Incorporating DIY Corpus into English Writing in a Higher Vocational Institute in China: Students' Outcomes and Their Perceptions

by

GUAN Chao

A Thesis Submitted to

The Education University of Hong Kong

in Partial Fulfillment of the Requirement for

Degree of Doctor of Education

July 2023

Statement of Originality

I, GUAN Chao, hereby declare that I am the sole author of the thesis and the material presented in this thesis is my original work except those indicated in the acknowledgement. I further declare that I have followed the University's policies and regulations on Academic Honesty, Copyright and Plagiarism in writing the thesis and no material in this thesis has been submitted for a degree in this or other universities.

Abstract

General and DIY corpora have been demonstrated to be effective in improving university students' EFL or ESL writing skills. Often, these students have already achieved a high level of proficiency in the target language English. In recent years, there has been an increasing demand to support Chinese higher vocational institute students who generally possess lower English proficiency. Assisting these students in improving their English learning to meet future career requirements presents a significant challenge at higher vocational institutes in China.

This doctoral study examines the effects of implementing teacher-compiled DIY corpora on writing quality, vocabulary knowledge, and learner autonomy. Prior research has mainly focused on corpus-based approaches in EAP courses for advanced English learners (e.g., Charles, 2014; Smith, 2020). However, this study extends the application of DIY corpora to low-proficiency higher vocational institute students. In the mixed-method research, participants were divided into an experimental group of 46 students and a control group of 40 students. Both groups were required to complete four writing tasks using 6-8 target verbs each and a final writing exam. The experimental group followed corpus training adapted from the four-step corpus-based language pedagogy (CBLP) lesson design model from Ma et al. (2022), which included vocabulary knowledge tests, studying DIY corpus printouts, writing essays, and retaking vocabulary knowledge tests.

The quantitative results revealed a significant difference in writing quality during the fifth writing task conducted in the examination condition without reference to materials or using aids. Similarly, the frequency of correct target verb collocations exhibited statistically significant differences in the fourth and fifth writing tasks. Moreover, it is shown that accurate use of verb collocation is positively correlated to quality in English writing. In terms of vocabulary knowledge within the writing tasks, the t-test comparing the immediate post-test and the pre-test showed significant difference in the fourth vocabulary knowledge test. The analysis of the mean difference between the delayed post-test and pre-test indicated that both groups retained target vocabulary knowledge, but only the experimental group retained significantly more vocabulary knowledge. Regarding learner autonomy, the questionnaire results implied that participants in the experimental group significantly improved their perceived responsibilities towards English learning. However, both groups enhanced their frequency of English learning activities outside the classroom without a significant difference.

In the semi-focused group interviews, participants mentioned that referring to the DIY corpus printouts was an effective writing strategy, and target vocabulary knowledge was reinforced during corpus-aided writing process. As they talked about their voluntary learning supported by the DIY corpus, they also started to understand the value of learner autonomy. They also highly appreciated the comprehensibility of the DIY corpus printouts and provided suggestions for improving the teacher-compiled DIY corpus.

Pedagogical implications were made regarding the DIY corpus and the implementation of the CBLP model among lower-proficiency students. The study acknowledges several limitations and offers recommendations for future research. By demonstrating the effectiveness of incorporating DIY corpus to improve students' writing, vocabulary (verb noun collocations) and learner autonomy, this study makes a valuable contribution to the research on corpus-based language learning among lower English proficiency students by filling a significant research gap.

Keywords: corpus-based writing, vocabulary knowledge, DIY corpus, learner autonomy, corpus-based language pedagogy

Acknowledgement

I could not have finished my thesis without the direction and support of many people because the thesis could not have been realised without the assistance from my supervisors, friends, family and staff at the Education University of Hong Kong. First, I sincerely thank my principal supervisor, Dr. Ma Qing, Angel. She has given me tons of academic suggestions when I am lost in the field of data or research interests. Thankful for her patience when I am slow in processing the study. Grateful for her kind reminder when I lag behind the schedule. Her insightful comments on this thesis's content and careful, detailed checking of my work have always been priceless. I would also want to thank my co-supervisor, Dr. Chen, Hseuh Chu, Rebecca, for her continual and instructive suggestions in facilitating the thesis throughout the process. I also owe my deepest gratitude to Dr. Wu Wenli, Wendy. She helped proofread versions of some chapters and offered many constructive advice as a third party to guide my writing. Her life experience and attitude towards academic work impressed me greatly.

My wife, parents, daughter, and in-laws deserve my gratitude. My wife gave birth to my daughter during the epidemic era and suffered a lot when I was absent from work on the proposal and experiment preparations. My mom helped me a lot by sharing the family duty of caring for my daughter and daily housework, which saved me time to devote to my thesis. My daughter, who is unbelievably well-behaved and understanding of my work at night, profoundly promoted my determination for graduation because I hope she can see her father

succeed in persisting in his academic career at difficult time. I owe a debt of gratitude to my older relatives for their ongoing moral and material assistance as well as their encouragement during my EdD journey.

Great appreciation goes to professors, classmates, and friends because of their idea sharing and invaluable discussions, such as Dr. Lu Jinmiao, Ms. Fannie, Mei Fang. The colleagues and students the research supported my research and cooperated with me unconditionally, like Ms. Guo Xiaoli, Ms. Ji Jie, and so on. Their trust and engagement encouraged me to persist in the research.

Finally, I want to express my gratitude to the faculty and personnel of the Graduate School of the Education University for their generous help and the support team offered many constructive suggestions and hints for accomplishing the degree.

Table of Contents

List of Abbreviations

BFSU	Beijing Foreign Studies University
BNC	British National Corpus
CEE	College Entrance Examination
CEFR	Common European Framework of Reference for Languages
CELST	Computer-based English Listening and Speaking Test
CET4	College English Test Band Four
CET6	College English Tests Band Six
CLEC	Chinese Learner English Corpus
COCA	Corpus of Contemporary American English
CQP	Corpus Query Processor
DDL	Data Driven Learning
EAP	English for Academic Purposes
EFL	English as a Foreign Language
ESL	English as a Second Language
HVI	Higher Vocational Institute
HREC	Human Research Ethics Committee
ILH	Involvement Load Hypothesis
KWIC	Keyword in Context
L1	First Language
L2	Second Language

LSP	Language for Specific Purposes
MOE	Ministry of Education
NMET	National Matriculation English Test
PRETCO	Practical English Test for Colleges
QR Code	Quick Response Code
SD	Standard Deviation
SLA	Second Language Acquisition
SPSS	Statistical Package for the Social Sciences
SZIIT	Shenzhen Institute of Information Technology
TESOL	Teaching English to Speakers of Other Languages
TFA	Technique Feature Analysis
TOEFL	Test of English as a Foreign Language
VKS	Vocabulary Knowledge Scale
VLT	Vocabulary Levels Test
VPN	Virtual Private Network

List of Tables

List of Figures

Chapter 1: Introduction

This thesis focuses on the application of do-it-yourself (DIY) corpus in students' writing practice during their regular English class hours in a higher vocational institute in China. In addition to highlighting the overview of the background and objectives of the research, Chapter 1 also gives the research questions that will be investigated. A discussion of the study's importance is also included, and the chapter's conclusion includes an overview of the thesis structure.

1.1 Background of the Study

The current state of the employment market, industrial restructuring and the shortage of technical personnel have led to a significant increase in enrolments at vocational colleges, with an additional one million students enrolled (Li, 2020). This growth has presented both opportunities and challenges for vocational colleges, including a shortage of teaching faculty and a decline in the overall academic proficiency of students. Despite these challenges, the recently revised *Curriculum Standards for College English in Higher Vocational Education*, completed in April 2021, requires freshmen in higher vocational institutes to learn approximately 500 new words and a certain number of phrases. They also need to master a total of 2300–2600 words within their first two semesters, even though they may have struggled with English in previous years and their packed schedules of vocational skills training with other major-related courses. In light of these circumstances, the present research

was conducted at a prestigious higher vocational institute in Shenzhen, China. While this institute is renowned for its information technology majors, it is representative of higher vocational institutes due to its multidisciplinary approach and its ranking as the second-best institute in Guangdong Province and among the top 20 in China.

1.1.1 English Education in China

English language education in China has undergone several phases of development over the past four decades. Early in the 1980s, English was once again required for the College Entrance Examination (CEE) (Adamson, 2004). The College English Test (CET) Band 4 and Band 6 were introduced by the Ministry of Education (MOE) in 1987 and were intended for all college and university students whose specialties did not include English. Passing this test became essential for graduation, with higher-scoring institutions requiring better CET scores.

