



ChatGPT as a Facilitator of Vocabulary Learning Strategy Use for Lexical Bundles Across Different Difficulty Levels

Ryo Sawaguchi

*Graduate School of Foreign Language Education and Research,
Kansai University, Japan
rswgch@gmail.com*

Abstract

While there has been significant research into single-word vocabulary learning strategies (VLSs), few studies have examined effective VLSs for multi-word units, such as the lexical bundles (LBs) crucial for English for academic purposes (EAP). This study explores the VLSs used by Japanese university students to learn LBs. The participants were at B1 (independent) and A2 (basic) on the Common European Framework for Reference (CEFR). The participants, who received instruction in social (questioning), cognitive (gap-fill practice), and memory (functional grouping and L1 translations) strategies, studied 40 LBs (20 LBs for A2 and B1 each) and took A2- and B1 LB tests. Their VLS use was mediated and tracked in interactive out-of-class learning using the ChatGPT, artificial intelligence (AI) tool. The results found that combining strategies (e.g., social and cognitive) and the associated deeper cognitive processing are effective for both A2 and B1 LB learning, while the importance of the questioning (social)-based strategy combination and analytical learning increases for more difficult B1 LBs. The theoretical and practical applications of the findings are discussed.

Keywords: vocabulary learning strategies; multi-word units; lexical bundles; ChatGPT; difficulty.

Background

A key area in English for academic purposes (EAP) vocabulary learning/teaching is lexical bundles (LBs), such as *the fact that*. Because of the significance of LBs in EAP, their nature has been extensively studied (Biber et al., 2004; Cortes, 2004). For example, the LB *as a result* signals that a writer is connecting the preceding and subsequent sentences, creating logical transitions. The importance of LBs in academic contexts requires specific EAP learning/teaching materials and methods (Jones & Haywood, 2004). Specifically, wordlists have been found to be useful for intensive EAP vocabulary learning (Durrant, 2019). Wordlists also allow for autonomous learning using vocabulary learning strategies (VLSs) suited to students' learning styles, and can also be tailored to learners' ability with lexical sophistication variables (e.g., Nicklin et al., 2025).

Oxford (1990, p. 8) defined VLSs as a subcategory of learning strategies, which involves learner-directed behavior "to make learning easier, faster, more enjoyable, more self-directed, more effective." Because vocabulary tends to be learned without the direct involvement of instructors, VLSs and self-directed learning are highly compatible as they can enhance a learner's "own independent vocabulary-building skills" (Chiew & Ismail, 2021, p. 1301). While Alahmadi et al. (2018) and Paton (2020) differ in their views on learners' VLS use, with the former regarding VLSs as a matter of frequency and the latter motivation for sustained learning, they both agree that VLSs should be introduced to students. Ueno and Takeuchi's (2023) study, in which students perceived the effectiveness of VLS use, also supports the benefits of VLS instructions.

Recently developed artificial intelligence (AI) tools, such as ChatGPT, can assist in autonomous learning in ways not possible using traditional learning tools. For example, ChatGPT provides learners with vocabulary practice opportunities through real-time interactions (Ngo, 2024), such as question-answer exchanges, and can also develop gap-fill exercises tailored to learners' needs. Vocabulary learning practice using ChatGPT also aligns with social strategies (asking others to clarify the meaning of words) and cognitive strategies (practicing in meaning-focused contexts). Schmitt's (1997) VLS taxonomy includes social strategies in discovery strategies and cognitive strategies in consolidation strategies. Because the ChatGPT interactions with learners are recorded as chat logs, learners' strategy use can be accurately tracked, which overcomes the difficulties in measuring VLS use using only descriptive frequency questionnaires discussed in Ueno and Takeuchi (2023). This feature of ChatGPT also makes it possible to directly observe how learners study the items set at beyond ($i + 1$), current (i), or below ($i - 1$) their proficiency levels. This could be useful to explore whether previous studies' findings regarding difficulty apply to LB learning across different difficulty levels. Krashen's (1982) " $i + 1$ principle" and Laufer and Hulstijn's (2001) "Involvement Load Hypothesis" are similar in that language (e.g., vocabulary) learning is more effective when learners are involved in more cognitively demanding tasks, such as search (deeply process the meaning) and evaluation (using words in sentences). Keshmirshekan's (2019) reading (including vocabulary processing) intervention study supports these theories, with learners' performance declining when completing " $i - 1$ " reading tasks.

The close examinations of learners’ learning behaviors could provide insights into the appropriate VLSs required for various learning phases, which aligns with the globally accepted step-by-step learning evaluation scale, the Common European Framework for Reference (CEFR: Council of Europe, 2020). In Japanese English education contexts, setting LBs at A2 (basic) and B1 (independent) levels corresponds to the average university students’ proficiency levels measured by the Test of English for International Communication: TOEIC (institutional program test: 470, public test: 600, The Institute for International Business Communication [IIBC], 2025a). To explore the potential of ChatGPT-mediated VLSs effective for different LB learning stages (A2 and B1), this study set the following research questions that employed the hierarchical cluster analysis:

- RQ1: What ChatGPT-mediated VLSs are effective for learning A2 LBs?
- RQ2: Do effective ChatGPT-mediated VLSs change from A2 to B1 LBs?

Methods

Participants

The participants were 45 undergraduate students enrolled at a private university in Japan. All students voluntarily participated in the study after providing informed consent, which covered both privacy protection and the publication of the research findings. The participants were divided into two CEFR levels based on their self-reported TOEIC scores: 15 A2 students and 30 B1 students. Table 1 shows the participants’ TOEIC scores and corresponding CEFR proficiency levels (IIBC, 2025b).

