PASAA Journal
Volume 68, January – June 2024, 203–228

PASAA

Lexical Level in English Major EFL Students' Writing: A Learner Corpus Study

Pong-ampai Kongcharoen, Jiraporn Dhanarattigannon, and Tirote Thongnuan*
Department of Foreign Languages, Kasetsart University, Bangkok, Thailand
**Corresponding author: tirote.t@ku.th*

Article information

Abstract

This study aimed to investigate the lexical competence of English-major EFL students. The learner corpus comprised 552 pieces of writing by sophomore English majors during five academic years between 2017 and 2021, containing 190,506 words in total. The results from VocabProfile program showed that these students used words contained in the academic word list (AWL) at lower rates than what has been suggested by experts. This is probably because they were in their second year studying their first writing course. With less exposure to higher-level English, they might not have developed an advanced vocabulary repertoire. However, when considering the AWL words used in each type of paragraph, it was found that the students used AWL words to a higher degree in comparison-contrast, cause-effect, and problem-solution paragraphs, suggesting that the type of paragraphs can affect the lexical level of the students. When considering the British National Corpus/Corpus of Contemporary American English (BNC/COCA) lexical level, the words used were mostly in Base List 1, Base List 2, and Base List 3, respectively. The results from VocabProfile program indicated that they used vocabulary at a very typical level compared to non-English majors, suggesting that they needed more input to stimulate them to use higher vocabulary levels in other advanced writing courses to attain an effective vocabulary level upon completion of the program of studies.

E-ISSN: 2287-0024

Keywords	learner corpus, lexical level, students' lexical competence
APA citation:	Kongcharoen, P., Dhanarattigannon, J., & Thongnuan, T. (2024). Lexical level in English major EFL students' writing: A learner corpus study. *PASAA. 68.* 203–228.

1. Introduction

Previous studies (Abduh & Rosmaladewi, 2017; Crosthwaite, 2016; Huang et al., 2010; Tiliakou & Frantzi, 2021) have been conducted on learners' corpus because it is believed that vocabulary size of EFL learners can enable them to achieve a higher level of English proficiency and thus improve their English language performance. Abduh and Rosmaladewi (2017) provided several reasons why it is essential for both teachers and learners to study lexical frequency (word lists):

> First, it helps educators in vocabulary teaching, and it is necessary to establish what vocabulary means to focus on teaching it. Second, it provides a useful academic word pool for non-native English learners who need to read and publish articles in English. Third, it assists learners in dramatically enhancing reading power for a relatively modest learning investment (p. 283).

According to Abduh and Rosmaladewi (2017), the English vocabulary repertoire of ESL and EFL students is essential to enhance their English proficiency. Schmitt et al. (2001) have analyzed the Vocabulary Level tests and suggested that the tests serve to measure general or academic vocabulary size of second language (L2) learners of English. The estimation of vocabulary size of learners at different frequency levels has also been studied. For instance, Hsueh-chao and Nation (2000) have reported that knowledge of 98%-99% of the lexical items in a written text is required to avoid comprehension problems caused by new words. Using statistics derived from the British National Corpus (BNC), Nation (2017); Hsueh-chao and Nation (2000); and Schmitt et al. (2001) suggest that EFL students need to master 8,000 to 9,000 word families to reach 98% text coverage and comprehend English texts. Laufer and Ravenhorst-Kalovski (2010) has revisited the lexical threshold for "adequate reading comprehension" (p. 15) and suggested an optimal threshold of 8,000 word families and a minimum threshold of 4,000-5,000 word families, yielding 98% and 95% coverage, respectively. In addition, Alfatle (2016); Sun et al. (2010); and Mungkonwong and Wudthayagorn (2017) suggest that size of vocabulary and years of study are related, and the more students are exposed to English, the larger the vocabulary size they will build.

In the Thai context, a number of studies have been conducted on vocabulary size of university students (Mungkonwong & Wudthayagorn, 2017; Nirattisai & Chiramanee, 2014; Pringprom & Obchuae, 2011; Wiriyakarun, 2018; Zhiying, 2007) mainly to figure out how many words Thai university students knew, while other studies examined vocabulary size in other dimensions. For example, Pringprom and Obchuae (2011) explored the relationship between vocabulary size and reading comprehension. Zhiying (2007), on the other hand, studied the relationship between passive recognition vocabulary knowledge, active recall vocabulary knowledge, and free active written vocabulary knowledge. Nirattisai and Chiramanee (2014) looked into the relationship between vocabulary size and vocabulary learning strategies. Moreover, two of the most recent studies on this topic were conducted by Wiriyakarun (2018) who examined Thai EFL learners' knowledge of academic English vocabulary using the academic vocabulary test and by Mungkonwong and Wudthayagorn (2017) who investigated Thai university freshmen's vocabulary size related to years of English study using the Bilingual English-Thai version of the Vocabulary Size Test (VST).