English became an obligatory subject beginning in Grade 3 in accordance with the *Guidelines for Curricular Reforms in Basic Education* in the autumn of 2001. In some developed regions, English instruction begins in Grade 1. All elementary school years must now include English instruction, according to the 2011 Elementary School Standards for English Curriculum. When children are 12 or 13 years old, they take an exam that includes English as a required subject. For approximately 63 million primary school students seeking admission into junior middle school, English once more becomes a required foreign language when pupils, who are between the ages of 15 and 16, pass the entrance exams to enroll in senior

high school. Since students must take the National Matriculation English Test (NMET), English as a foreign language instruction becomes more exam-focused, emphasizing reading, writing, and listening abilities. The People's Education Press, however, released the New English Language Curriculum for Senior Secondary Schools in 2003, which added a humanistic objective in addition to the conventional instrumental justifications. Following that, English speaking assessments started to be used in many developed regions and provinces. For instance, in Guangdong, computer-based English listening and speaking examinations (CELST), which account for 15 points out of a total of 150 points for English, were adopted annually in March. The MOE of China published the "Guidelines for Improving Teaching for University Undergraduate Students" in 2001, advising that 5% to 10% of all instruction at the undergraduate level be done in a foreign language. Higher education has also undergone modifications.

The CEE underwent revision in September 2014 based on the State Council's implementation recommendations for furthering the reform of the examination and enrolment system. This reform allowed students in Shanghai and Zhejiang Provinces to take the NMET twice, including listening and speaking tests. Similar reforms were implemented in Hunan Province in 2015. Reduced English lesson hours in the educational system has recently been the subject of a contentious discussion, although no official comments or actions have been taken.

After CEE (known as Gaokao), held every June, high school graduates proceed to higher

education. Only those who achieve a certain baseline score (at least 437 out of 750 in Guangdong Province in 2022) are eligible to pursue a four-year bachelor's degree. Other secondary school graduates may choose to continue their studies in a higher vocational institute to obtain a three-year certificate or opt for alternative paths, such as joining the army or entering the workforce. Freshmen in higher vocational institutes are still required to study English for at least two semesters, as stipulated in the newly revised *Curriculum Standards for College English in Higher Vocational Education* in April 2021 (MoE, 2021). The basic English module is a compulsory or limited elective course for students in higher vocational institutes, and it is offered in the first and second semesters, with a total of 128–144 class hours. Each credit is typically 16–18 class hours, totalling 8 credits. The curriculum standards also specify vocabulary size requirements for students in higher vocational institutes. Building upon the 1800–1900 words from secondary vocational education and 2000–2100 words from general school education, the basic module's goal is to teach students about 500 new words and a certain number of phrases, resulting in a total mastery of 2300–2600 words. Furthermore, autonomous learning has become a core competency in the curriculum standards for higher vocational education. Students are encouraged to develop awareness and abilities for lifelong learning based on the characteristics of English language learning and to enhance the accuracy and richness of their vocabulary use in expression. Moreover, teachers are required to assist students in learning vocabulary through various resources and methods, combined with thematic approaches for understanding and expressing relevant information.

English proficiency is also essential for students in higher vocational institutes who wish to

pursue further education. Currently, obtaining a degree from a higher vocational institute represents the highest level within this educational pathway. However, students who wish to enhance their academic qualifications can upgrade from a higher vocational college to an undergraduate programme by first entering the general education sequence. Qualified junior students in higher vocational institutes must pass a provincial unified examination, similar to the CEE, prior to graduation. This examination covers four subjects, with political theory and English each accounting for 100 points. The professional basic course, selected from nine subjects, including higher mathematics, management, economics and college Chinese, is worth a total of 100 points. In addition, a professional comprehensive course aligned with the requirements of the chosen major contributes 200 points. While the examination is relatively strict, the content of examination is simpler and test-takers have fewer university choices compared to CEE. After admission, they enroll in undergraduate colleges and universities for their third year. Following two years of full-time study, they earn an undergraduate diploma from a full-time general college and university. This represents the only pathway for junior students in higher vocational institutes to enter general undergraduate colleges and universities.

1.1.2 Importance of Incorporating DIY Corpus into English Writing

Every year, approximately 50% of high school graduates pursue undergraduate studies, 25% enrol in higher vocational institutes and the remaining join the army, retake the Gaokao or seek employment. The Gaokao score threshold for entry into higher vocational institutes is

typically no more than 50% of the total score of 750 (with Chinese, math and English each accounting for 150 points, and three other elective subjects accounting for 300 points), indicating relatively lower performance in the exam. As noted by Sakai and Takagi (2009), while successful learners may share similarities, unsuccessful learners fail in their own unique ways. Among secondary school graduates, those who choose science-oriented elective subjects tend to have poorer English proficiency, making them the weakest English learners among higher vocational institute enrollers.

While very little can be said without grammar, nothing can be said without vocabulary, according to David Wilkins in 1972, highlighting the crucial function of vocabulary in language communication. The most crucial goal for lower-level English learners is to build an adequate vocabulary, which primarily consists of prefabricated chunks of various kinds, with collocations being a typical type of chunk. The predictable ways that words are put together are referred to as collocations (Lewis, 2000, p. 48). However, traditional vocabulary teaching methods often focus on decontextualised individual word learning, such as translation or paraphrasing common words (Lewis, 2000, p. 33). The challenge of learning vocabulary lies in factors such as the *codability* of word morphological forms and the *arbitrariness* of form-meaning links (Hulstijn, 2001). Moreover, learning requires rearranging the learner's current interlanguage; it is not merely an additive procedure. Compared to rote learning or the direct teaching of new vocabulary, how can learners learn vocabulary easily? Lewis (2000) suggested that rather than focusing exclusively on helping students improve their grammar or learning a large number of new words that may only have

a few specific applications, teachers could draw students' attention to helpful collocations, helping them remember and utilise them successfully. Language input plays a significant role in enabling learners to adjust their internalised knowledge. When learners notice or become aware of language input, it becomes imperative for them to operationalise the elaborative processing. Two theoretical frameworks that use various attentional component characteristics and weights have sought to explain and quantify the depth of processing. In the context of intentional learning, where awareness is emphasised as a key factor in second-language vocabulary learning (Laufer & Hulstijn, 2001), the quality of elaborate processing is of great value to English as a Second Language (ESL) or English as a Foreign Language (EFL) learners. These frameworks analyse the significance of different learning tasks and task-induced involvement, which later became widely known as the involvement load hypothesis (ILH). Three elements are highlighted by ILH that stand for various levels of processing brought on by the tasks themselves (Laufer & Hulstijn, 2001). Drawing upon the framework of Incidental Language Learning Hypothesis (ILH), Nation and Webb (2011) further advanced this construct by introducing technique feature analysis (TFA), which encompasses five crucial parameters, namely "motivation, noticing, retrieval, generation, and retention". While comprehensible input (Krashen, 1995) has been emphasised for facilitating language input, achieving comprehensibility can be challenging, whereas output control is more manageable. How a material could be made to be a piece of comprehensible input to learners is a context-dependent topic, yet designing language output with a similar grading standard on a similar basis is more plausible for teachers. This is why entrance or application requirements often stipulate a standard language proficiency. Moreover, the output hypothesis

(Swain, 1985) asserts that the process of language output itself serves as a trigger for learning, as output fulfils three functions: “noticing, hypothesis testing and metalinguistic functions”. When language learners must develop a language, they become conscious of their knowledge and ignorance. Language learners use their interlanguage resources while experimenting with new structures as they build their utterances. They test their hypotheses while writing, either consciously or unconsciously, by revising and proofreading their work.

Corpus, as a reference resource, allows users to observe, analyse, hypothesise and formulate their own rules. Through trial-and-error procedures, learners can gain insights and broaden their understanding. More importantly, concordance lines in the Key Word in Context (KWIC) style highlight the frequency and salience of the target vocabulary, enabling students to focus their attention on achieving ‘noticing’. After noticing the target forms, they can consciously generate rules or patterns from these examples (Todd, 2001). Students can explore and exploit the corpus independently, thus becoming explorers of language learning.

