely link*	811	5.73
spoken language	583	7.43	*increasingly impatient*	1197	3.12	*increase greatly*	1124	5.67

An internal comparison within either genre would reveal the fact that learners do not grasp these three types equally well. From the oral production lists, we can see that the overall strength of Adj-N collocations is the strongest, with an MI over 6 for virtually every case (except for *final exam* which barely falls short). The second strongest group is the Adv-V type, the majority of which have an MI of 4 to 6. The Adv-Adj type, being the weakest, have an MI that is consistently below

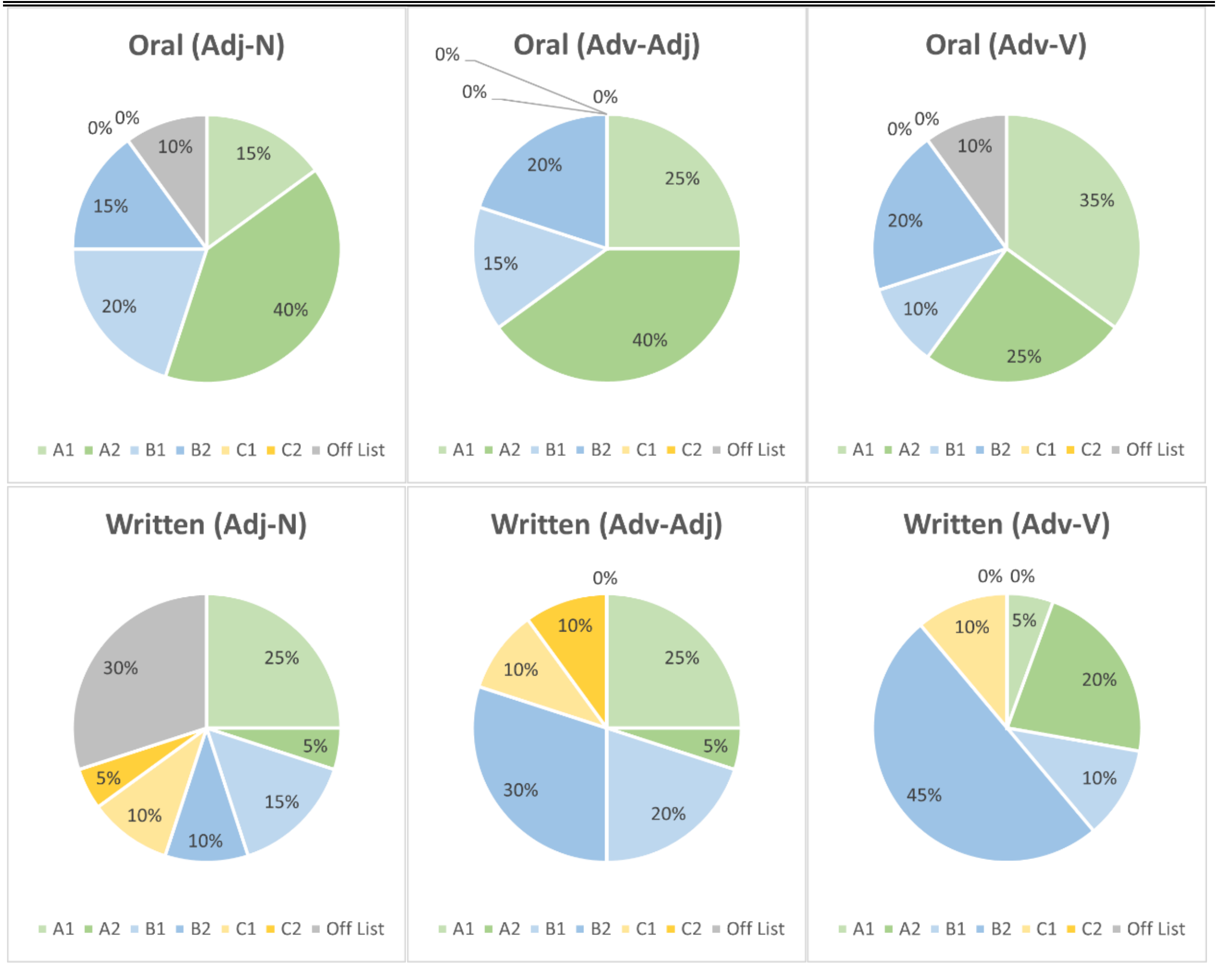
4, even for the top one *pretty good*. This order of sequence holds true for the written production as well. This indicates that, although these three types of collocations belong to the same "modifier-head" category, learners may not notice them and produce them in the same manner. More specifically, they seem to overlook strong Adv-Adj ones the most and should be advised to look out for this type and acquire them holistically during their daily encounter with input.