Table 1 Participants’ CEFR Proficiency Levels

CEFR level	Corresponding score	N	Percentage	Mean	SD
A2	225–545	15	33.3%	525.3	19.0
B1	550–780	30	66.6%	575.1	19.8

Selecting Instructional LBs

The instructional LBs in this study were taken from the Essay Bundles List (EBL) developed by Sawaguchi (2025a). This list includes the high-frequency LBs commonly found in argumentative essays written by L1 English university students. The EBL was designed specifically for use in college-level English classes that emphasize basic academic writing skills development. As Wingate (2012) states, argumentative essays are an important writing form for the development of critical/logical thinking and argumentation skills and provide the basis for undergraduate students’ future academic English use, such as writing research papers at graduate schools. This study defines

LBs as bundles of two to four words. While previous studies have defined LBs as three or more word clusters (Biber et al., 2004; Cortes, 2004) that occur 20-40 per million times in five or more texts, some two-word blocks, such as *for example*, also have important discourse functions (Granger, 2019). Therefore, this study included two-word in addition to three to four-word LBs.

Forty LBs (see Appendix) were selected for the actual VLS intervention. Because the purpose of this study was to explore the effective VLSs for A2 and B1 LB learning, it was necessary to label each item with its corresponding CEFR difficulty. Therefore, A2 and B1 LBs in Sawaguchi (2025b) were consulted, which classified these LBs based on A2 and B1 examinee performances (i.e., θ values) on the test that had assessed their productive (gap-filling) knowledge of LBs. θ values of LBs in Sawaguchi (2025b) show the degree of possibility the items can be correctly answered by test takers. With 0 as the mean, test takers with higher proficiency (i.e., B1) are placed on the positive side (e.g., 0.1), while those with lower proficiency (i.e., A2) on the negative side (e.g., -2). Thus, an LB with an assigned θ value of 0.1 is difficult for A2 examinees, which distinguishes it from B1 examinees who could answer the item correctly.

The labeling resulted in 20 A2 and B1-level LBs each. Eighteen of the 20 A2-level LBs and 11 of the 20 B1-level LBs were selected directly from Sawaguchi (2025b) according to the LB's θ values. The remaining LBs were chosen primarily based on their frequency, followed by functional categorizations as a supplementary criterion. Sawaguchi (2025b) found that LBs with higher frequencies are generally easier, whereas those with lower frequencies are often more difficult, as evidenced in the study's θ values of LBs (A2: $-1.4 \leq \theta < -0.3$; B1: $-0.3 \leq \theta < 0.8$). One possible reason for the difficulty is that less frequent (i.e., B1) LBs often contain polysemous content words that express different senses of meaning and parts of speech, depending on contexts. In a B1 LB *a great deal of*, the content word *deal* functions as a quantifier, while in a *trade deal*, *deal* means a business (countable noun).

Regarding frequency, the term *frequency* refers to the estimated input frequency for Japanese university students, and it was determined from the English N-gram List for Japanese Learners of English (ENL-J: Ishikawa, 2019) and the Corpus of Contemporary American English (COCA: Davies, 2008). The ENL-J compiles frequency counts normalized to occurrences per 100,000 words from what can be considered major sources of input for Japanese university students: English language textbooks (middle and high school) and the National Center Test, an exam that evaluates high-school learning and for which many Japanese high school students prepare. Meanwhile, the COCA contains frequency information for modern American English. Both ENL-J and COCA can be useful for an assessment of the input frequencies for Japanese university students, as their English language education focuses on American English. To ensure alignment with the ENL-J, the raw COCA frequencies were normalized to a per 100,000-word basis. The study used the mean of ENL-J and COCA normalized frequency. If the ENL-J (textbook and national center exam) had the normalized frequency of 2 and 7, and the COCA had the normalized frequency of 3, the mean frequency of the LB was 4. The normalized mean frequency was 6.37 for A2 LBs and 1.78 for B1 LBs from Sawaguchi (2025b). This mean difference supports the earlier claim regarding the θ value in Sawaguchi (2025b) that B1 LBs (frequency mean 1.78)

can be considered more difficult than A2 LBs (frequency mean 6.37). Thus, the mean frequency for both levels was set as the first frequency filtering criterion to assign CEFR levels to the remaining LBs (2 LBs for A2 and 9 LBs for B1). The LBs with a frequency of 6.37 (A2 mean) or more were regarded as A2 LBs, while those less than 6.37 were categorized as B1 LBs. Table 2 presents the basic frequency statistics for the total 40 LBs (20 A2 and B1 each). The unpaired Mann-Whitney U-test found a statistically significant difference ($p = .003$) in frequency between the 20 A2 and B1 LBs, as shown in Table 3. Following the aforementioned difficulty discussion (less frequent = more difficult), B1 LBs were considered more difficult than A2 ones.

Table 2 Basic Frequency Statistics for A2 and B1 LBs

CEFR	N	Mean	SD
A2	20	6.55	6.74
B1	20	1.64	1.94

Table 3 Mann-Whitney U-Test Result for Frequency Differences between A2 and B1 LBs

Comparison	U	r	p-value
A2 vs. B1 LBs	79	.47	.003

Filtered A2 and B1 LBs underwent functional categorizations. The LBs were categorized into three functional categories: referential, stance, and discourse. Referential bundles specify the following objects (e.g., *aspect of, the presence of*), stance bundles express the writer's attitude on the topic (e.g., *argue/note that*), and discourse bundles create a smooth transition between sentences (e.g., *as a result, in order to*). Although the primary criterion of selecting LBs was frequency-based difficulty, the study made maximum efforts to balance the number of LBs in the above three functional categories. This was done to ensure compatibility with the grouping VLS (discussed later) and to explore VLSs applicable to LBs with varying functional categories. The author consulted Simpson-Vlach and Ellis's (2010) categorizations that extensively covered more than 200 LBs (as they termed *formulas*). If the LBs in this study were found in the categories in Simpson-Vlach and Ellis (2010), the study applied their functional categorizations. Thus, the LBs, such as *whether or not* and *in order to* were categorized as discourse bundles in the same way as Simpson-Vlach and Ellis (2010). When the LBs were not in Simpson-Vlach and Ellis's (2010) functional categorizations, the author manually categorized the LBs by referring to Simpson-Vlach and Ellis's (2010) functional categorizations in terms of syntactics and discourse functions. For instance, although *degree of* and *aspect of* were not in Simpson-Vlach and Ellis (2010), they share the same syntactic (noun + of) features as *a variety of* in Simpson-Vlach and Ellis (2010). *Degree of* and *aspect of* also specify the following nouns, which conforms to a feature of referential bundles. From a discourse perspective, an LB *in turn* can express

a similar meaning as a discourse bundle *as a result* in Simpson-Vlach and Ellis (2010), depending on the context. Thus, the LB was considered a discourse bundle in this study. The study applied these syntactic and discourse categorizations to the rest of the LBs not found in Simpson-Vlach and Ellis's (2010) functional categorizations, including stance bundles. The grouping resulted in 18 referential (45%), 12 stance (30%), and 10 (25%) discourse bundles in total.