Considering the unsatisfactory English learning experiences of enrollers in high vocational institutes, DIY corpus-based learning, a relatively new approach, provides an avenue worth exploring. It enables them to take more responsibilities of their education and develop as autonomous learners (Charles, 2012). DIY corpus-building refers to the construction of small-scale collections of e-texts by teachers or students for personal, specific, limited and local purposes. These corpora are also referred to as personal, self-compiled, disposable or local corpora (Zhang et. al., 2017; Charles, 2018, 2019). The process of building a DIY

corpus can enhance students' learner autonomy. First, students have the agency to choose the resources they want to use and the content in the corpus. Second, consulting their chosen corpus can reduce their reliance on other translation tools, which may result in multiple clicks to produce a piece of writing. Third, DIY corpora are not heavily reliant on internet speed, as they are stored as text files on a disc, allowing for seamless reference without obstacles.

1.2 Statement of the Problem

Even among advanced Chinese (English as a Foreign language) EFL learners, verb collocation errors make up a large share of all collocational faults, such as doctoral students (Wang & Li, 2018), and undergraduates majoring in English (Wang & Zhou, 2020). In addition, verb collocation errors account for approximately 11.61% of errors in the Chinese Learner English Corpus (CLEC) (Yang et al., 2005, p. 15). Second Language (L2) learners also use only approximately half the number of verb–noun collocations (as part of verb collocations) compared to native speakers, with almost 30% of their collocations being erroneous (Laufer & Waldman, 2011). Research analysing English argumentative essays written by 60 doctoral students revealed that approximately 25% of the most frequent verb + object constructions contained incorrect verb and noun usage (Wang & Li, 2018). Similarly, among 72 Chinese undergraduates majoring in English, approximately 30% of the most frequently used constructions, verb + object and verb + preposition, were incorrect (Wang & Zhou, 2020). The present researcher conducted careful proofreading of over 300 written parts of final exams from vocational institute students in China over two consecutive school years,

and the results confirmed that verb-related mistakes constituted a significant proportion of all writing errors.

Students in higher vocational institutes are often perceived to be low achievers based on their relatively lower admission threshold of the academic records. They are required to complete compulsory English public courses within a single academic year, as their schedules comprise practical training courses and other major-related courses that usually occupy at least 4 weeks of a 16-week semester for first-year students. Autonomous learning has become a core competency that is emphasised by new curriculum standards for higher vocational education. The autonomous learning can also help students fulfil their English course requirements with the challenge of their heavy major-related courses.

The existing literature on corpus application as a reference tool primarily focuses on English for academic purposes (EAP) courses for more advanced learners (e.g., Charles, 2012, 2018; Smith, 2020) and language for specific purposes (LSP) courses (e.g., Charles, 2019). However, the incorporation of corpus data into public English courses for less advanced L2 learners is less prevalent, with the exception of EFL writing courses (Luo & Liao, 2015). Moreover, studies often involve participants who are doctoral students (Charles, 2018), English majors (Luo, 2016) or upper-intermediate level learners (Sun & Hu, 2020), while inclusion of participants at lower L2 levels is relatively rare in the field.

The application of DIY corpora in higher vocational institutes is rare. The utilization of

corpus data need not be limited to small groups of students in EAP courses, as Charles (2012, 2018) pointed out, and incorporating corpus data in English writing could be an effective strategy to improve public English courses. In China, DIY corpora are mainly employed to explore language features and phenomena. This type of research starts with searching terms and collocations in concordance lines, from which regular patterns can be summarised. After discovering the characteristics and laws of language formation, the meaning can also be translated by analysing the semantic tendencies of its collocations. For instance, Sun (2017) gathered 400-word English writing assignments on six different subjects from non-English majors at a famous institution in China and examined their use of productive language. According to the Common European Framework of Reference for Languages (CEFR), it was discovered that undergraduates majoring in subjects other than English can develop a level B1 or B2 of productive vocabulary. Additionally, some researchers, such as Lee and Lin (2019), have experimented with incorporating corpus data into daily teaching contexts. The two researchers used the Corpus of Contemporary American English (COCA) to teach eight target terms to 27 EFL learners by randomly dividing them into deductive and inductive groups. Moreover, the DIY corpora in the empirical studies focused on students' hands-on application. Usually, these tailor-made corpora were compiled by students with high language proficiency for self-learning (e.g., Charles, 2012, 2014, 2018; Zhang et al., 2017) rather than by teachers. Since the target participants of these studies were low-proficiency students who may lack motivation and strategies for compiling their own DIY corpora, the teacher and researcher would compile a corpus suitable for their proficiency level and guide them in learning using the DIY corpus during class hours. In this sense, the current research extends

the exploration of the feasibility of teacher-compiled DIY corpora for low-proficiency EFL Chinese students enrolled at higher vocational institutes.

1.3 Purpose of the Study

In the different phases of students' English learning journeys, writing has become increasingly important, not only due to the writing section is the most heavily weighted part of English tests but also due to the fact that writing has great importance in communication in professional and potentially international settings. Among the various grammatical and lexical constructions, verb collocation errors are highly prevalent. Concordancing in the KWIC style can assist in addressing this issue by identifying and analysing patterns, including common collocates (Ackerley & Coccetta, 2007). Thus, the first purpose of this research is to assess how incorporating a DIY corpus into students' learning impacts verb collocation errors in writing. Writing, as an output skill, aligns with output hypothesis theory (Swain, 1995), in which learners redraft or proofread their work. The Output Hypothesis in no way minimizes the significance of Input. Instead of using input-based language learning techniques as a replacement, the goal is to absorb more than just the essentials of the message (Swain & Lapkin, 1995; Izumi & Bigelow, 2000). These conscious or unconscious actions involve hypothesis testing, and some of the verbs can become target words for students to use in their writing. Therefore, the second purpose of this study is to investigate the effects of including a DIY corpus on the acquisition of target vocabulary in writing assignments. Additionally, using a corpus encourages students to take more responsibilities of their writing

and develop into more independent learners (Charles, 2012). The DIY corpus, as a relatively recent academic tool, has the potential to foster autonomy development among less-motivated English learners in higher vocational institutes. Thus, the third purpose is to analyse the impact of DIY corpus uses on students' learner autonomy. Finally, to further research in this area of DIY corpus application, feedback from participants is invaluable, such as concerning their perceptions of the corpus, their use habits with the corpus, their perceived advantages of the corpus, and so on. Therefore, the fourth purpose is to gather participants' perceptions and attitudes towards using corpus data in English writing.

The research questions guiding this study are as follows:

(1) To what extent does incorporating a DIY corpus improve writing quality and the use of verb collocations?

(2) To what extent do participants improve their knowledge of target vocabulary within writing tasks after incorporating a DIY corpus?

(3) To what extent does the use of a corpus facilitate participants' learner autonomy?

(4) What are the participants' attitudes and perceptions towards using corpus data in English writing?

1.4 Significance of the Study

Increasing student enrolment in higher vocational institutes has led to lowered entrance thresholds. Consequently, there is a higher possibility of admitting students with lower

English proficiency. However, these students are still required to meet the same English graduation requirements, such as the CET Band 4 for non-English majors. This exam often puts more pressure on students than their major-related courses do, as English proficiency is crucial for their future careers. This study uses a corpus-based strategy to help L2 students with low competence levels write better overall and with more accurate collocations.

This study also addresses issues related to vocabulary retention to assist diligent students who may not achieve desirable outcomes within a short timeframe. Despite their best efforts, some students may find it difficult to acquire the necessary level of English proficiency in the allotted time of the first two years of university or the first year of higher vocational institute, as no English lessons are offered in later years. As a result, balancing students' compulsory courses with their English foundations poses a challenge. DIY corpora can serve as an appropriate tool for low-proficiency English learners in higher vocational institutes.

The feasibility of teacher-compiled DIY corpora can also be explored. Due to limited Internet access and hardware in higher vocational institutes, the present study involves an exploration of the teacher-compiled DIY corpus from the teacher's perspective. The DIY corpus not only enables teachers faced with teaching a new or unfamiliar course to familiarise themselves with necessary specialised discourse and reference resources but also provides illustrations and lexico-grammatical support to help create instructional materials that are appropriate and relevant (Charles, 2019). However, the widespread application of DIY corpora has usually been independently compiled by high-proficiency students for self-learning (Charles, 2012,

2014, 2018; Zhang et al., 2017).

Promoting learning autonomy is an important yet complex process. Autonomous learners take charge of their own education, choosing, for example, their English learning objectives and the frequency of English learning activities outside of the classroom. When the need for English proficiency arises, such as graduation or better job opportunities, students seek methods and tools for language acquisition. Exploring corpus data is an understudied approach that allows students to discover their own interests and develop their hypotheses about knowledge independently. Using corpus data from resources like the British National Corpus (BNC) and the Corpus of Contemporary American English (COCA), students not only gain access to new tools but also develop a habit of consulting and studying English in their spare time. DIY corpora can continue to support their progress as they aim to reach higher levels of English proficiency for other purposes. Once students become familiar with and accustomed to these language tools, their development as more autonomous language learners is promoted.