It is generally assumed that students majoring in English possess a larger vocabulary size than non-English majors at a sufficient level to ensure lexical competence, and their academic vocabulary size is necessary for them to develop quality English writing products. However, it is noteworthy that several studies that investigated the students' size of vocabulary in Thai contexts (Mungkonwong & Wudthayagorn, 2017, for instance) rarely focused on Thai students majoring in

English or used the learner corpus of their writing assignments which are considered authentic texts as an instrument. In fact, students' writing assignments can represent their actual English proficiency and their use of the English language. Therefore, this study aimed to explore the vocabulary mastery of students majoring in English (EFL) based on their writing assignments by applying the VocabProfile program for data analysis to investigate the lexical level in their writing based on Academic Word List (AWL) and the British National Corpus/Corpus of Contemporary American English (BNC/COCA) frequency word lists. The aim of the study led to the following research questions:

1. How many AWL words are used in the learner corpus compared to the reference corpus, namely the corpus of three well-known Scopus-indexed journals in Thailand?
2. What is the vocabulary level of English major students compared to the reference corpus?

The results of the present study would yield empirical evidence of English major students' level of lexical competence so that writing courses in the Bachelor of Arts Program in English could be properly designed to more effectively help students develop lexical competence. Moreover, based on the study findings, teachers should be able to construct instructional materials to more appropriately enhance students' writing abilities, particularly lexical competence—vocabulary—in order to help them reach the advanced vocabulary level required for further studies since writing proficiency is one of the key elements to ensure students' academic success (Gallagher, 2006).

2. Literature Review

2.1 Lexical Levels

2.1.1 Academic Word List (AWL)

Coxhead (2000) developed the Academic Word List (AWL), comprising 570 word families commonly found in academic texts. Coxhead (2000) created the AWL by exploring the corpus of about 3.5 million words in four academic fields, including Arts, Commerce, Law, and Science. Each field is divided into seven subject areas and contains approximately 875,000 running words.

To extract the AWL, Coxhead (2000) initially screened the General Service List (GSL) created by West (1953). After that, the criteria of frequency and dispersion were employed. Any words that appeared more than 100 times in the whole corpus and ten times in each sub-corpus were included in the list. After the final session, the AWL comprises 570 word families and has widely been used as a useful list to help students' accomplish academic purposes. The AWL accounts for about 10% of words in academic texts. When combined with words in the GSL, which covers approximately 80% of words in written texts, it would encompass around 90% of words in texts.

These AWL words are not associated with any specific subject, making them valuable for all students. The 570 word families in the AWL are categorized into ten sublists based on their frequency. Sublist 1 includes the most frequent word families, followed by sublist 2 with the next most frequent, and so on. Each sublist consists of 60 word families, except for sublist 10, which comprises only 30 word families. In this study, AWL was used as one indicator to identify the English major students' vocabulary level which reflected their vocabulary repertoire.

2.1.2 BNC/COCA Vocabulary List

Nation (2016) created the BNC/COCA (British National Corpus/Corpus of Contemporary American English) word family lists, which are sets of lists categorized by frequency level in each of the 1000 word families. The word families in the BNC/COCA lists span from the first 1000 to the 25th 1000 words in English. These lists were specifically designed to assist English language learners with a focus on supporting those learning English as a foreign language. The higher frequency lists such as the first 1000 and second 1000 include vocabulary relevant to foreign travel, studying in English, and Internet usage.

The BNC/COCA lists were generated through two distinct methods. The first two lists are derived from a ten million-word corpus. This corpus consists of six million words from spoken British and American English, including content from films and TV programs. The additional four million words are from written British and American English. This approach is adopted to prevent the first two lists from being overly influenced by the written corpus used for subsequent lists. This ensures the inclusion of very common spoken words like 'pardon,' 'hello,' 'dad,' and 'bye.' Moreover, it encompasses word sets such as numbers, days of the week, and months of the year. Additionally, some essential vocabulary for foreign travel, terms like 'survival' are included. The third 1000 lists onward are generated using rankings in the BNC (University of Oxford, 2015) and COCA (Davies, 2020), after removing the 1k and 2k words.

2.2 Learner Corpus

The learner corpus was developed from corpus linguistics. Therefore, it shares some common features of corpus linguistics. For example, both are used as tools to analyze languages. McEnery et al. (2006) define the term "corpus" as a "collection of machine-readable authentic texts (including transcripts of spoken data) which is sampled to be representative of a particular language or language variety" (p. 5, as cited in Meunier, 2021, p. 23). The learner corpus has subsequently been developed and defined based on the concept of corpus linguistics.

Granger (2008) has defined learner corpora in comparison with corpora in that "[l]earner corpora have all the characteristics commonly attributed to corpora, the only difference being that the data come from language learners" (p. 259). She also roughly defines it as "electronic collections of texts produced by language learners" (p. 259). Based on Granger's definition of the learner corpus, language learners are foreign language learners, which subsequently applies to L2 and EFL learners.