Figure 1 illustrates the whole LB selection process in the study.

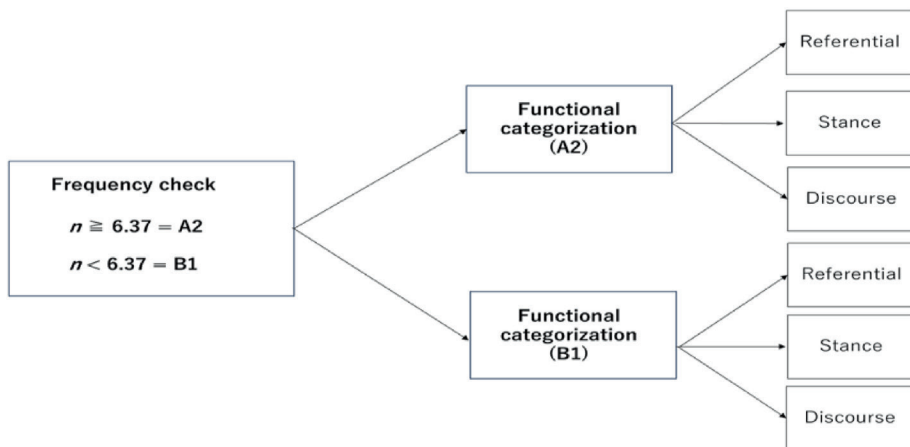


Figure 1 *Instructional LB Selection Process.*

Instructional VLSs

This study taught four VLSs to participants, which were divided into the discovery and consolidation strategies (Schmitt, 1997) found to be most influential in recent VLS studies (Chiew & Ismail, 2021). Discovery strategies involve the use of social strategies (Mizumoto, 2010) that involve asking others to clarify the meaning of newly encountered words. Consolidation strategies strengthen the learner's knowledge of words they already know. The discovery strategy in this study was (1) interactive questioning (social strategy), and the consolidation strategies were (2) gap-filling practice (cognitive strategy: Mizumoto, 2010), (3) grouping (memory strategy: Oxford, 1990), and (4) L1 Japanese translations (memory strategy). The rationales for choosing these four strategies are explained in the following.

- (1) Interactive questioning (social strategy): ChatGPT has significant potential as a social strategy use facilitator because it offers "more personalized and interactive vocabulary learning experiences" (Ngo, 2024, p. 3). This function corresponds to a key social strategy that assists in the learning of target words. For example, while learning the LBs, learners can check their correct understanding of the LBs by asking ChatGPT about their meanings.
- (2) Gap-fill practice (cognitive strategy): Mizumoto (2010) describes cognitive strategies as "those that deal directly with information and process it" (p. 8),

which could be interpreted as any form of cognitive effort, such as production. Specifically, Mizumoto (2010) introduces “activation strategies” (p. 42), which require learners to primarily use the new words by writing sentences or in gap-fill activities. This study employed gap-fill activities to align with previous studies that explicitly taught LBs using gap-fill exercises to complete sentences (Shamsabadi et al., 2017).

- (3) Grouping (memory strategy): Oxford (1990) includes grouping in memory strategies (creating mental linkages) to foster the retention of new words. Because LBs have been categorized into three functional categories (i.e., referential, stance, and discourse), they are compatible with grouping strategies that focus on the explicit teaching of LBs.
- (4) L1 translations (memory strategy): Chiew and Ismail (2021) regard learning vocabulary using flashcards as a consolidation strategy, as this can encourage the quick and straightforward repetition of newly learned words with L1 translations. Although ChatGPT cannot produce flashcards, this study took advantage of its quick creation of L1 translation tables readily available to learners. The effectiveness of this L1-L2 (bilingual) wordlist learning is supported in Sonbul and Macis (2021).

Procedure

The instructional intervention and data collection took place over three weeks: one week for the pilot experiment and two weeks for the actual intervention. During this time, the participants (henceforth learners) were introduced to 50 LBs (10 for pilot experiments and 40 for actual intervention) and provided with instructional prompts for the four VLSSs. Because this study focused on the learning of LBs useful for basic academic writing (i.e., argumentative essays), the learning outcomes were measured using form-recall tests, which indirectly assessed the learners’ productive (writing) knowledge of the LBs. The tests presented learners with partially blank bundles and asked them to recall and complete the missing parts. An example test item is provided below:

She can’t eat dairy products s____ a____ milk and cheese. (*for example*)
 Answer: *such as*

The learners took two tests with 20 questions each, and they earned one point for each correct answer. Thus, the possible score range was 0–20. For a correct form knowledge assessment, all errors in spelling (e.g., *based* on: correct; *baced* on: incorrect) were marked incorrect even if they were inferable as LBs. Meanwhile, because the study’s focus was on LBs (vocabulary), grammatical (e.g., verb inflectional) errors were not penalized as long as they were the possible correct LB forms. For example, both *appear to* and *appears to* are correct depending on subject-verb agreements.

At the beginning of the intervention, the four strategies and prompts for ChatGPT were explained and modeled to ensure learners understood their employment. The learners were free to choose one or more of their preferred strategies.