1.5 Organization of the Thesis

The thesis consists of seven chapters. Chapter 1 provides an introduction to the research, including the background of the study, a brief historical review of English education in China, new policies in higher vocational institutes and the theoretical background of the proposed study. The research questions are presented and key terms are also clearly defined, followed

by an explanation of the research's significance.

In Chapter 2, a comprehensive literature review is conducted. The chapter covers topics such as vocabulary knowledge and English writing, corpus-based studies in China and overseas, the relationship between corpus-based studies and learner autonomy and the use of DIY corpus instead of larger general corpus. Research gaps in the existing literature are also identified.

Chapter 3 reports on a pilot study conducted to test corpus-based approaches in an institute with similar participants. A quasi-experimental study was conducted to pilot the effect of corpus use on students' verb collocation errors in writing tasks. Large general corpora, such as COCA and BNC, were adopted to collect students' feedback. In addition, instruments, such as vocabulary knowledge tests and learner autonomy measurement tools, were used to ensure reliability. The main study mentioned in Chapters 4 and 5 is informed by the pilot study's findings.

Chapter 4 focuses on the research methodology. The rationale for choosing an explanatory sequential mixed-methods design is explained, followed by a description of the research context and participants. The chapter also provides a discussion of the experimental research method, including the instruments, data collection process and data analysis procedures. Additionally, the use of questionnaires and semi-focused group interviews for data gathering are discussed, as well as content analysis and grounded theory for qualitative data analysis.

Chapter 5 presents the results of the study. The results from the quantitative and qualitative analyses are presented in combination to address the research questions proposed in Chapter 1.

Chapter 6 focuses on the discussion. Key findings are presented, and key terms in chapter 2 are also discussed in relation to the results, such as writing quality, verb collocations, vocabulary knowledge, learner autonomy and feasibility of DIY corpus and CBLP for low-proficient students.

Chapter 7 concludes with an overview of the research process, a summary of the major findings. The pedagogical implications along with the research limitations and recommendations for future studies are presented before a short conclusion.

Chapter 2: Literature Review

This chapter provides an overview of relevant studies and theories that form the theoretical foundation for the DIY corpus-based approach to English learning. The chapter begins by explaining key terminologies related to vocabulary study. It then provides a comparison of intentional vocabulary learning with incidental vocabulary learning with reference to language input activities such as reading, listening and watching. The theoretical frameworks behind incidental vocabulary learning through language output, specifically writing, are also reviewed. Furthermore, the chapter provides a discussion of the significance of collocations in EFL writing, particularly verb collocations. Corpus-based empirical studies are also evaluated to demonstrate the benefits and drawbacks with regards to English instruction and to justify the selection of a DIY corpus. The chapter concludes with a summary of the research gaps.

2.1 Vocabulary Knowledge and EFL Writing

Vocabulary has long been recognised as a critical aspect of language proficiency since the 1990s. The amount of words that native speakers and language learners of other languages know varies significantly, as noted by Laufer (1998). Lewis (2000) also asserted that developing a sufficiently broad vocabulary presents a major challenge for language learners. Words are crucial to every area of life since they are the cornerstone of language and are required for speaking, reading, writing, and listening (Webb & Nation, 2017).

Activities for learning vocabulary are typically divided into intentional and incidental

learning (Nakata, 2008). According to Rieder (2003), incidental learning is the process of acquiring new words through various vocabulary-learning activities, such as reading (Coady & Huckin, 1997), listening (Vidal, 2003), watching (Danan, 2004) and so on. In contrast, intentional vocabulary learning involves activities specifically designed to memorise words (Hulstijn, 2001), including direct vocabulary acquisition. The current study will address both incidental and intentional vocabulary acquisition, taking into account the concept of vocabulary knowledge and the needs of the participants.

English writing is considered an essential component of EFL learning for Chinese students (Huo, 2014). Even in a student's native language, writing can be a slow and challenging process, and these difficulties are magnified when writing in a second language (Gilmore, 2009). Foreign language writing is often perceived as "non-native-like", and lexical poverty and collocation errors are common in writing (Luo, 2016). Vocabulary knowledge plays a central role in writing and significantly impacts the quality of written text (Nation, 2022, p. 226). One of the key elements influencing the quality of writing is regarded to be vocabulary (Yılmaz, 2017). Insufficient vocabulary knowledge is a major obstacle in foreign language writing (Huang, 2014; Mao et al., 2018), and lexical and grammatical precision can enhance L2 writers' overall writing quality (Huang, 2014). In fact, many students simply string words together to complete their English writing assignments (Mao et al., 2018), and even non-native English scholars face language-related challenges, such as vocabulary, when writing research articles (Chang, 2014).

However, in some cases, although L2 learners may only need a partial understanding of a

term to understand it, having a larger lexicon is often advantageous (Qian, 2008). This part concludes with a discussion of the sufficiency and nature of vocabulary knowledge necessary for effective language and use in various contexts.

2.1.1 Vocabulary Knowledge

Vocabulary knowledge is a multifaceted concept that encompasses different types of knowledge about words (Nation, 2022, p. 49). Understanding a word involves more than just knowing its basic form, as words can have different forms and variants. For example, *employs, employers, employees, employed, employment, employing, unemployment*, and other versions of the word could be derived from *employ*. Hence, if we believe that understanding the word *employ* involves understanding all of its variants, managing studies could become challenging. This raises questions about the scope of word knowledge: Should all proper nouns and alternate spellings be considered separate words? How do we explain the discrepancies in spelling between British and American English? To address these complexities, lexical researchers have developed different criteria to define and categorise word knowledge (Qian, 2008). One approach is to categorise word knowledge into specific domains, such as grammatical and word association knowledge (Schmitt & Meara, 1997). Grammatical knowledge includes understanding the class, morphological traits and affixes of a word, and word associations refer to the connections or associations that learners make between words. For instance, words like interrupt, bother, disturb and intervene may be considered synonyms in Mandarin because of their similar meanings in English. The

vocabulary knowledge of the research subjects was evaluated by van Zeeland and Schmitt (2013) using form identification, grammatical recognition, and meaning recall.

To the literature on vocabulary knowledge, many second-language acquisition (SLA) researchers have made contributions, examining it from linguistic, psycholinguistic and sociolinguistic perspectives (Richards, 1976; Nation, 1990; Wesche & Paribakht, 1996; Webb, 2005). A thorough examination of vocabulary knowledge was suggested by Richards (1976), which included word frequency, vocabulary growth in non-native speakers, collocations, register, case connections, underlying forms, word associations, and semantic structures. But this definition disregards some lexical knowledge components, such as spelling and pronunciation. Wesche and Paribakht (1996) introduced the concepts of breadth and depth of vocabulary knowledge. Depth refers to understanding the form, meaning and use of a word, while breadth relates to the extent of knowledge coverage. For English language learners, Laufer (1988) suggested that a vocabulary of 3000–5000 word families is necessary to achieve basic text comprehension, with 95% coverage deemed sufficient. A word family consists of a headword, its inflected forms and related derived forms (Nation, 2022, p. 11). However, this vocabulary requirement can pose a challenge for students in higher vocational institutes, as their curriculum standards (*Curriculum Standards for College English in Higher Vocational Education* issued in April 2021) typically aim for a mastery of 2300–2600 words, assuming a mastery of approximately 1900 words from secondary vocational education or 2100 words from general high school education. Additionally, the CET Band 4 (CET-4) requires a minimum vocabulary size of 4200 words. As a result, there is

a sizable discrepancy between the vocabulary demands of the curriculum and those of the students, which must be taken into account when acquiring new words.

Vocabulary knowledge can be further classified into receptive and productive knowledge, which apply to each aspect of vocabulary (Webb, 2005; Ma & Sin, 2015; Nation, 2022, p. 49). Receptive knowledge involves understanding language input through reading or listening, while productive knowledge involves creating language through speaking and writing to communicate with others. However, the terms “receptive” and “productive” are not entirely accurate, as receptive skills also involve creating meaning when we listen or read. Recognizing a word’s form when reading or listening and remembering its meaning are the main goals of using receptive vocabulary. By using a vocabulary effectively, one wants to convey meaning in speech or writing, and they search for, create and employ the right words to do so (Nation, 2022, p. 52). Vocabulary knowledge will be assessed in terms of both reception and production in the present study.