Meunier (2021) defines a learner corpus as "a specific type of corpus which, to follow up on McEnery et al.'s definition, can broadly be defined as a collection of machine-readable texts consisting in representative samples of the language written and/or spoken by learners of an additional language (viz. not their mother tongue, but a foreign/second/nth target language)" (p. 23).

In this study, the learner corpus refers to a collection of authentic texts, particularly written texts, produced by EFL learners (Thai students). The learner corpus used in the present study was a representative sample of the language written by English major students in a Thai context. According to Meunier (2021), learner corpora are a useful input for applied research projects because of their main features. They are considered authentic aspects of the language produced by learners and they are selected based on certain criteria which include learners, types of written texts, and contextual conditions of task setting.

2.3 Previous Related Studies

Since learner corpus analysis emerged, several research studies have been conducted based on learner corpus, particularly in the fields of second language acquisition (SLA) and foreign language teaching. Based on the *Corpus Learner Bibliography* (Granger, 2009), studies on the learner corpus can be grouped into three main areas:

1. A corpus-based analysis of output linguistics in written and spoken learner texts, e.g., Agerström (2000), Aijmer (2002), Abe (2003), Aktas (2005), Benso (2000), Eriksson (2008), etc.
2. Learner corpus analysis and the development of foreign language proficiency (data-driven learning approach) for L2 teaching and learning, e.g., Axelsson (2000), Allan, (2002), Axelsson and Berglund (2002), Belz (2004), etc.

3. Creating English corpus data produced by EFL learners, which are useful for the SLA research on the development of learners' English language proficiency, such as Bączkowska (2000).

The focus of this research was on a learner corpus analysis of output linguistics in written texts produced by Thai students majoring in English. Therefore, the following sections address previous related studies in the first area, particularly in EFL and Thai contexts.

2.3.1 Studies on Learner Corpus Analysis of Output Linguistics in Written Discourse of EFL Learners

Granger (2008) reviewed studies on a learner corpus analysis and summarized the two methods mostly used for the learner corpus analysis as "contrastive interlanguage analysis and computer-aided error analysis" (p. 265) to understand how learners acquire the second language and to apply for EFL learning and teaching. In this section, studies on learner corpus analysis in SLA and EFL teaching and learning undertaken in the past decade will be addressed.

Many studies used a learner corpus to examine some linguistics features of L2 or EFL learners using a native speaker (NS) corpus as a benchmark to understand how they acquire and produce English in spoken and written discourses and use the information for EFL teaching (Shirato & Stapleton, 2007; Gilquin et al., 2008; Granger et al., 2015). These studies focused on using a learner corpus of written assignments to study vocabulary produced by Thai university students majoring in English, compared with an NS corpus, to explore the frequency of the English vocabulary they employed and how they used such vocabulary as well as to investigate cause/effect verbs and verb phrases in English.

2.3.2 Related Studies on Vocabulary Employed by EFL Students and Thai Students

Many empirical studies have been conducted, such as Qilichevna (2020), Ma and Mei (2021), Pu (2018), Chanchanglek and Sriussadaporn (2011), and Liangpanit (2010) on corpus linguistics and learner corpus in terms of corpus-based approach in vocabulary teaching to help EFL students improve their vocabulary proficiency, while few studies have been conducted on learner corpus to explore how EFL students employed English vocabulary in their spoken and written discourses.

Among them is a study by Shirato and Stapleton (2007), who investigated the nature of vocabulary in informal conversations used by Japanese adults, compared with the English NS corpus. The data comprised the spoken learner corpus. The participants were 117 Japanese-native speakers ranging in age from 17 to 74 years old, from non-English major college students to Japanese adults who engaged in five EFL classrooms. Data were collected through 1) face-to-face conversations among small groups (up to nine participants at a time), and 2) informal interviews of some participants. Their findings revealed that the Japanese adults in this study tended to underuse the lexical items representing interactive functions such as modal words, discourse markers, and hedges. The findings also showed that they tended to overuse "some high frequency of auxiliary verbs and some common adjectives" (p. 393). They suggested that the study of vocabulary used by EFL learners by using learner corpus helps teachers know their students' acquired vocabulary so that they can use this information to provide them with suitable lexical tools that will help them produce English texts close to the native norm.

In the Thai context, Veerachaisantikul and Chootarut (2016) and Wiriyakarun (2018) conducted a learner corpus study to investigate the academic English vocabulary of Thai EFL learners. Wiriyakarun (2018) studied the relationship between academic English vocabulary knowledge and English reading proficiency. She created “Academic Vocabulary Tests” based on Coxhead’s AWL to evaluate Thai EFL students’ receptive and productive academic English vocabulary. The participants of this study comprised 53 Thai undergraduate students studying Engineering, Science, and Industrial Education at a public university. The findings revealed that there was a moderately positive relationship between their receptive and productive academic English vocabulary. Interestingly, the findings showed that the students’ first ten receptive and productive academic words were from different sublists of the AWL. She explained that the students may acquire those words from other sources such as from classroom learning. Moreover, the most recognized words that the students produced in their test were from AWL’s sublist 2, not sublist 1 as they should. In her conclusion, although the scores on receptive and productive academic words on the test were a little different, the Thai EFL students in this study seemed to “perform better on the productive test than the receptive test” (p. 129) and they tended to know more high-frequency words on the AWL’s word list. The results also revealed the relationship between the students’ knowledge of academic vocabulary and their success in learning English, and it was found that an increase in the number of word lists seemed to relate to their learning experience. Therefore, it was concluded that Thai university students should be provided with academic vocabulary, and the AWL is one of the most useful resources.