A cross-genre comparison between their oral performance and written performance would shed light on any discrepancies in learners' implicit and explicit knowledge. For the Adj-N type, it is quite clear that the ones produced in writing are stronger than those in speaking. While there is only one oral Adj-N collocation with an MI higher than 8 (*fluent English*), there are as many as eight written ones passing this threshold (the MI of the top one, *rote memorization*, is even as high as 14.25!). The comparisons concerning the other two types (i.e., Adv-Adj and Adv-V) tell the same story. Overall, it seems that the explicit collocational knowledge of the learners is superior to their implicit one, as far as associative strength goes. Since implicit knowledge is largely from incidental learning (Loewen, 2020), we would suggest that a potential weak point lies in learners' insufficient ability to notice and acquire collocations as whole units during their extensive exposure to input. Awareness raising activities might be beneficial in this case.

Besides the matter of strength, we shall also investigate the level of words contained in these collocations to understand to what extent learners are capable of making use of their vocabulary to compose collocations. The following six pie charts show the distribution of percentages across CEFR levels of A1 (light green), A2 (dark green), B1 (light blue), B2 (dark blue), C1 (light yellow), C2 (dark yellow), and off list (grey) for the 20 words of each type of each genre. The external comparison across the two genres shows a clear trend: learners use much more advanced vocabulary to make collocations in writing than in speaking. On the one hand, no C level word (i.e., C1 and/or C2, colored in yellow) is found in any type of oral collocations, yet they show up in all three types of written collocations, ranging from 10% to 20%. On the other, the percentages of A level words (i.e., A1 and/or A2, colored in green) are consistently above 50% for all types of oral collocations, but they never account for more than 30% for written collocations.

Figure 13

Word Level Distribution of the Top Ten Collocations



Providing that the overall lexical levels of the two genres are quite the same (a finding from the first section of quantitative analysis), we tend to conclude that the depth of the learners' explicit lexical knowledge is better than that of their implicit one. To narrow such a gap, a possible solution, we would suggest, is still to establish the habit of acquiring collocations holistically during extensive reading or listening, especially for those containing advanced words.

Next, we shall examine the collocational errors learners made, starting from overt errors and then covert ones.

Overt Errors

Some errors seem to be due to the influence of the learners' first language, since the corresponding L1 collocations are acceptable (Nesselhauf, 2003). Examples include: **deep communication 深度交流, *extracurricular book 课外书籍, *fast-paced age 快节奏的时代, *fragmented reading 碎片化阅读, *whole daytime 整个白天, *whole person 全人, *remember mechanically 机械式记忆, *officially learn 正式学习, *learn seriously 认真学习, *master firmly 牢牢掌握, *face positively 积极面对, *follow tightly 紧紧跟随,* etc. Learners are well advised to heed the danger of mental translation for making collocations. However, there does exist an L1 facilitating effect as well, as *ancient era 古时, basic expression 基本的表达*, etc. could testify. Therefore, learners should watch out for discrepancies between their L1 and L2 with regards to collocation making, but not discarding their L1 completely.

Certain errors seem to result from learners' inadequate word knowledge. These inadequacies are related to countability of nouns (e.g. **appropriate audios -> appropriate audio*), POS of words (e.g. **calm word -> calming word, *extrovert classmate -> extroverted classmate*), and semantic range of words (e.g. **personal culture*, in which *culture* entails a mass of people, not an individual).

Clearly, in order to produce sound collocations, one must possess solid word knowledge.

Some learners are insensitive to word order when making collocations, especially for Adv-V type (e.g. **learn seriously -> seriously learn, *remember firmly -> firmly remember, *grasp patiently -> patiently grasp*). Such unorthodox order may not seriously impede communication, but a rectification would definitely make the process smoother. This suggests the necessity to raise learners' awareness of the subtlety regarding the flexibility of word order in collocations and definitely avoid taking it too freely.

Covert Errors

As widely acknowledged, a correct form does not guarantee correct usage. Errors are not limited to those that have an unorthodox form but also include those used in a deviant manner. Unfortunately, such covert errors have been seldomly, if ever, researched in other collocational studies so far. Below is a list of covert errors identified from our NNS corpora.