2.1.2 Input-based Approaches for Vocabulary Learning

Language learners typically think that their struggles with receptive and productive language use are primarily caused by a lack of vocabulary (Nakata, 2008). Adequate language input places a high priority on vocabulary knowledge (Nation, 1990). This section outlines four typical methods for Chinese EFL students to study vocabulary. Older-style direct vocabulary learning is an example of intentional vocabulary learning, whereas reading, listening, and

watching are instances of incidental vocabulary learning.

2.1.2.1 Direct Vocabulary Learning

One of the most popular methods for increasing vocabulary is by using word lists and word cards (Nakata, 2008). A word list is a piece of paper having L2 words and their First Language (L1) translations or definitions printed on it. A set of playing cards called "word cards" has L2 terms written on one side and their L1 definitions or translations written on the other. This method eliminates the requirement for a threshold, such as the 3000-word vocabulary size proposed by Nation (1985), which requires a coverage of at least 95% of a text before students can effectively learn from the context with unsimplified material. Moreover, learners can memorise a list of between 30 and 100 L2 terms and their L1 equivalents in an hour and remember them for weeks afterwards (Nation, 1982, 1990). Thus, the main advantages of direct vocabulary learning are efficiency and focus.

However, Oxford and Crookall (1990, pp. 9–10) argued that these decontextualised techniques "remove the word completely from any communicative context that might help the learner remember and that might provide some notion as to how the word is used as a part of the language". This implies that they may not be suitable for memory or actual usage. Therefore, to bridge the gap between syllabus requirements and students' needs, other methods that facilitate long-term memory and usage should be considered.

Referring to L1 vocabulary acquisition, L1 students can understand language by listening to it

and processing it for comprehension, as well as by reading at the beginning of their schooling (Vidal, 2011). I will review vocabulary acquisition through reading and listening in the following sections.

2.1.2.2 Reading to Learn Vocabulary

Vocabulary acquisition through reading has more disadvantages than advantages, and several factors affect its effectiveness. When L2 learners desire to read to learn new words, they first notice unfamiliar words and then try to discover the meanings of the target words by inferring based on contextual clues or by consulting a dictionary. According to Vidal (2011), low-proficiency students needed longer processing time since they had more trouble keeping up with academic lectures. Thus, they can benefit more from written text and focus on words they do not understand and backtrack if necessary. However, acquiring incidental vocabulary through in-depth reading is a time-consuming and error-prone process, with minimal vocabulary gains (Peters et al., 2009). Moreover, it is impossible to predict which words will be learned or the extent of vocabulary acquisition (Coady & Huckin, 1997, chapter 9) because when reading a text, L2 learners sometimes fail to recognize unusual words. Even if they do, their general lack of vocabulary knowledge may prevent them from always being able to deduce the meaning of the words. L2 learners do not acquire enough information or come across enough new words in formal language learning environments to achieve this (Peters et al., 2009). Additionally, learning the vocabulary of a second language is a gradual process, and encountering unknown words 10 times in context can result in significant

learning gains. However, more than 10 repetitions may be necessary to fully understand a word (Webb, 2007). Therefore, Peters et al. (2009) suggested three factors for successful L2 vocabulary acquisition through reading: learners should discover the meaning of unfamiliar words, process the lexical information elaborately and reinforce the form-meaning connections of these words through repetition. Later, Vidal (2011) discovered four factors: "frequency of occurrence, type of word, type of elaboration and predictability from word form and parts"; these factors made an important contribution to the encoding of new lexical items. Moreover, Zhao et al. (2016) provided further evidence that learners' linguistic, affective and cognitive characteristics influence the acquisition of incidental L2 vocabulary. These learner factors include L2 proficiency, motivation, anxiety and mastery techniques.

To address these factors, researchers have developed various approaches to accelerate vocabulary learning. This research has identified four enhancement strategies: glosses, dictionaries, reading comprehension and reading comprehension combined with fill-in-the-blank vocabulary exercises to improve vocabulary learning (Zhao & Guo, 2012). The findings indicate that glosses outperformed dictionaries in terms of receptive knowledge. Additionally, reading comprehension, when combined with vocabulary exercises as a post-reading activity, proved more effective for both receptive and productive knowledge. Furthermore, Liu (2017) investigated the impact of different glosses (i.e., Chinese, English and bilingual) on incidental vocabulary learning through reading. These findings showed that bilingual glosses were more effective at fostering a depth of vocabulary knowledge than Chinese glosses were in fostering immediate retention. Bilingual glosses, on the other hand,

stood out in delayed retention and helped people broaden their vocabulary knowledge.

Another approach for enhancing vocabulary through reading is word-focused exercises. Vocabulary worksheets developed by Wesche and Paribakht (2000) are carefully examined and divided into five stages—comprehended input, manipulation, interpretation, output and perceived input or notice. These stages are evident in various tasks, such as matching definitions between L1 and L2, collocational matching and responding to queries. The process begins with establishing the form-meaning association to learn a new word, and auditory ethnography from rehearsals plays a crucial role in aiding memory. Through subsequent knowledge expansion, the word becomes readily accessible and is ultimately stored in long-term memory through frequent usage or practice in diverse contexts (Wesche & Paribakht, 2000).

Table 1. Typology of word-focused exercises (Wesche & Paribakht, 2000)

1. *Selective attention:* Attract learners' attention to the target words using specific techniques, such as underlining and circling the words.
2. *Recognition:* Building upon the previous step, learners gain a general understanding of the spelling of target words and then connect the word form with one of its meanings. This phase may involve matching activities with definitions or synonyms or undertaking multiple-choice exercises.

3. *Manipulation:* Involves structural analysis and understanding of target words. Learners are required to change the grammatical category of the target word or construct words using affixes.

4. *Interpretation*: By analysing the semantic and syntactic relationship of the target word with other words in given contexts, such as collocations, synonyms and antonyms, learners develop a deeper comprehension of the target words. Exercises may include guessing meanings in contexts and multiple-choice cloze exercises.

5. *Production*: Learners are expected to use the target vocabulary in new contexts, ensuring the proper retrieval of meanings. Examples of exercises include L2 to L1 sentence translations and open cloze exercises.

Tasks with word focus also support ILH (Hulstijn & Laufer, 2001). Empirical experiments have demonstrated that the level of task participation load is linked to language retention. For example, the memory rate decreases, progressing from composition work to reading combined with fill-in exercises and finally to reading alone.

However, research also reveals that learning a new language, whether through reading or other methods, is a gradual and complex process (Paribakht & Wesche, 1996). Four factors affect students' ability to learn new vocabulary from the texts they read: interest in the subject, understanding of the target passages, understanding of the new vocabulary, word meanings and contextual signals, and generative processing of the target words (Nation, 2001, p. 117). It is advisable for target words to be closely connected to the core theme of the

passages, presenting them in a narrative or involving students in the stories. This method helps students handle a certain density of unfamiliar words. The comprehension threshold, however, continues to be difficult for students to meet because they need to comprehend at least 95% of the text's words for basic comprehension, with a preference for 98% coverage, or one in every 52 to three words every minute (Laufer, 1988; Hu & Nation, 2000).

In conclusion, successful vocabulary acquisition through reading requires several prerequisites, such as noticing unfamiliar words, elaborating on their meanings and reinforcing learning through repetition (Peters et al., 2009). Additionally, learner factors (Zhao et al., 2016) can influence the effectiveness of this method. Despite the various approaches developed to overcome these limitations (Wesche & Paribakht, 2000), the threshold of lexical coverage required for basic understanding often discourages learners, making the process slow and prone to errors (Peters et al., 2009).

2.1.2.3 Listening to Learn Vocabulary

Research findings indicate that L1 learners have better language comprehension through listening than through reading. As they progress through school grades, their reading skills improve, eventually narrowing the gap between their listening and reading abilities (Vidal, 2011). Additionally, for adequate listening comprehension, non-native language speakers only need to be familiar with between 2000 and 3000 word families (van Zeeland & Schmitt, 2013), whereas reading requires a basic knowledge of 3000–5000 word families (Laufer,

1988).

However, due to its implicit nature, which is characterized by the transient nature of acoustic information and the challenge in accessing the underlying processes, L2 listening continues to be the least explored of the four language skills (Vandergrift, 2013). Top-down listeners construct a conceptual framework for comprehension using context and prior knowledge, such as topic, culture, and other schema knowledge stored in their long-term memory. The bottom-up method, in contrast, involves building meaning incrementally by gradually merging ever-larger units of meaning starting at the phoneme level and working your way up to the discourse level.