Veerachaisantikul and Chootarut (2016) studied the general vocabulary frequently used by Thai EFL tertiary students in their English writing, compared with the New General Service List (NGSL). They created the learner corpus (TEFL corpus) from 1,233 students’ writing tasks (661,596 words) using “WordSmith Tool Version 6”. The results revealed that the 50 most frequently produced words in the TEFL corpus were similar to general words in the NGSL. To illustrate, high-frequency words employed by the students in this study were “article, pronouns, and the verb to be” (p. 55), and the top five words were “be,” “the,” “and,” “a,” and “I” (p. 55). They emphasized the significance of students' vocabulary knowledge in writing and suggested that the teachers of EFL learners should provide their students with useful vocabulary resources such as concordance software to enhance their vocabulary development.

Additionally, Mungkonwong and Wudthayagorn (2017) studied vocabulary size of Thai university freshmen and the relationship between their vocabulary size and years of study. They used VST as an instrument, and the results revealed that the vocabulary size of Thai university freshmen was about 4,200 words, which was considered sufficient for the basic use of the English language. They also found that the vocabulary size was significantly related to each student's years of study.

To date, empirical studies using a learner corpus to investigate Thai EFL university students, particularly English-major students, are rare. Veerachaisantikul and Chootarut (2016) have reported that knowledge of vocabulary is necessarily a fundamental element for EFL learners to succeed in language learning, and it is one of many elements used to determine each student's level of English proficiency. They have concluded, "[w]ithout adequate vocabulary students cannot comprehend others or express their ideas" (p. 52). Therefore, by using the learner corpus of Thai EFL writing assignments, this study aimed to explore the academic vocabulary knowledge of Thai English-major students at a university, particularly their vocabulary level, to gain more insight into the vocabulary repertoire of Thai students majoring in English and to determine their level of English proficiency.

3. Methodology

3.1 Setting

The data collected in this research were authentic writing assignments written by English major students at a public university in Thailand. These students were sophomores enrolled in their first English Writing course. The writing assignments comprised listing, sequence, comparison-contrast, cause-effect, and problem-solution organizational patterns.

3.2 Data Collection

The learner corpus for this study consisted of four types of writing assignments for the English Writing course each academic year. They were collected from five academic years between 2017 and 2021. The data for each academic year are shown below.

2017: 38 students; four assignments (listing, sequence, comparison and contrast, and cause-effect); 152 total assignments (50,496 words)

2018: 18 students; four assignments (listing, sequence, comparison and contrast, and cause-effect); 72 total assignments (21,470 words)

2019: 25 students; four assignments (listing, sequence, comparison and contrast, and cause-effect); 100 total assignments (30,827 words)

2020: 22 students; four assignments (listing, sequence, comparison and contrast, and cause-effect); 88 total assignments (27,938 words)

2021: 35 students; four assignments (listing, comparison and contrast, cause-effect, and problem-solution); 140 total assignments (59,760 words) (The sequence organizational pattern was excluded and replaced with the problem-solution one because 1) the curriculum was revised, resulting in changes in the course content and 2) problem-solution was considered more essential in developing students' analytical and argumentative skills. However, sequence was included in Introduction to English Reading and Writing Skills to develop their writing skills at a paragraph level.)

There were a total of 552 assignments with a total of 190,506 words of the learner corpus.

This research was approved by the Institutional Review Board (IRB), Kasetsart University Research Ethics Committee, on November 21, 2022 (COE No. COE65/147).

3.3 Reference Corpus

In order to ensure that the Bachelor of Arts Program in English heads in the right direction, the reference corpus was created. This corpus consists of three well-known Scopus-indexed journals in Thailand. The journals were selected and compiled as a reference corpus because they are highly regarded by graduate students and scholars in the field of language teaching and learning in Thailand. Due to the limited space in the VocabProfile program, the data taken from each journal were only from one issue in the year 2023. This comparison aimed to further forecast how much more the students in the program needed to develop their lexical competence. Therefore, the results in this research from the learner corpus and the reference corpus were compared for the purpose of further forecasting how far and in which direction the students in the program needed to develop to achieve lexical competence.

3.4 Data Analysis

The data were analyzed using the VocabProfile program to identify the level of Academic Word List (AWL) and BNC/COCA vocabulary list. The VocabProfile program was created by Paul Nation. The program provides the screening of AWL with 570 word families and 25 vocabulary base lists. Each base list comprises the most 1000 words from BNC/COCA. The program is available for free at https://www.lextutor.ca/vp/eng/.