Table 8

Covert Errors

Subject	Corpus	Concordance	Correction	Note
No. 2	Oral	Because I think teacher is a **good job** …	good profession	What the subject means is really good profession while *good job* is a compliment on someone's performance.
No. 15	Oral	I want to be an English teacher because it is a **good job**.	good profession	See previous note.
No. 18	Oral	I think teacher … it is a **good work**.	good profession	See previous note.
No. 7	Oral	…my first choice is **Chinese literature**.	Chinese language	The subject is talking about her major, and what she meant is Chinese language.
No. 13	Oral	I **got well** with all my English teachers ...	got along well	She really meant having a good relationship with the teachers.
No. 5	Oral	Fortunately, I passed the **first exam** and met so many friends in it...	first interview	The subject was talking about her interview to join a local club.
No. 17	Oral	I did some **voluntary activities**...	volunteer activities	The subject participated activities as a volunteer.
No. 1	Oral	I tried riding an **electric car** again...	electric bike	The subject was talking about learning to ride an e-bike on campus.
No. 1	Written	The **old word** said, "it is when you are using what you have learned from books..."	old saying	She meant old saying.

Other insufficiencies

Overuse of simple collocations. As we have seen from the diversity analysis earlier, learners generally fall short in this aspect. It is not too hard to find some examples, mostly composed of simple words (e.g., the "*very* + Adj" structure occurs 13 times in a 366-word speech!) Learners are advised to expand their collocation repertoire and diversify their production as a daily habit.

Failing to make the best choice. This lies in the grey area of right or wrong. Some collocations learners produced are arguably acceptable, but there is a better and more conventional way to say it (e.g., *convictive way* -> *convincing way*, *comprehensive application* -> *broad application*, *small action*

-> *small-scale action*, *white sky* -> *gray-white sky*, etc.) This possibly results from the bottom-up mode as well. Again, a switch of mode is called for.

Conclusion

Collocation, despite its seemingly straightforward concept, is indeed mysterious and tricky to profile. Having collected their natural production, both oral and written, and with clear criteria for collocation counting and judging as explained earlier, this study has strived to draw a picture of learners' collocational competence that is as comprehensive, accurate and enlightening as possible.

All together six aspects are investigated, namely, general lexical level, collocation lexical level, collocation density, collocation diversity, collocation accuracy rate, and collocation associative strength. For each aspect, two kinds of comparisons were made, both external (i.e., NNS vs. NS) and internal (i.e., NNS_oral vs. NNS_written). The former should cast light on the areas where learners need improvement, and the latter should reveal any discrepancies between learners' implicit knowledge and explicit knowledge.

The NNS vs. NS comparisons confirm that native speakers are more proficient than their non-native counterpart across all six collocational aspects. To be more specific, native speakers use a higher percentage of advanced vocabulary both in a general way and in collocation making; they produce more, stronger collocations; their production demonstrates better diversity. All of these indicate that the lack of collocational competence for intermediate learners is an all-round matter. Furthermore, as Figure 7 (concerning collocation density) and Figure 10 (concerning accuracy rate) show, the gap is wider by lemma than by token, which suggests that learners resort to overly repetitive use of their limited repertoire of collocations. This rather comprehensive, overall profile of their collocational competence carries two implications. First, a shift of language acquisition mode can be beneficial (Wray, 2002). When encountering a new word, learners should heed not only its individual information, but also the surrounding context along with boundaries of various levels, in order to naturally develop their multi-word unit repertoire. Second, in production, learners are advised to make more use of modifiers (e.g., Adj/Adv) to vivify their language, and the best way to do this is to accumulate and apply the thousands of readily available modifier-head types of collocations from NS sources, at least during the intermediate stage.

The internal comparisons between learners' oral and written production help to address the second research question: how different is their implicit collocational knowledge from their explicit one? From the results, we would conclude that they do not differ much in general lexical level, collocation diversity, and collocation strength. On the other hand, they demonstrate clear differences in collocational lexical level, collocation density, and collocation accuracy rate, with learners' written performance being superior to their oral one in these three aspects. It seems plausible to us that one's explicit knowledge plays an indispensable, rather than negligible, role in her second language acquisition. As for the quality of knowledge, although both kinds are not perfect, we are inclined to suggest that learners' implicit knowledge needs more improvement work, especially from the perspective of covert errors. Based on the general notion that the nature of learning largely corresponds to the nature of resulting knowledge (Loewen, 2020), an implication of such findings above is to adopt some balanced pedagogical approach, such as Focus on Form (Long, 1991), to develop both learners' implicit and explicit collocational knowledge. The incidental learning aspect should be properly emphasized when it comes to the development of collocational lexical level, collocation density, and accuracy rate. Potential remedial strategies include extensive exposure to L2 input with a keen eye on how words collocate, as well as a shift from bottom-up mode to a top-down one as mentioned earlier.