The process of learning vocabulary through listening (Vidal, 2003) includes an auditory presentation, phonological memory and vocabulary acquisition. Initially, the listener hears the speech as raw words temporarily retained in the memory before being replaced by a complete understanding of the phrases, which are subsequently stored in the long-term memory.

However, listeners may encounter several difficulties (Vandergrift, 2013) during different phases of the listening process. In the perceptual processing phase, they may miss the beginning of the text, neglect what follows or struggle to chunk the stream of speech, among other challenges. In the parsing phase, listeners may struggle to comprehend subsequent parts because they do not understand the earlier missed part. In the utilisation phase, the L2 listener may face the dilemma of understanding individual words but not the overall message.

Furthermore, studies have revealed the knowledge gains that would not have been discovered using a solely conventional form-meaning test format. It seems that while certain types of knowledge (such as word form) are relatively easier to acquire through L2 listening, others (such as meaning) are not (van Zeeland & Schmitt, 2013).

In conclusion, vocabulary acquisition through listening seems to lower the threshold of vocabulary size compared to reading, but it brings along more difficulties throughout the listening comprehension process.

2.1.2.4 Watching to Learn Vocabulary

The availability of videos with subtitles and soundtracks in multiple languages provides L2 listeners with the option to receive assistance in either L1 or L2 to improve their comprehension (Vandergrift, 2007). With advancements in virtual reality technology, watching videos has become a common activity. Learners can acquire new vocabulary while enjoying captivating movies accompanied by sound, with or without subtitles. Similarly, TV shows are often referred to as audiovisual input in the SLA field. In 1983, Karen Price introduced subtitles or captions to make foreign movies or TV shows more accessible to people who are hard of hearing or deaf. Recent research has indicated that students can read the captions or subtitles displayed below or beside the screen while simultaneously listening to the dialogue spoken by the characters. Subtitling helps listeners visualise what they hear, improves language comprehension and offers other cognitive advantages (Danan, 2004).

Therefore, in addition to reading and listening, watching TV might be a useful strategy for boosting students' vocabulary knowledge (Peters & Webb, 2018).

The effectiveness of using audiovisual materials in language learning has been demonstrated through numerous studies (Yuksel & Tanriverdi, 2009; Peters & Webb, 2018). In Price's extensive study, 500 participants watched four video segments with or without captions, and the findings showed that the participants were more likely to understand the material when captions were available. In a study by Yuksel and Tanriverdi (2009), 120 college students were randomly assigned into two groups. While Group B viewed the same video without captions, Group A watched a movie clip with subtitles. Both groups demonstrated improved vocabulary knowledge, but Group A showed a statistically higher improvement in the post-test using a 20-item vocabulary knowledge scale. This finding aligns with the dual coding theory (Mayer, 2014), which suggests that language input in both oral and written modes stimulates the verbal and imagery systems, leading to a deeper level of processing and recall.

One research area focuses on how different subtitle delivery techniques affect students' learning outcomes in terms of vocabulary and other language abilities. Subtitle formats can be broadly categorised into non-captions, full captions and target-word captions. There were also variations in subtitle types to differentiate languages during reading and listening. Options include bimodal subtitling (L2 in audio, L1 in subtitles), normal subtitling (L2 in both audio and subtitles) and multilingual subtitling (L2 in audio, L1 and L2 in subtitles). The effects of these different modes can be further investigated across various age groups,

educational levels and other factors. Zarei (2009) found that bimodal subtitling significantly outperformed the other two types in L2 vocabulary recognition and recall, contrary to the findings of Gorjian (2014), who carried out research among 90 first-year Bachelor of Arts students at an Iranian university concentrating in English Translation. In terms of learning new vocabulary words, Gorjian discovered that the reversed subtitling group greatly outperformed the conventional subtitling group. As a result, not all products and viewers with all levels of language competency may benefit from captioning (Danan, 2004).

The main advantages of using audiovisual media lie in providing visual clues, context (Danan, 2004) and discourse (Bal-Gezegin, 2014), as it closely resembles real-life situations. It has been discovered that using video as a resource that represents the target language and culture is a successful teaching technique (Bal-Gezegin, 2014), as its authenticity helps compensate for the limited access to authentic materials in EFL environments. However, learners' prior vocabulary knowledge and the frequency of occurrence influence their learning gains through watching videos (Peters & Webb, 2018). Additionally, the time constraints of audiovisual materials may pose another challenge, as the treatments in the aforementioned empirical studies are limited, such as 2- to 3-minute videos (Bal-Gezegin, 2014), watching a 9-minute and 14-second clip twice (Yuksel & Tanriverdi, 2009) or a maximum of 15 minutes (Peters & Webb, 2018).

2.1.2.5 Summary of Input-based Methods for Vocabulary Learning

The literature reviewed above highlights the limited vocabulary gains achieved through input-based methods, such as listening and reading (Vidal, 2011). The decontextualised nature of direct vocabulary learning through word lists and word cards hinders long-term memory retention (Oxford & Crookall, 1990, pp. 9–10). Vocabulary acquisition through reading is a slow and error-prone process influenced by numerous factors (Peters et al., 2009). Despite the adoption of different approaches to address these issues, the threshold of lexical coverage in reading texts remains a significant consideration. Meanwhile, listening lowers the vocabulary threshold compared to reading (van Zeeland & Schmitt, 2013), but the process of listening comprehension itself is more complex than vocabulary acquisition (Vandergrift, 2013). Meanwhile, watching, although a relatively new method, combines reading and listening with visual cues in the vocabulary learning process (Danan, 2004). However, it is impractical to play video clips throughout the entire class hours, and there is lack of "ecologically valid" materials for practical use (Peters & Webb, 2018).

The researcher of the present study began to question the popularity of *Comprehensible Input*, as a crucial question remained unanswered: How comprehensible can educators make language input when providing it to students? Since this question depends on knowledge of the learners' proficiency and many other aspects, it is ambiguous and, in some cases, impossible to answer. Based on long-term observations of a French immersion programme in Canada, Swain (1985, 1993) argued that providing comprehensible input alone is not sufficient for acquisition to take place; learners also need to engage in output production to process language more deeply.

2.1.3 Output-based Theories for Vocabulary Learning

Given that vocabulary development is a prerequisite for language development and that it is a key indicator of reading comprehension and second language acquisition, researchers, teachers and students have been working to develop more effective vocabulary strategies (Hu & Nassaji, 2016; Rahmanian & Soleimani, 2018). In recent years, increasing attention has been paid to understanding the best approaches for learning vocabulary in a second language (Laufer & Hulstijn, 2001), with SLA distinguishing between implicit and explicit learning, both of which rely on the learners' attention. These concepts are rooted in the groundbreaking study by Craik and Lockhart (1972), who introduced the idea of processing depth. Perception involves rapid analyses of stimuli with early and later stages of analysis (Craik & Lockhart, 1972; Kim, 2011). When learning a language, learners initially focus on sensory or physical aspects, such as colour, pitch, shapes and numbers. Later stages involve establishing connections between language input and prior knowledge, emphasising pattern-finding and meaning-extraction. In order words, Craik and Lockhart (1972) proposed a two-level processing model, with shallow processing for surface-level features like spelling and sound and deep processing for semantic meaning (Fløan, 2015). In summary, effective second-language learning requires learners to process various characteristics of target words and consciously pay attention to them. *Elaborative processing*, commonly known as deeper processing, is crucial for learning L2 vocabulary (Hu & Nassaji, 2016), and learners' interpretation of information determines whether it is retained in their long-term memory.

However, the first step in language learning is to notice the stimulus.

2.1.3.1 Output Hypothesis

It is notable that the output hypothesis does not downplay the significance of input. Instead, it tries to push learners over the point of minimal message understanding by enhancing and reinforcing input-based approaches to language acquisition (Swain & Lapkin, 1995; Izumi & Bigelow, 2000).

Prior to the proposal of the output hypothesis, output was seen as a mere indicator of an already-acquired second language, serving no significant function in the language acquisition process (Izumi, 2003). Swain's (1985) output hypothesis was based on her long-term observations of a French immersion programme in Canada, where she discovered that immersion learners acquired comprehension and phonological skills comparable to native speakers, but trailed behind them in production abilities, particularly grammatical accuracy. Russell (2014) attributed these findings to the nature of immersion education, in which students receive large amounts of comprehensible input in second-language instruction but engage in minimal production in the classroom.