Pilot Test

A pilot test was first conducted using 10 LBs whose normalized mean frequency was 9.47. They were considered to be easier than the actual A2 LBs (mean frequency 6.55). This was to familiarize the learners with the test format and assess whether they were able to apply the strategies and correctly record the study logs, with easier LBs as a starting point. The learners were given one week to study the 10 LBs for the pilot test and were also told to record their study logs in ChatGPT. All used the free ChatGPT model, which at the time of the experiment, initially ran on GPT-4o and automatically switched to 4-o mini when the usage cap was exceeded. As all learners appropriately recorded their study logs, the study proceeded to the main tests.

Actual Intervention

The learners studied 20 A2- and B1-level LBs each for two weeks. Learners submitted study logs via a URL on both test dates. The A2 LBs were tested on the first date (hereafter A2 test), followed by the B1 test one week later. The tests were administered without time limits. The learners were permitted to ask the author for the meanings of unknown words in the sentences during the test to ensure they could understand the meaning of the whole sentence and produce the appropriate LBs. Figure 2 outlines the whole process of the study, from the pilot test to the actual intervention.

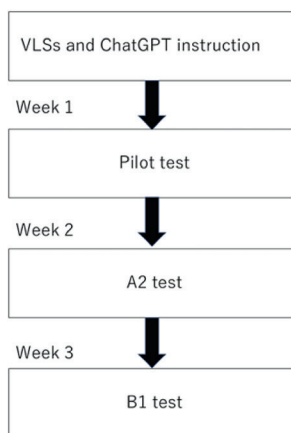


Figure 2 *The Whole Process of the Study.*

Statistical Analysis

To investigate RQ1 and RQ2 (effective learning strategies for the A2 and B2 LBs), hierarchical cluster analyses were conducted. The cluster analysis is a statistical method that can group learners and their learning behaviors into certain categories, and is often used to explore how VLSs affect learning outcomes (Alahmadi et al., 2018). This study used five variables to categorize learners and their VLS use on A2 and B1 tests. With the test scores as the first variable, four VLSs (questioning, gap-fill, grouping, and L1 translation) followed.

To quantify the four VLSs for the analysis, learners’ use of these VLSs was observed in their learning chat logs submitted via URL. The author manually checked whether the instructed prompts for the four VLSs appeared and how many times they were used in the learners’ chat logs. For instance, if learners used the Japanese order verb 教えて (teach me) for questioning (social) strategy and 穴埋め問題を作って (make gap-fill questions for me) for gap-fill practice (cognitive) strategy, these were counted as strategy uses. The counting used raw frequency. The count 0 means the strategy was never used, and 10 means the strategy was used 10 times. All of the variables, including test scores and raw frequency of VLS use, were standardized so that they were compatible on the same scale.

Hierarchical cluster analyses were performed for the A2 and B1 tests separately. The cluster analyses grouped A2 and B1 learners together. This was to observe widely applicable VLSs for A2 and B1 LBs, regardless of learners’ proficiency levels. The cluster analyses were performed using the Ward method with squared Euclidean distance.

All of the statistical analyses were conducted on langtest.jp (Mizumoto, 2015), a free multiple-application website.

Results

RQ1: Effective VLSs for A2 LBs

RQ1 investigated the effective VLSs for the A2 LBs. The term “effectiveness” in the RQs of the study refers to applicability (how widely VLSs are used by learners) and learning outcomes (test scores). The study conducted the hierarchical cluster analyses to classify learners by their VLS use. Table 4 provides the descriptive statistics for five variables (test scores and four VLSs) for the A2 test. The learners who both completed the tests and submission of learning logs (43 of 45) were subject to the analysis. The first row (*N*) lists the number of learners in each proficiency group, and the rows after the second show the mean, median, and standard error for the 20-point test scores. The rows questioning, gap-fill, grouping, and L1 translation provide the mean number of times the four VLSs were used.

Table 4 Descriptive Statistics of Variables in Cluster Analysis for A2 Test

Variables	Proficiency groups		
	A2 and B1	A2	B1
<i>N</i>	43	15	28
Mean	16.35	17.60	15.68
Median	18.00	19.00	17.50
SE	0.74	1.05	0.98
Questioning	4.93	5.73	4.50
Gap-fill	4.47	4.73	4.32
Grouping	1.00	1.47	0.75
L1 translation	1.19	0.60	1.50

In Table 5, the first row (N) lists the number of learners, and the second and third rows show the mean and standard error of the test scores in each cluster. Questioning to L1 translation rows show the mean number of times VLSs were used.

Table 5 *Descriptive Statistics for Each Cluster for A2 Test*

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
N	11	24	3	4	1
Mean score	18.55	17.00	18.67	3.75	20.00
SE	0.55	0.62	1.33	0.75	NA
Questioning	5.00	3.71	18.00	1.50	8.00
Gap-fill	2.82	3.67	21.67	2.00	0.00
Grouping	2.82	0.33	1.00	0.25	0.00
L1 translation	0.64	1.04	0.33	0.75	15.00

Cluster 2 in Table 5 demonstrates a remarkable feature of the learners' VLS use. The balanced combination of questioning (social: 3.71) and gap-fill (cognitive: 3.67) strategies can be considered the most widely applicable VLS to the learners regardless of their proficiency levels (A2 and B1), as they are most widely used among the five clusters ($N = 24$). Thus, this combined use of VLS is the most effective in terms of applicability. This supports the effectiveness of Schmitt's (1997) discovery and consolidation strategies in A2 LB learning.

From a view point of the most successful learning outcomes (mean test score: 18.67), cluster 3 reinforces the effectiveness of combined questioning and gap-fill strategies. The cluster shows the largest mean number of times the two strategies were used (questioning: 18.00, gap-fill: 21.67). The cluster also supplemented the two VLSs with grouping, whose mean number of times is the second-highest among the five clusters. This high degree of the learners' cognitive involvement (large number of times social and cognitive VLSs used and supplementary memory strategy) could have contributed to the highest mean test score of 18.67, which conforms to the effectiveness of Laufer and Hulstijn's (2003) Involvement Load Hypothesis. However, this heavier cognitive load (i.e., large number of times VLSs used) can be challenging and may not be preferred by the most of the learners, because cluster 3 had only three learners out of 43.