This study is not without certain limitations. First, as we briefly discussed in Methodology, the topics for the oral task and the written task may not be exactly on the same level as far as formality is concerned. Although the influence of this factor on collocational performance is not clear from previous studies, we concede that the comparability between oral and written production could be less than ideal. Another limitation lies in the measurement of the two kinds of knowledge. While instant accessibility is a valid distinguishing factor (Krashen, 2009), we admit that it is virtually impossible to elicit either type of data on an absolute basis. In other words, our participants were likely relying on both knowledge types simultaneously for each task, and it is only a matter of how much implicit/explicit knowledge contributed to their particular performance. Besides, we only examined the production side of learners' competence. It would be of value for others to probe into their comprehension side as well in future. It is our hope that the current study, standing on the shoulders of many before us, would pave the way for SLA practitioners to follow up with sound pedagogical approaches to effectively develop learners' collocational competence to the fullest extent possible.

Acknowledgements

We would like to express our gratitude to the two anonymous reviewers for their insightful comments and suggestions. We are also grateful to Nicole Lasas for being our language consultant for this study.

About the Authors

Zhiliang Yue: a full-time lecturer at Zhanjiang University of Science and Technology, China. His research interests include Applied Linguistics, English Language Teaching (ELT), and English for Specific Purposes (ESP).

Sugunya Ruangjaroon: a linguist specializing in Thai syntax and phonology, including topics like wh-questions and consonant-tone interaction. She currently researches syllable epenthesis in Thai compound words. Dr. Ruangjaroon also contributes to language teaching, co-authoring research in theoretical linguistics, second language acquisition and language pedagogy.

References

Alangari, M. A. (2019). *A corpus-based study of verb-noun collocations and verb complementation clause structures in the writing of advanced Saudi learners of English.* [Unpublished Doctoral Dissertation]. University of Reading.

Anthony, L. (2022a). *AntConc* (4.2.1). Waseda University.

Anthony, L. (2022b). *TagAnt* (2.0.5). Waseda University.

Anthony, L. (2023). *AntWordProfiler* (2.1.0). Waseda University.

Bestgen, Y., & Granger, S. (2014). Quantifying the development of phraseological competence in L2 English writing: An automated approach. *Journal of Second Language Writing*, *26*, 28–41. https://doi.org/10.1016/J.JSLW.2014.09.004

Biber, D. (2009). A corpus-driven approach to formulaic language in English. *International Journal of Corpus Linguistics*, *14*(3). https://doi.org/10.1075/ijcl.14.3.08bib

Chang, Y. C., Chang, J. S., Chen, H. J., & Liou, H. C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning*, *21*(3). https://doi.org/10.1080/09588220802090337

Chen, W. (2019). Profiling collocations in EFL writing of Chinese tertiary learners. *RELC Journal*, *50*(1). https://doi.org/10.1177/0033688217716507

Chen, Y. (2017). Dictionary use for collocation production and retention: A call-based study. *International Journal of Lexicography*, *30*(2). https://doi.org/10.1093/ijl/ecw005

Chomsky, N. (2014). *Aspects of the theory of syntax* (Issue 11). MIT press.

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment - Companion Volume*. Council of Europe Publishing.

Cowie, A. P. (2013). *Semantics*. Oxford University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2). https://doi.org/10.2307/3587951

Davies, M. (2008). *The Corpus of Contemporary American English (COCA)*. https://www.english-corpora.org/coca/

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL - International Review of Applied Linguistics in Language Teaching*, *47*(2). https://doi.org/10.1515/iral.2009.007

Eguchi, M., & Kyle, K. (2023). L2 collocation profiles and their relationship with vocabulary proficiency: A learner corpus approach. *Journal of Second Language Writing*, *60*, 100975. https://doi.org/10.1016/J.JSLW.2023.100975

Eisenbeiss, S. (2010). Production methods in language acquisition research. In E. Blom & S. Unsworth (Eds.), *Experimental methods in language acquisition research* (pp. 11–34). John Benjamins Publishing Company.

Ellis, R. (1990). A response to Gregg. *Applied Linguistics*, *11*(4). https://doi.org/10.1093/applin/11.4.384

Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.

Ferraro, G., Nazar, R., Alonso Ramos, M., & Wanner, L. (2014). Towards advanced collocation error correction in Spanish learner corpora. *Language Resources and Evaluation*, *48*(1). https://doi.org/10.1007/s10579-013-9242-3

Firth, J. R. (1969). *Papers in linguistics: 1934-1951*. Oxford University Press.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in corpus-based language learning research: Identifying, comparing, and interpreting the evidence. *Language Learning 67*(1). https://doi.org/10.1111/lang.12225

Gass S. (1988). Integrating research areas: A framework for second language studies. *Applied Linguistics*, *9*(2), 198–217.