Output was believed to serve four functions: developing automaticity, testing hypotheses, metalinguistic reflection and consciousness-raising (Izumi, 2002, 2003). First, output provides opportunities to improve language use automatically. Second, language learners

experiment with new structures, drawing on resources from other languages as they construct utterances. Moreover, when writing, learners edit or proofread their work and compensate for missing meaning in their spoken language. All of these intentional or unintentional behaviours involve hypothesis testing. Third, learners' output serves as a metalinguistic function, as they reflect on their own use of the target language, helping them control and internalise linguistic knowledge. Finally, learners need to be aware of what they know and do not know when they need to produce a language. This contributes to recognising and addressing problems and prompts learners to attend to relevant information in the input.

Among the various empirical studies (e.g., Izumi, 2002) investigating the consciousness-raising function of output, writing requires more attention regarding its impact on EFL vocabulary knowledge. In contrast to exposure to textual augmentation when learners were directed to read for meaning, Russell (2014) discovered that when learners were instructed to read texts and then reconstruct them, pushed output followed by exposure to targeted forms in subsequent input resulted in greater learning gains. This finding aligns with the research conducted by Izumi (2002) after replicating the same study. Both studies confirmed that output facilitates the noticing function of the output hypothesis in SLA. However, Izumi and Bigelow (2000) found that essay-writing tasks were much more susceptible to individual variations compared to text reconstruction tasks. Further research is necessary to determine more precise and effective uses of output in L2 teaching (Izumi & Bigelow, 2000). For EFL vocabulary understanding, a definition task was found to be more effective than other tasks including matching, choosing, and combining (Bao, 2008). Therefore, the impact of writing

on vocabulary learning warrants further exploration.

2.1.3.2 Involvement Load Hypothesis

For ESL or EFL learners, the significance of elaborative processing is emphasized, especially in the context of intentional learning. In order to learn vocabulary in a second language, awareness is once again stressed (Laufer & Hulstijn, 2001).

Formally known as ILH, Laufer and Hulstijn (2001) proposed that task-induced involvement comprises three distinct components: "need, search and evaluation". These components describe different processing depths and work together to enhance the retention of target words.

Table 2. The ILH's components and level of involvement index (Alcaraz Marmol & Almela Sanchez-Lafuente, 2013)

Components	Involvement Index	Definition
Need	0 (none)	Learners have no need to learn the word.
	1 (moderate)	Learners are required to learn the word.
	2 (Strong)	Learners decide to learn the word.
Search	0 (absence)	Learners do not look for word form or meaning.
	1 (existence)	Word form and meaning are found by learners.
Evaluation	0 (none)	The word is not compared.
	1 (moderate)	The word is compared in a provided context.
	2 (Strong)	The word is compared in self-created context.

There are three different levels of processing depth for the word "need", which serves as the

motivation for learning new words. When students have a genuine need to acquire knowledge, there is a strong desire and a high likelihood of retaining that knowledge. If someone else, such as a teacher, imposes the need to learn something, such as by giving them instructions to look up new vocabulary or expressions in an upcoming course, then the need is only moderate. Meanwhile, when the meaning of target vocabulary is given in margins or glosses, the need is almost non-existent. The "search" element focuses on how learners establish connections between the form and meaning of words, and this can be significantly influenced by whether the search is used for productive or receptive retrieval (Nation & Webb, 2011). For instance, learners engage in more search activities when trying to find the word form than when finding the meaning. The third component, "evaluation", examines how learners compare new words with other words to expand their vocabulary knowledge and retain them in their long-term memory. If students determine whether the meanings of words are applicable in various contexts rather than comparing different meanings of words, the evaluation of this component becomes more significant (Hu & Nassaji, 2016).

Researchers can use this framework to categorise the need and evaluation factors into high, moderate and low levels of involvement, quantifying them as 0 for low, 1 for moderate and 2 for high. In contrast, the search factor can be quantified as 1 or 0 to indicate its presence or absence, respectively. By considering these factors together, researchers and teachers can design various types of activities to assess and enhance learners' engagement. Overall, the likelihood of language retention increases with the level of involvement.

Hulstijn and Laufer (2001) conducted experiments to examine the impact of involvement load on the retention of target words. They developed two tests on incidental vocabulary acquisition involving 10 English terms by young adult EFL learners in the Netherlands and Israel.

Table 3. Three activities with varying levels of involvement loads (Hulstijn & Laufer, 2001)

Tasks	Need	Search	Evaluation	Index
Reading comprehension with marginal glosses	1	0	0	1
Reading comprehension plus 'fill-in'	1	0	1	2
Writing a composition and incorporating the target words	1	0	2	3

During the experiment, they observed variations in the time spent on each task, which ranged from 40 to 45 minutes for Task 1, 50 to 55 minutes for Task 2 and 70 to 80 minutes for Task 3. The composition assignment yielded the most effective outcome based on the results. Kim (2008) investigated the impact of different task combinations under similar levels of involvement load. The study focused on writing compositions and writing sentences, which had comparable involvement loads but different task combinations. The findings indicated that new words were initially learned and retained in similar ways. Kim (2011) conducted two additional experiments, one replicating Hulstijn and Laufer's (2001) study with two proficiency levels and another focusing on different task types (reading with gap-filling and composition writing) with the same involvement load index scores. The descriptive data showed that the composition group outperformed the other proficiency levels in terms of vocabulary retention. Therefore, equal engagement loads resulted in comparable learning

outcomes, while proficiency level did not interact with task types.

Kim (2008) also emphasised the need for further studies on the importance of each ILH component. Subsequent research focused on the search factor (Alcaraz Marmol & Almela Sanchez-Lafuente, 2013). 28 ten-year-old students participated in the study. They were split into four groups and given various activities, including reading with gap-filling, writing with glosses, and creating sentences using bilingual dictionaries. For Tasks 1 through 3, the ILH index values ranged from 1 to 4. Both the receptive and productive tests showed that the group that finished Task 3 fairly better than the others. The researchers suggested that this could be attributed to the participants' developing metacognitive abilities, such as using a dictionary, despite their young age. However, the results did not fully align with ILH, indicating the need for further refinement of the theory (Alcaraz Marmol & Almela Sanchez-Lafuente, 2013).

2.1.3.3 Technique Feature Analysis

Recognising that the time spent on a task may have influenced the research outcomes and considering that the three ILH components were insufficient (Hu & Nassaji, 2016), Nation and Webb (2011) modified and expanded the ILH to form technique feature analysis (TFA).

Table 4. Technique Feature Analysis Checklist (Nation & Webb, 2011, p. 7)

Criteria		Scores	
Motivation	Is there a clear vocabulary learning goal?	0	1
	Does the activity motivate learning?	0	1
	Do the learners select the words?	0	1
Noticing	Does the activity focus attention on the target words?	0	1
	Does the activity raise awareness of the new vocabulary learning?	0	1
	Does the activity involve negotiation?	0	1
Retrieval	Does the activity involve retrieval of the word?	0	1
	Is it productive retrieval?	0	1
	Is it recall?	0	1
	Are there multiple retrievals of each word?	0	1
	Is there spacing between retrievals?	0	1
Generation	Does the activity involve generative use?	0	1
	Is it productive?	0	1
	Is there a marked change that involves the use of other words?	0	1
Retention	Does the activity ensure successful linking of form and meaning?	0	1
	Does the activity involve instantiation?	0	1
	Does the activity involve imaging?	0	1
	Does the activity avoid interference?	0	1
	Maximum score		18

Table 4 provides more specific guidelines for each criterion. Whether the vocabulary exercise incorporates negotiating and has a distinct learning purpose is taken into account by the "motivation" component. The "noticing" component emphasises students' awareness of and emphasis on words. Receptive and productive retrieval, memory as a distinct item, and the gaps between retrievals are all included in the "retrieval" component. "Generation" distinguishes between the receptive and productive aspects, and "retention" examines whether the task successfully links form and meaning, incorporates instantiation and imagery and avoids interference.

TFA builds upon ILH but expands and clarifies it further, leading to a series of comparative studies. The TFA framework has received significant attention in numerous studies. One study involved 96 adult EFL students with six to seven years of English learning experience and a business major. The students were divided into four groups and instructed to learn the definitions of 14 new words. Each group received one of four tasks, and their involvement load was assessed using ILH and TFA. The tasks included reading a text with multiple-choice items, reading a text and selecting definitions, reading with fill-in-the-blanks and reading and rewording sentences (Hu & Nassaji, 2016).