4. Results and Discussion

This section presents and discusses the results of the study on the students' use of AWL and their vocabulary level in the learner corpus, compared to the reference corpus.

4.1 Results of AWL Vocabulary from the Learner Corpus

Table 1 belows reveals that students majoring in English in this study used the AWL at quite a low level in all organizational patterns in 2017, with the highest of only 3.74% in a cause-effect organizational pattern. In the year 2018, the highest percentage of the AWL used by students rose to 5.12% in a compare-contrast organizational pattern, while the lowest was in a sequence organizational pattern. In the years 2019 and 2020, the highest percentages of AWL used by students were 5.60% and 5.20%, respectively, in a cause-effect organizational pattern. However, the highest percentage of AWL used by the students shifted to a problem-solution organizational pattern in the year 2021 at 5.35%.

Table 1

Total AWL Vocabulary Found in Students' Writing Assignments Each Academic Year

2017		**Families**	**Types**	**Tokens**	**Percent of Tokens**
Listing	AWL Words:	125	158	329	3.03%
Sequence	AWL Words:	116	142	261	2.30%
Compare-contrast	AWL Words:	134	180	455	3.08%
Cause-effect	AWL Words:	150	212	505	3.74%
					Percent
2018		**Families**	**Types**	**Tokens**	**Of Tokens**
Listing	AWL Words:	55	67	123	3.39%
Sequence	AWL Words:	62	68	92	2.48%
Compare-contrast	AWL Words:	107	139	366	5.12%
Cause-effect	AWL Words:	97	121	222	3.71%
2019		**Families**	**Types**	**Tokens**	**Percent of Tokens**
Listing	AWL Words:	96	124	199	3.38%
Sequence	AWL Words:	90	108	173	2.82%
Compare-contrast	AWL Words:	162	208	444	4.77%
Cause-effect	AWL Words:	156	223	532	5.60%

2020		**Families**	**Types**	**Tokens**	**Percent of Tokens**
Listing	AWL Words:	89	105	162	3.31%
Sequence	AWL Words:	75	88	125	2.61%
Compare-contrast	AWL Words:	157	218	412	4.46%
Cause-effect	AWL Words:	181	241	468	5.20%
2021		**Families**	**Types**	**Tokens**	**Percent of Tokens**
Listing	AWL Words:	201	277	572	3.88%
Compare-contrast	AWL Words:	196	291	649	4.29%
Cause-effect	AWL Words:	212	325	733	4.93%
Problem-solution	AWL Words:	228	327	803	5.35%

On the overall, the students seemed to use AWL the least in a sequence organizational pattern almost every year, which was about 2%, lower than what has been suggested by experts (Coxhead, 2000), while using the highest AWL in a cause-effect organizational pattern. This is probably because the cause-effect organizational pattern requires a higher level of thinking and thus affects topic choice. To illustrate, when the students write a cause-effect essay, they need to find concrete evidence to support or explain their claims so as to make their writing more convincing and logical. To do this, they have to carefully choose a topic that contains sufficient and reliable information from other sources. This type of reading may help them gain a higher level of vocabulary for their writing. The results seem to be compatible with those of Veerachaisantikul and Chootarut (2016) in that Thai university students tended to use general academic words, and they acquired a higher level of vocabulary from other reading sources.

Interestingly, from the academic years between 2019 and 2021, the students seemed to have a higher level of AWL. This is probably because they were screened through TCAS (Thai University Central Admission System). The Bachelor of Arts Program in English required a minimum ONET score of 50% or higher in English, which was higher than the requirement in previous years. They also had to participate in the Freshmen Preparation Program, where they were required to study Online English to obtain at least a B1 level.

The data from the reference corpus showed that the AWL used in these three journals is at the percentage level of about 10% coverage, which was suggested by experts. The highest percentage of AWL coverage was in Journal 1 (11.03%) with Journal 2 also possessing almost equivalent coverage of 11.02%, while the AWL coverage in Journal 3 was 9.30%, which was also considered high. See Table 2.

Table 2

Total AWL Vocabulary Found in the Reference Corpus

Journals	Families	Types	Tokens	Percent	All tokens
Journal 1	391	832	4329	11.03%	39254
Journal 2	424	970	5814	11.02%	52765
Journal 3	369	737	2763	9.30%	29702

The data shown in Table 1 and Table 2 seem to suggest that, in terms of AWL, the students in this study used a very low AWL in the learner corpus. They need to expose themselves more to AWL to develop their lexical competence, and the teacher should provide them with more useful and higher-level vocabulary.

4.2 Results of BNC/COCA List from the Learner Corpus

It was found in this study that the coverage of vocabulary used in students' writing varied. In every organizational pattern, students used only the first 3,000 frequent words from the BNC/COCA list to gain 95% coverage. At a coverage of 98%, the number increased to 4,000 or 5,000 frequent words from the BNC/COCA list. In the year 2020, they used up to 6,000 frequent words in a cause-effect organizational pattern. This aligned with the findings by Veerachaisantikul and Chootarut (2016), who conducted research with engineering students; the 50 most frequently produced words of their learner corpus were similar to general words in the NGSL. The high-frequency words employed by the students in their study were "article, pronouns, and the verb to be" (p. 55). This suggested that both English-major students and non-English major students used more general words in their work.