Cluster 1 also supports the importance of cognitive involvement. The cluster shows the balanced use of social, cognitive, and memory strategies. Cluster 1 equally uses gap-fill and questioning (mean number of times: 2.82). These VLSs also require deeper processing of the LBs compared to L1 translation, which provides superficial translations. Cluster 1 has the second-largest number of learners (11) and the second-highest test score (18.55) among the five clusters. This suggests that the VLS use tendency in the cluster has both applicability and direct effects on successful learning outcomes (test scores).

Clusters 1 and 3 also imply the significance of grouping the LBs into functional categories, as they both exhibit relatively higher grouping (1.00 or more) mean number

of times and test scores than clusters 2 and 4. This also shows that the theoretical LB categorizations can be put into LB learning practice.

Cluster 5 is exceptional among the clusters. While the other four clusters favor questioning, gap-fill, or grouping, cluster 5 (only 1 of 43 learners) preferred L1 translation but scored 20. To prevent this one exceptional cluster from affecting the overall clustering, this learner was categorized as cluster 5.

Cluster 4 shows that even B1 learners would not succeed on the A2 (i – 1) test without sufficient VLS use. Although they used all four VLS and had balanced use of questioning (1.50) and gap-fill (2.00), their overall mean number of VLS use is lower than all of the other clusters, which could have led to their lowest test scores (3.75).

Figure 3 depicts the five clusters for the A2 test. A2 and B1 learners were labelled with their CEFR levels (A2 (a)...B1 (c')). Each cluster is marked C (e.g., C1).

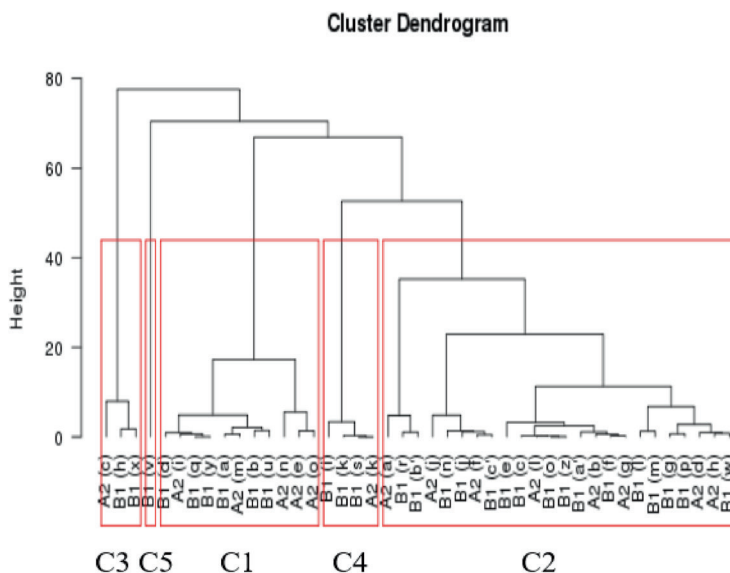


Figure 3 Clustered Learner Groups for A2 Test.

The “height” separates the clusters. The height in Figure 3 first separates cluster 3 around 80, as the cluster demonstrated exceptionally high mean values in questioning and gap-fill strategies, along with relatively frequent use of grouping (second-highest mean number of times). Its mean test score was also the highest. As a result, its overall mean values across all variables placed it furthest from the other clusters, leading to its early separation in the dendrogram. Following this, cluster 5, characterized by its extreme specialization in translation strategy, was also separated between 80 and 60, due to its distinctiveness and higher test scores. The remaining clusters (1, 2, and 4) shared more balanced mean values across strategy use (mean number of times each strategy is close compared to clusters 3 and 5) and test scores, and were therefore grouped more closely, except that cluster 1’s noticeable feature of grouping strategy

use still separates it from clusters 2 and 4. Although cluster 4 had significantly lower mean test scores (3.75), their similarities of relatively balanced VLS use of questioning (1.50) and gap-fill (2.00) placed it closely together with cluster 2. Key observations are summarized below:

- Combining questioning and gap-fill strategies is the most effective VLS in terms of applicability (cluster 2: $N = 24$), and learning outcomes (cluster 3: mean test score 18.67), regardless of learners' proficiency levels.
- Higher cognitive involvement associated with VLSs, including grouping, leads to better learning outcomes (mean test scores = cluster 1: 18.55).

RQ2: Changes in VLSs for B1 LBs from A2

RQ2 explored the changes in the learners' VLS use from A2 LBs. The study employed the hierarchical cluster analysis to group the learners by their VLS use. Table 6 provides the key descriptive statistics for the cluster analysis for the B1 test.

Table 6 *Descriptive Statistics of Variables in Cluster Analysis for B1 Test*

Variables	Proficiency groups		
	A2 and B1	A2	B1
<i>N</i>	43	15	28
Mean	16.81	17.67	16.36
Median	18.00	18.00	18.50
SE	0.65	0.67	0.93
Questioning	7.70	10.88	7.57
Gap-fill	3.16	5.06	3.11
Grouping	1.07	1.22	1.15
L1 translation	0.70	1.06	1.07

Table 7 shows the descriptive statistics for each cluster.

Table 7 *Descriptive Statistics for Each Cluster for B1 Test*

Variables	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
<i>N</i>	6	23	4	5	2	3
Test score	18.17	17.52	15.25	17.20	2.00	20.00
SE	0.48	0.58	2.25	0.86	1.00	0.00
Questioning	7.33	2.70	9.50	23.40	9.50	17.00
Gap-fill	0.83	2.22	2.00	2.40	2.00	18.67
Grouping	3.33	0.70	0.50	0.40	1.50	1.00
L1 translation	0.33	0.61	3.50	0.00	0.00	0.00

Figure 4 shows that the height separates cluster 6 at around 80, then cluster 3 at around 60. As in the A2 test, these clusters are distinctive in their strategy use and test scores, with the former using a large number of questioning and gap-fill (test score 20.00), the latter focusing on L1 translation (test score 15.25). Cluster 1 is different from clusters 2, 4, and 5 in that it favors grouping. Lastly, clusters 4 and 5 are closely grouped due to their similarity (not balanced questioning strategy use with the other strategies).