Harmer, Jeremy. (2015). *The practice of English teaching language* (2nd ed.). Pearson Education ESL.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1). https://doi.org/10.1093/applin/19.1.24

Hymes, D. (1972). On communicative competence. *Sociolinguistics*, *269293*, 269–293.

Jiang, N., & Nekrasova, T. M. (2007). The processing of formulaic sequences by second language speakers. *Modern Language Journal*, *91*(3). https://doi.org/10.1111/j.1540-4781.2007.00589.x

Khonamri, F., & Roostaee, S. (2014). The impact of task-based extensive reading on lexical collocation knowledge of intermediate EFL learners. *Procedia - Social and Behavioral Sciences*, *136*, 265–270. https://doi.org/10.1016/j.sbspro.2014.05.326

Kim, S. H., & Kim, J. H. (2012). Frequency effects in L2 multiword unit processing: Evidence from self-paced reading. *TESOL Quarterly*, *46*(4). https://doi.org/10.1002/tesq.66

Krashen, S. D. (2009). Principles and practice in second language acquisition (1st Internet ed.). In *Sistem Otot* (Vol. 1).

Laufer, B., & Waldman, T. (2011). Verb-Noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning*, *61*(2). https://doi.org/10.1111/j.1467-9922.2010.00621.x

Loewen, S. (2020). *Introduction to instructed second language acquisition* (2nd ed.). Routledge, Taylor & Francis Group.

Long, M. H. (1991). Focus on form: A design feature in language teaching methodology. In K. de Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39–52). Amsterdam: John Benjamins.

McIntosh, C. (2009). *Oxford collocations dictionary* (2nd ed.). Oxford University Press.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, *24*(2), 223–242.

Öksüz, D., Brezina, V., & Rebuschat, P. (2021). Collocational processing in L1 and L2: The effects of word frequency, collocational frequency, and association. *Language Learning*, *71*(1), 55–98. https://doi.org/10.1111/lang.12427

Peng, X. (2016). *Use of verb-noun collocations by advanced learners of Chinese*. [Unpublished Doctoral Dissertation]. University of Pennsylvania.

Saeedakhtar, A., Bagerin, M., & Abdi, R. (2020). The effect of hands-on and hands-off data-driven learning on low-intermediate learners' verb-preposition collocations. *System*, *91*. https://doi.org/10.1016/j.system.2020.102268

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Siyanova-Chanturia, A. (2015). On the 'holistic' nature of formulaic language. *Corpus Linguistics and Linguistic Theory*, *11*(2). https://doi.org/10.1515/cllt-2014-0016

Stubbs, M. (2002). Two quantitative methods of studying phraseology in English. *International Journal of Corpus Linguistics*, *7*(2). https://doi.org/10.1075/ijcl.7.2.04stu

Tarone, E. E. (1990). On variation in interlanguage: A response to Gregg. *Applied Linguistics*, *11*(4). https://doi.org/10.1093/applin/11.4.392

Taylor, D. S. (1988). The meaning and use of the term 'competence' in linguistics and applied linguistics. *Applied Linguistics*, *9*(2). https://doi.org/10.1093/applin/9.2.148

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. *Language Learning and Language Teaching, 9*. https://doi.org/10.1075/lllt.9.09und

Vilkaite, L., & Schmitt, N. (2019). Reading collocations in an L2: Do collocation processing benefits extend to non-adjacent collocations? *Applied Linguistics*, *40*(2). https://doi.org/10.1093/applin/amx030

Vu, D., & Peters, E. (2023). A longitudinal study on the effect of mode of reading on incidental collocation learning and predictors of learning gains. *TESOL Quarterly*, *57*(1). https://doi.org/10.1002/tesq.3111

Walker, C. P. (2011). A corpus-based study of the linguistic features and processes which influence the way collocations are formed: Some implications for the learning of collocations. *TESOL Quarterly*, *45*(2). https://doi.org/10.5054/tq.2011.247710

Webb, S., & Chang, A. C. S. (2022). How does mode of input affect the incidental learning of collocations? *Studies in Second Language Acquisition*, *44*(1). https://doi.org/10.1017/S0272263120000297

West, M. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. Longman Publishing.

Widdowson, H. G. (1983). *Learning purpose and language use.* Oxford University Press.

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press.