Table 3

Total BNC/COCA Vocabulary Found in the Students' Writing Assignments Each Academic Year

Academic Year	Listing		Sequence		Compare-contrast		Cause-effect		Problem-solution	
	95%	**98%**	**95%**	**98%**	**95%**	**98%**	**95%**	**98%**	**95%**	**98%**
2017	3000	5000	3000	5000	3000	4000	3000	4000	N/A	N/A
2018	3000	5000	3000	5000	3000	4000	3000	4000	N/A	N/A
2019	3000	4000	3000	4000	3000	5000	3000	5000	N/A	N/A
2020	3000	4000	2000	3000	3000	5000	3000	6000	N/A	N/A
2021	3000	5000	N/A	N/A	3000	4000	3000	5000	3000	5000

The results from BNC/COCA also suggested that the students tended to have general academic words at the first 3,000-word level (95%). This is probably because this was their first English writing course and they were in their second year, so they were infrequently exposed to higher levels of English. As explained by Alfatle (2016), Sun et al. (2010), and Mungkonwong and Wudthayagorn (2017), the size of vocabulary and years of study are related in that the students' vocabulary size will increase when they spend more time learning English.

After taking a closer look, it was found that the types of organizational patterns seemed to affect the word level the students used. When comparing the data for each academic year, the students reached 98% coverage by using the first 4,000 words for compare-contrast and cause-effect organizational patterns in the academic years 2017 and 2018. For the academic years 2019 and 2021, the students reached 98% coverage by using the first 5,000 words for the cause-effect organizational pattern. Remarkably, the students in the academic year 2020 gained 98% by using the first 6,000 words for the cause-effect organizational pattern. This may have been because those organizational patterns (compare-contrast, cause-effect, and problem-solution) required students to not only apply logical thinking to their writing, which is considered a high level of cognitive skills, but also provide solid evidence to support their claims. To do this, they needed to read more from other reliable sources, which led to gaining a higher level of vocabulary. These patterns also affected their topic choice. The topics that they chose were not only based on their own experience but also on academic or scientific grounds.

Still, the data from BNC/COCA seemed to fluctuate. One possible explanation is that the criteria for university admissions change every year. The Bachelor of Arts Program in English needs to adjust the criteria according to university requirements. The recruitment criteria each year affect the quality of the recruited students. Simply put, the program cannot recruit students with the same level of English proficiency every year.

In terms of the vocabulary used from the BNC/COCA list, the results from the reference corpus revealed that the coverage of 95% in Journal 1 and Journal 3 fell into the first 4,000 words while only in Journal 2 could the first 3,000 words cover at 95%. To reach 98%, Journal 2 required 5,000 words; Journal 3, 7,000 words; and Journal 1, 9,000 words from the BNC/COCA list. Comparing the data in Tables 3 and 4, the students in this study had a similar coverage of 95% to that found in Journal 2, but the coverage was lower than those found in Journal 1 and Journal 3. This suggests that the students required more advanced vocabulary to get their papers published in high-quality journals, as shown in Table 4.

Table 4

Total BNC/COCA Vocabulary Found in Reference Corpus

Reference corpus	95%	98%
Journal 1	4000	9000
Journal 2	3000	5000
Journal 3	4000	7000

From the academic years 2019 and 2021, however, the results showed that the students reached 98% coverage by using the first 5,000 words for the cause-effect organizational pattern and reached 98% in the academic year 2020 by using the first 6,000 words for the cause-effect organizational pattern. This suggests that students used vocabulary at the level where they almost reached the standard of some journals when they wrote a more academic essay using more logical thinking in their writing, such as in cause-effect and problem-solution organizational patterns.

5. Limitations and Recommendations for Future Research

Data for this study were collected from writing done by second-year students majoring in English at a university in Thailand. According to the study plan in the B.A. English curriculum, the only writing course is scheduled to be studied in the second year, though a few other courses are focused on integrated skills (reading and writing).

Further studies on vocabulary size or lexical level need to be conducted using more samples of Thai English-major students and writing assignments to gain more insight into this topic. Also, as the results of this study have suggested that the organizational patterns seem to affect the students' lexical level, more studies need to be done to identify how they are related.

6. Conclusion

The results of the study revealed that, overall, sophomore students majoring in English had a lower level of academic vocabulary than expected. The results suggested that they would need more input on academic English to reach a higher academic word level when they graduate. Higher-level English courses should implement higher academic words to at least 7,000-10,000 words based on the academic articles of reliable academic journals. It is also necessary to provide first-year students with preparation courses to ensure the same level of English proficiency required to start learning English courses, which will also provide a solid base to improve their English proficiency to meet the requirements for graduate studies.