The above cluster examples suggest that, in B1 LB learning, it is crucial to combine questioning effectively with gap-fill or grouping, which could contribute to deeper processing of LBs. The followings are the summary of both the changes and commonalities observed in the B1 test cluster analysis compared to A2:

- The combined use of questioning and gap-fill strategies, along with grouping itself, remains effective (applicability and score outcomes) VLSs for B1 LB learning.
- The effectiveness of the questioning strategy depends largely on how it is combined with the other strategies.

Overall, the results for the A2 and B1 test cluster analysis (RQ1 and RQ2) support the effectiveness of the Involvement Load Hypothesis (Laufer & Hulstijn, 2001) in LB learning at both levels. This is evident in the cluster tendencies in both A2 and B1. The more learners engage in deeper cognitive processing (combined use of VLSs), the more effective vocabulary learning can become (i.e., applicable for a wider range of learner groups or higher score outcomes).

Discussion

RQ1: Effective VLSs for A2 LBs

RQ1 investigated which ChatGPT-mediated VLSs contributed to the A2 test performance.

The A2 cluster analysis revealed that combining questioning and gap-fill is the most effective. Figure 5 is an example of a gap-fill strategy that a B1 learner skillfully used by asking for hints and a random question order. The learner adjusted the cognitive load to make the learning easier (with hints) or more difficult (random order) for A2 LBs. Although the learners' prompts and ChatGPT's explanations were in Japanese, the author translated the original chat logs in all of the following Figures to ensure both anonymity and an accurate interpretation for English readers.

As mentioned earlier, the gap-fill strategy can be more effective when used with questioning. This conforms to the discovery and consolidation strategies in Schmitt's (1997) taxonomy. This combination is the most widely applicable (cluster 2: $N = 24$) strategy for learning A2 LBs. Remarkably, some learners with full marks (20/20) followed the same order as Schmitt's (1997) discovery (social) and consolidation (cognitive) strategy, as shown in Figure 6. The learner asked ChatGPT to first explain the meaning of the A2 LB *whether or not*, and provide gap-fill afterwards.

I want some random fill-in-the-blank questions with English hints.

Great! This time, I'll give you **five random fill-in-the-blank questions** based on the expressions we've learned so far.

Each one comes with an **English hint** to help you out!

【Random Fill-in-the-Blank Quiz (with Hints)】

1. She likes sports ____ swimming and tennis.
Hint: used to give examples (like "for example")
2. He'll go to the event ____ his friends go or not.
Hint: shows "no matter if something happens or not"
3. I stayed home ____ the heavy rain.
Hint: shows the reason or cause (similar to "because of")
4. ____ I think this is a good decision.
Hint: means "overall" or "when looking at the whole situation"
5. You should sleep ____ 7 hours a night.
Hint: means "no less than," shows a minimum amount

Figure 5 Tailored Gap-fill Strategy for the A2 LBs (e.g., 1: such as).

Explain the meaning of "whether or not" and provide a gap-fill exercise.

Ok! Here's an explanation of the meaning of "**whether or not**" along with a gap-fill exercise!

■ Meaning of "whether or not"

"whether or not" means:

"Regardless of whether something happens or not" or "no matter if something happens or not."

- **whether** = whether or not
- **whether or not** = even if it does or doesn't, in short: "either way"

Often used when expressing things like "do something regardless of the result."

■ Examples of Usage

- I'm going to the party **whether or not** it rains.
- She'll do it **whether or not** you help her.

■ Fill in the blanks

1. I'm going jogging ____ it's cold.
2. He'll join the trip ____ his friends go.



Figure 6 Meaning Explanations and Gap-Filling Exercises for the A2 LB "whether or not".

When the learner understands (discovers) the meaning of *whether or not*, the learner then can practice (consolidate) using the LB in actual sentences, which encouraged meaning-focused output. This could have reinforced the learner’s knowledge of the LB, even though it was below the learner’s proficiency level ($i - 1$). As the cluster analysis for the A2 test found the same strategy use pattern among A2 learners as well, combining the questioning and gap-fill can be considered equally effective regardless of the target LB levels ($i - 1$ for B1 learners; i for A2 learners).

Another effective use of questioning is to combine it with the grouping (memory) strategy. Learners first learn the meanings and functions of the LBs, and then organize them for better retention. Figure 7 provides an example of a ChatGPT-generated table for a syntactically similar (noun + of) A2 referential bundles.