In terms of data collection, the results from this study suggest that using the writing assignments of the students can be a useful and reliable learner corpus for the study since the assignments are authentic, revealing students' existing vocabulary repertoire and their vocabulary levels not being interfered with the topic or vocabulary chosen for the test. Unlike VST, the learner corpus of writing assignments also shows how the students use the vocabulary in context, not just through memorization. Additionally, letting the students choose their own topic should ensure that their vocabulary repertoire is genuine.

7. About the Authors

Pong-ampai Kongcharoen is an assistant professor of English at the Department of Foreign Languages, Faculty of Humanities, Kasetsart University, Bangkok, Thailand. Her research interest lies in corpus linguistics, semantics, discourse analysis, vocabulary learning and teaching, and second language acquisition. She can be reached at pongampai.k@ku.th.

Jiraporn Dhanarattigannon, Ph.D. is an assistant professor in the Department of Foreign Languages at Kasetsart University, Bangkok, Thailand. Her areas of research interest are EFL writing (process-based approach), extensive reading, online language learning and teaching, and corpus-based research. She can be reached at jiraporndh2000@yahoo.com.

Tirote Thongnuan is a lecturer of English at the Department of Foreign Languages, Faculty of Humanities, Kasetsart University, Bangkok, Thailand. His areas of interest are corpus linguistics and corpus-based research, discourse analysis, English syntax and grammar, and translation studies. He can be reached at tirote.t@ku.th.

8. Acknowledgement

This research was funded by the Faculty of Humanities, Kasetsart University.

9. References

Abduh, A., & Rosmaladewi, R. (2017). Taking the Lextutor on-line tool to examine students' vocabulary level in business English students. *World Transactions on Engineering and Technology Education, 15*(3), 283–286.

Abe M. (2003). A Corpus-based contrastive analysis of spoken and written learner corpora: The case of Japanese-speaking learners of English. *Proceedings of the Corpus Linguistics 2003 conference*, 1–9. https://ucrel.lancs.ac.uk/publications/cl2003/papers/abe.pdf

Agerström J. (2000). Hedges in argumentative writing: A comparison of native and non-native Speakers of English. In Virtanen T. and Agerström J. (Eds.), *Three studies of learner discourse. Evidence from the international corpus of learner English*. (Vol. 10, pp. 5–42). Rapporter Från Växjö Universitet.

Aijmer K. (2002). Modality in advanced Swedish learners' written interlanguage. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 55–76). John Benjamins.

Aktas N. (2005, May 12–15). Functions of 'Shell Nouns' as cohesive devices in academic writing: A comparative corpus-based study [Paper presentation]. *The 26th International Computer Archive of Modern and Medieval English and the 6th American Association of Applied Corpus Linguistics conference*, University of Michigan, Ann Arbor, Michigan, United States.

Alfatle, A. B. M. (2016). Investigating the growth of vocabulary size and depth of word knowledge in Iraqi foreign language learners of English. [Master's Thesis, Missouri State University]. MSU Graduate Theses. https://bearworks.missouristate.edu/theses/2230

Allan, Q. G. (2002). The TELEC secondary learner corpus: A resource for teacher development. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 195–212). John Benjamins.

Axelsson, M. W. (2000). The use of a corpus of students' written production in university English teaching. *Korpusar I Forskning Och Undervisning, 13*, 293–303. https://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-39162

Bączkowska A. (2000). Creating a corpus of spoken English of Polish EFL learners. *PALC'99: Practical Applications in Language Corpora*, 221–232.

Belz J. A. (2004). Learner corpus analysis and the development of foreign language proficiency. *System, 32*(4), 577–591. https://doi.org/10.1016/j.system.2004.09.013

Benso B. (2000). *Adjective intensification in present-day English: A native vs. non-native corpus-based analysis* [Unpublished doctoral dissertation]. University of Torino.

Axelsson, M. W., & Berglund, Y. (2002). The uppsala student English corpus: A multi-faceted resource for research and course development. In L. Borin (Ed.), *Parallel corpora, parallel worlds* (pp. 79–90). Brill.

Chanchanglek, S., & Sriussadaporn, N. (2011). A corpus analysis of English vocabulary input in course materials used for Engineering students. *Journal of Studies in the Field of Humanities, 18*(1), 141–149.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

Crosthwaite, P. (2016). L2 English article use by L1 speakers of article-less languages. *International Journal of Learner Corpus Research, 2*(1), 68–100. https://doi.org/10.1075/IJLCR.2.1.03CRO

Davies, M. (2020). The COCA corpus. Corpus of Contemporary American English. https://www.english-corpora.org/coca/

Eriksson, A. (2008). *Tense and Aspect in Advanced Swedish Learners' Written English*. Göteborgs Universitet.

Gilquin, G., Papp, S., & Díez-Bedmar, M. B. (2008). *Linking Up Contrastive and Learner Corpus Research*. Brill.

Granger, S. (2008). Learner corpora. In Lüdeling, A. & Kytö, M. (Eds.), *Corpus Linguistics. An International 5 Handbook* (pp. 259–275). W. de Gruyter.

Granger, S. (2009, December 22). *Learner corpus bibliography*. https://sites.uclouvain.be/cecl/projects/Downloads/Learner%20Corpus%20Bibliography.pdf

Granger, S., Gilquin, G., & Meunier, F. (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press. https://doi.org/10.1017/CBO9781139649414

Gallagher, K. (2006). *Teaching Adolescent Writers*. Stenhouse Publishers.

Hsueh-chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403–430.

Huang, C., Cheng, W., Cheung, H., Harada, Y., Hong, H., Skoufaki, S., & Chen, H. K. Y. (2010). English learner corpus: Global perspectives with an Asian focus. In T. E. Kao & Y. F. Lin (Eds.), *A new look at language teaching and testing: English as subject and vehicle*. The Language Training and Testing Center (LTTC). https://www.researchgate.net/publication/262070267_English_learner_corpus_Global_perspectives_with_an_Asian_focus

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language, 22*(1), 15–30.

Liangpanit, C. (2010). *The development of a corpus-based vocabulary package for business English majors* [Master's thesis, Suranaree University of Technology]. http://sutir.sut.ac.th:8080/jspui/bitstream/ 123456789/3469/1/Fulltext.pdf

Ma, Q., & Mei, F. (2021). Review of corpus tools for vocabulary teaching and learning. *Journal of China Computer-Assisted Language Learning, 1*(1), 177–190. https://doi.org/10.1515/jccall-2021-2008

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book*. Routledge.

Meunier, F. (2021). Introduction to learner corpus research. In N. Tracy-Ventura & M. Paquot (Eds.), *The Routledge handbook of second language acquisition and corpora* (pp. 23–36). Routledge.

Mungkonwong, P., & Wudthayagorn, J. (2017). An investigation of vocabulary size of Thai freshmen and its relationship to years of English study. *LEARN Journal: Language Education and Acquisition Research Network, 10*(2), 1–9. https://so04.tci-thaijo.org/index.php/LEARN/article/view/111681

Nation, I. S. P. (2017). The BNC/COCA Level 6 word family lists (Version 1.0.0). https://www.victoria.ac.nz/__data/assets/pdf_file/0004/1689349/Information-on-the-BNC_COCA-word-family-lists-20180705.pdf

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins.

Nirattisai, S., & Chiramanee, T. (2014). Vocabulary learning strategies of Thai university students and its relationship to vocabulary size. *International Journal of English Language Education, 2*(1), 273–287.

Pringprom, P., & Obchuae, B. (2011, March 11–12). Relationship between vocabulary size and reading comprehension. *Proceeding of the 2nd International Conference on Foreign Language Learning and Teaching*, *1*(1), 182-191.

Pu, F. (2018). Research on corpus-based college English vocabulary teaching. *Advances in Social Science, Education and Humanities Research (ASSEHR), 300*, 688–692.

Qilichevna, T. M. (2020). Corpus based approach in vocabulary teaching. *European Journal of Research and Reflection in Educational Sciences, 8*(2), 172–176. https://www.idpublications.org/wp-content/uploads/2020/02/Full-Paper-CORPUS-BASED-APPROACH-IN-VOCABULARY-TEACHING.pdf

Schmitt, N., Schmitt, D., & Chapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88. https://doi.org/10.1177/026553220101800103

Shirato, J., & Stapleton, P. (2007). Comparing English vocabulary in a spoken learner corpus with a native speaker corpus: Pedagogical implications arising from an empirical study in Japan. *Language Teaching Research, 11*(4), 393–412. https://doi.org/10.1177/136216880708

Sun, Y., Zhang, J., & Scardamalia, M. (2010). Knowledge building and vocabulary growth over two years, Grades 3 and 4. *Instructional Science, 38*(2), 147–171.

Tiliakou, C., & Frantzi, K. T. (2021, September 10 - 12). Investigation of vocabulary in a corpus of written production of Greek learners of English [Paper presentation]. *4th International Conference on Teaching, Learning, and Education*, Zurich, Switzerland. https://www.researchgate.net/publication/362208851_Investigation_of_Vocabulary_in_a_Corpus_of_Written_Production_of_Greek_Learners_of_English

University of Oxford. (2015). *British national corpus*. http://www.natcorp.ox.ac.uk/

Veerachaisantikul, A., & Chootarut, S. (2016). General vocabulary in Thai EFL university students' writing: A corpus-based lexical study. *Journal of Advanced Research in Social Sciences and Humanities, 1*(1), 52–57. https://dx.doi.org/10.26500/JARSSH-01-2016-0107

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

Wiriyakarun, P. (2018). Examining Thai EFL learners' knowledge of academic English vocabulary. *The Liberal Arts Journal Faculty of Liberal Arts, Mahidol University, 1*(1), 119–132.

Zhiying, Z. (2007). Three-modality vocabulary knowledge of South China Agricultural University and Thai Prince of Songkla University EFL learners. *CELEA Journal, 30*(2), 25–33.