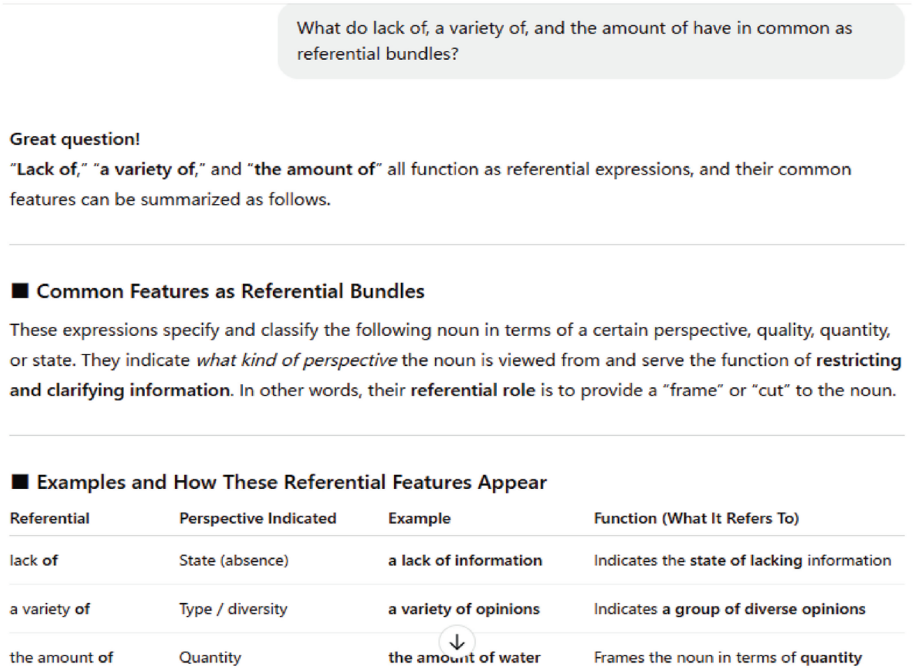


Figure 7 Meaning Elaborations and Table for A2 Referential Bundles.

As cluster 1 in the A2 test shows, combining questioning and grouping led to the second-highest learning outcomes (test score 18.55) and applicability ($N = 11$).

RQ2: VLS Use Changes in B1 LBs from A2

The cluster analysis found significant changes in effective VLS use. While the most effective VLSs remain the same from the A2 test (balanced or abundant use of questioning and gap-fill: clusters 2: 23 learners and 6: score of 20.00), some learners shifted to more analytical VLSs, with a focus on questioning. Cluster 1 (second-highest test

score) prioritizes questioning and grouping, and cluster 4 favors questioning (23.40). Cluster 1 implies that combining questioning and grouping continues to be important, and cluster 4 demonstrates that nearly exclusive use of questioning is almost as equally as effective, as cluster 4 has close mean test scores (17.20) to cluster 2 (17.52). As learners encounter more advanced B1 LBs, they start to ask ChatGPT for meaning elaborations more often. Figure 8 is ChatGPT's answer to a prompt by an A2 learner to check the understanding of the meaning of *degree of*. "A decrease in" is 減る in Japanese, so the learner might have confused "degree" with "decrease" as both share "de-".

Does "degree of" mean "a decrease in"?

👤 No! The true meaning of "degree of" is:

- "the level" or "extent of ~."

It does **not** mean "a decrease in ~"!

✅ Summary of the meaning of "degree of":

- degree means "level" or "extent."
- degree of ~
→ expresses the *extent*, *degree*, or *level* of something.

🌟 Check with examples:

- There is a high **degree of** risk.
- She showed a great **degree of** patience.
- Different people have different **degrees of** sensitivity.

Figure 8 *Meaning Correction of "degree of" for an A2 Learner.*

This interaction shows that ChatGPT assists the learner in reconstructing the initial incorrect understanding of *degree of*. Clarifications like this could have mediated the A2 learners' learning of the $i + 1$ (B1) LBs, characterizing their VLS use into the same cluster as B1 learners in cluster 4 in Figure 4 (increased number of questioning use).

Nevertheless, questioning should be accompanied by gap-fill or grouping for effective (applicability and score) outcomes to maximize the effectiveness of LB learning, as mentioned previously. Cluster 4 (inclined to questioning) has fewer learners and lower scores than clusters 1, 2, and 6.

Conclusion

This study explored how A2 and B1 Japanese university students learn the LBs set at their proficiency levels. Specifically, RQ1 focused on effective VLSs for the A2 LBs,

and RQ2 examined the changes in the VLSs for B1 LBs from A2. A three-week instructional intervention was conducted, during which 45 students learned 20 A2 and B1 LBs using ChatGPT-mediated VLSs. The learning outcomes were measured with gap-fill post-tests, and the use of four learning strategies: interactive questioning (social), gap-fill practice (cognitive), grouping (memory), and L1 translation (memory) was manually counted for the hierarchical cluster analysis to group the learners by their VLS use and test scores.

RQ1 found that combining social and cognitive strategies is the most effective for learning A2 LBs in terms of applicability and score outcomes. RQ2 indicated that the use of questioning (e.g., the number of times used or combining it with gap-fill or grouping) largely affects the learning of B1 LBs, as observed in the cluster changes. Also, learners continued to favor Schmitt's (1997) discovery and consolidation (questioning and gap-fill) strategies.

The results for RQ1 and RQ2 support Laufer and Hulstijn's (2001) Involvement Load Hypothesis, which argues that deeper cognitive processing is necessary for effective vocabulary learning.

Despite the above findings, this study has limitations. First, while the relatively small analyzed sample size ($N = 43$) found meaningful clusters and learning tendencies common to both A2 and B1 learners, the sample drawn from one site limits generalizations (e.g., Vitta et al., 2022). Future studies should replicate this research with a larger and more diverse population. Second, the strategy use was manually coded, which was labor-intensive and limited scalability. Developing semi-automated coding systems would reduce this effort and allow for larger-scale studies.

Nevertheless, the present study has several notable strengths. From a theoretical perspective, it is one of the first to apply a strategy-based method to LB learning. The learning logs directly recorded the learners' VLS use for the LBs at different difficulty levels ($i - 1, i, i + 1$), which was never attempted in the previous studies. The results indicated that deeper processing is important for LBs, regardless of difficulties. From a practical standpoint, the study provides some pedagogical guidance for instructors teaching A2–B1 learners. For example, teachers can guide students to engage in deeper processing (i.e., questioning, gap-fill, and grouping) for more complex LBs, while also encouraging them to balance strategies.

In conclusion, this study provides both theoretical and practical contributions to the underexplored field of EAP LB learning with AI-assisted VLSs.

Acknowledgments

This study was supported by JST SPRING (grant number JPMJSP2150).

References

- Alahmadi, A., Shank, C., & Foltz, A. (2018). Vocabulary learning strategies and vocabulary size: Insights from educational level and learner styles. *Vocabulary Learning and Instruction*, 7(1), 14–34. <https://doi.org/10.7820/vli.v07.1.alahmadi>
- Biber, D., Conrad, S., & Cortes, V. (2004). *If you look at ...*: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3), 371–405. <https://doi.org/10.1093/applin/25.3.371>

- Chiew, M. T. L., & Ismail, H. H. (2021). Exploring vocabulary learning strategies in a second language setting: A review. *International Journal of Academic Research in Business and Social Sciences*, 11(12), 1298–1309. <https://doi.org/10.6007/IJARBS/v11-i12/11376>
- Cortes, V. (2004). Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes*, 23(4), 397–423. <https://doi.org/10.1016/j.esp.2003.12.001>
- Council of Europe. (2020). *The common European framework of reference for languages: Learning, teaching, assessment*. Council of Europe Publishing.
- Davies, M. (2008) *The Corpus of Contemporary American English (COCA)*. <https://www.english-corpora.org/coca/>
- Durrant, P. (2019). Formulaic language in English for academic purposes. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 211–227). Routledge.
- Granger, S. (2019). Formulaic sequences in learner corpora: Collocations and lexical bundles. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 228–245). Routledge. <http://dx.doi.org/10.4324/9781315206615-13>
- Ishikawa, S. (2019). *English N-gram List for Japanese Learners of English (ENL-J)*. <https://language.sakura.ne.jp/s/voc.html>.
- Jones, M., & Haywood, S. (2004). Facilitating the acquisition of formulaic sequences: An exploratory study in an EAP context. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 269–292). John Benjamins Publishing Company.
- Keshmirshekan, M. H. (2019). Krashen's input hypothesis, reading comprehension and reading motivation: $i+1$ versus $i-1$. *Language in India*, 19(5), 10–26.
- Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon.
- Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: The construct of task-induced involvement. *Applied Linguistics*, 22(1), 1–26. <https://doi.org/10.1093/applin/22.1.1>
- Mizumoto, A. (2010). *Exploring the art of vocabulary learning strategies*. Kinseido
- Mizumoto, A. (2015). *langtest.jp* (Version 1.0) [Web application]. <https://langtest.jp/>
- Ngo, T. (2024). The use of ChatGPT for vocabulary acquisition: A literature review. *SSRN Electronic Journal*. <http://dx.doi.org/10.2139/ssrn.5059052>
- Nicklin, C., Bailey, D., McLean, S., Kim, Y. A., Kang, H., & Vitta, J. P. (2025). Applying lexical sophistication models to wordlist development: A proof-of-concept study. *Research Methods in Applied Linguistics*, 4(1), 1–18. <https://doi.org/10.1016/j.rmal.2024.100175>
- Oxford, R. (1990). *Language learning strategies: What every teacher should know*. Newbury House.
- Paton, S. (2020). Introducing mnemonics to Japanese students as a vocabulary learning strategy. *Vocabulary Learning and Instruction*, 9(1), 80–93. <https://doi.org/10.7820/vli.v09.1.paton>
- Sawaguchi, R. (2025a). Developing the preliminary essay bundles list (EBL) and its applicability to EAP. *English Corpus Studies*, 32, 21–40. https://doi.org/10.69193/ecs.32.0_21

- Sawaguchi, R. (2025b). Developing a CEFR-based diagnostic test to assess Japanese university students' productive knowledge of lexical bundles. *Language Testing in Asia*, 15(27), 1–23. <https://doi.org/10.1186/s40468-025-00361-0>
- Schmitt, N. (1997). Vocabulary learning strategies. In N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 199–277). Cambridge University Press.
- Shamsabadi, R., Ketabi, S., & Eslami Rasekh, A. (2017). Developing Iranian EAP students' writing skill through explicit instruction of lexical bundles. *Journal of English Language Teaching and Learning*, 9(19), 25–52.
- Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, 31(12), 487–512. <http://dx.doi.org/10.1093/applin/amp058>
- Sonbul, S., & Macis, M. (2021). The deliberate learning of L2-L1 wordlists: Effectiveness and learners' perceptions. In P. Szudarski & S. Barclay (Eds.), *Vocabulary theory, patterning and teaching* (pp. 89–98). Multilingual Matters.
- The Institute for International Business Communication. (2025a). *TOEIC program data & analysis 2025*. <https://www.iibc-global.org/hubfs/library/default/iibc/press/2025/p284/2025DAA.pdf?hsLang=ja>
- The Institute for International Business Communication. (2025b). *The guideline of TOEIC scores based on CEFR benchmark*. https://www.iibc-global.org/toEIC/official_data.html
- Ueno, S., & Takeuchi, O. (2023). Which vocabulary learning strategies are important and useful for Japanese university students? A text-mining approach. *Vocabulary Learning and Instruction*, 12(1), 18–30. <https://doi.org/10.29140/vli.v12n1.1213>
- Vitta, J. P., Nicklin, C., & McLean, S. (2022). Effect size-driven sample-size planning, randomization, and multisite use in L2 instructed vocabulary acquisition experimental samples. *Studies in Second Language Acquisition*, 44(5), 1424–1448. <https://doi.org/10.1017/S0272263121000541>
- Wingate, U. (2012). 'Argument!' helping students understand what essay writing is about. *Journal of English for Academic Purposes*, 11(2), 145–154. <https://doi.org/10.1016/j.jeap.2011.11.001>

Appendix

Instructional 40 Lexical Bundles (LBs) with CEFR Levels

A2 LBs	B1 LBs
1. such as	1. in terms of
2. based on	2. the presence of
3. the fact that	3. a great deal of
4. at the same time	4. take advantage of
5. long-term	5. degree of
6. lack of	6. with respect to
7. a variety of	7. associated with
8. the amount of	8. the relationship between
9. aspect of	9. suggest that
10. the concept of	10. note that
11. argue that	11. no way
12. ensure that	12. be the case
13. it is clear that	13. even more
14. at least	14. take into account
15. as a whole	15. to begin with
16. appear to	16. the point is
17. as well as	17. in turn
18. due to	18. vice versa
19. as a result	19. on the contrary
20. whether or not	20. on the one hand