Table 5. Different indexes based on ILH and TFA (Hu & Nassaji, 2016)

Task	ILH index	TFA index
1	3	6
2	3	6
3	2	7
4	3	6

Overall, the findings of the study indicated that TFA was a better predictor of lexical improvements than ILH (Hu & Nassaji, 2016). It was determined that ILH “could not be a good predictor”, while TFA showed promise in predicting pre-test score changes but not in during-task activity. As a result, the second research was significantly more favourable than the first in terms of the efficiency of TFA frameworks (Rahmanian & Soleimani, 2018). 120 Iranian students studying advanced English at an English institute between the ages of 15 and 25 were recruited for a different study. Only 90 underwent Test of English as a Foreign Language (TOEFL) screening before being randomly divided into three groups. There was a

total of three tasks. Reading the L1 translations of the target words in alphabetical order, together with an example sentence, was the first assignment. Writing a composition utilising the target words was the second assignment. The third activity required you to read the material that contained the target words and reply to the comprehension questions that were given. The first task was given to one experimental group, the second to the other experimental group and the third to the control group. Sentence-making and composition tasks showed no discernible differences in the data analysis. However, the TFA predictions were supported by the score decline in the second test (composition using 10 target words).

The researchers (e.g., Hu & Nassaji, 2016) came to the conclusion that ILH and TFA were unable to anticipate vocabulary acquisition accurately. However, after more research, they concluded that the composition task with the highest index in the TFA model's generation component would help with vocabulary growth and produce more intricate linkages between form and meaning. From a pedagogical perspective, they asserted that retention and generation might be more crucial in word activities (Rahmanian & Soleimani, 2018).

In another study by Zou and Xie (2018) conducted in a Chinese setting, 105 English speakers were randomly divided into three teams and assigned the task of learning 40 target words using one of three methods: a non-personalised approach, a personalised approach guided by some of the feature analysis's techniques, and a customised approach led by all of them. While all three approaches were found to help promote vocabulary learning in the trial, the individualised approach guided by the entire TFA list was the most effective. The researchers

concluded that comprehensive learning theories should be used to guide the design of e-learning systems. Although the second study focused only on the application and effectiveness of TFA, it confirmed the value of incorporating TFA into the design of learning materials.

2.1.3.4 Key Issues of Output-based Approaches to Vocabulary Knowledge

Pushed output activity has shown effectiveness in SLA, such as with relative clauses (Izumi, 2002) and the Spanish future tense (Russell, 2014), while the definition task was found to be more effective for EFL vocabulary knowledge in the research conducted by Bao (2018). However, it still requires further attention from researchers to determine whether writing, as a prominent form of language production, has similar effects on EFL vocabulary. Moreover, the task of writing a composition and incorporating target words has an ILH index of 3 (Hulstijn & Laufer, 2001). Additionally, writing sentences with bilingual dictionaries has an ILH index of 4 (Alcaraz Marmol & Almela Sanchez-Lafuente, 2013), which is among the highest indexes among the other activities. The composition task with the highest index in the TFA model's generation component could aid vocabulary development and result in more complex relationships between form and meaning (Rahmanian & Soleimani, 2018). Thus, while writing to learn vocabulary has a strong theoretical background, its effect on EFL vocabulary knowledge still requires further investigation.

2.1.4 Importance of Collocations in EFL Writing

The quality of collocation use is an important index of quality writing (Chang et al., 2008; Daskalovska, 2015; Siyanova-Chanturia, 2015; Zou, 2019). Words are not separate linguistic units but are part of several interconnected systems (Nation, 2022). Collocation per se is important because the way words combine in collocations is fundamental to all languages, as the lexicon is not arbitrary (Lewis, 2000, pp. 53–56). The predictability of collocations helps students anticipate their usage, and two, three, four and even five-word collocations account for "70% of everything we say, hear, read or write in some form of fixed expression". Moreover, the use of "ready-made language" enables native speakers to think more quickly and communicate more efficiently (Lewis, 2000, p. 54), and it serves as a hallmark of near-native-like language capability (Chang et al., 2008). A more recent study by Siyanova-Chanturia (2015) also confirmed that the appropriate use of collocations is one of the key prerequisites for proficient language use, as multi-word speech accounts for 52.3% and 40% of written discourse, respectively.

In terms of acquiring English language skills, mastering the language means not only knowing its lexical meaning but also understanding its collocation rules and the specific contexts in which collocations can be used (Zou, 2019). It is reasonable to presume that learners are more likely to know or employ collocations the more L2 vocabulary they have learned (Fan, 2009). Learning collocations should take a central place in vocabulary studies because they are crucial for proper and fluent language use (Daskalovska, 2015). Corrective and relevant collocations facilitate the fluidity and idiomaticity of L2 production (Wang &

Zhou, 2020). However, a significant number of errors are collocation errors, which are considered problematic to L2 learners (Sun & Wang, 2003; Fan, 2009), even among advanced L2 learners (Chan & Liou, 2005; Laufer & Waldman, 2011; Wang & Li, 2018; Wang & Zhou, 2020).

2.1.4.1 Definitions of Collocation

In the 1950s, J. R. Firth described collocation as "the way words join in predictable ways" or "the company words preserve their links with other words" (Lewis, 2000, p. 42). Collocation is also defined as "the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991, p. 170). Another common definition is "words which are statistically considerably more likely to emerge than random chance predicts" (Lewis, 2000). The researcher used Lewis's straightforward definition of a collocation— "two or more words that tend to occur together"—as it best encapsulates the phenomena that are the subject of this study.

2.1.4.2 Classification of Collocations

Following Benson et al. (1997), collocations can be classified into grammatical and lexical categories. Grammatical collocations (such as an infinitive or clause) typically consist of a main open word (noun, adjective or verb) plus a preposition or a specific structural pattern. The most common grammatical collocations include the following: (1) noun + preposition/to infinitive/that clause (e.g., key to, fact that), (2) preposition + noun (e.g., in detail), (3)

adjective + preposition/to infinitive/that clause (e.g., conscious) and (4) verbs combined with prepositions, infinitives with to, verbs without to, verb forms in -ing, and that clauses in various ways.

Table 6. Examples of collocations (Benson, 1985)

Type	Examples
Grammatical Collocations:	
• verb + preposition	• *(to) get at, (to) go for*
• adjective + preposition	• *different from, curious about, full of*
• adjective + preposition + preposition	• *fed up with*
• preposition + noun	• *for sale, on time*
• dative movement transformation	• *She sent the book to him/ She sent him the book.* • *He described the book to me/ *He described me the book.*
Lexical Collocations:	
•verb + noun (pronoun, prepositional phrase	• *(to) reach a verdict, (to) launch a missile, (to) lift a) blockade, (to) revoke a license*
•adjective + noun	• *reckless abandon, sweeping generalization*
•noun + verb	• *adjectives modify, alarms go off*
•noun + *of* + noun	• *a bunch of flowers, a piece of advice*
•adverb + adjective	• *deeply religious, fiercely independent*
•verb + adverb	• *(to) apologize humbly, (to) affect deeply*

Meanwhile, infinitives and clauses are typically absent from lexical collocations, which typically comprise open words (nouns, adjectives, verbs or adverbs). Benson et al. (2010) identified six major categories of lexical collocations: (1) verb + noun/pronoun/prepositional phrase (such as writing music and carrying out chores), (2) adjective + noun (such as tiny drop), (3) noun + verb (such as bees buzz), (4) noun + noun (such as a bouquet of flowers), (5) adverbs + adjectives (such as deeply thankful) and (6) verb + adverb (e.g., sharp increase).

Table 7. Classifications of lexical collocations by Benson et al. (2010)

Types	Examples
Verb + noun/pronoun (or prepositional phrase); with the verb denoting *creation* and/or *activation*	*Come to an agreement, make an impression, compose music*
Verb + noun; with the verb denoting *eradication* and/or *nullification*	*Reject an appeal, lift a blockade, break a code*
Adjective + noun	*Strong tea, warm regards, reckless abandon*
Noun + verb	*Adjectives modify, alarms go off, bees buzz*
Noun + of + noun	*A herd of buffalo, a pack of dogs, a bouquet of flowers*
Adverb + adjective	*Deeply absorbed, strictly accurate, sound asleep*
Verb + adverb	*Affect deeply, amuse thoroughly, argue heatedly*

However, it might not be easy to distinguish between idioms, collocations and free combinations. According to Howarth (1998), these categories of lexical pieces are structured on a gradient from pure idioms to free combinations, with no real boundaries.

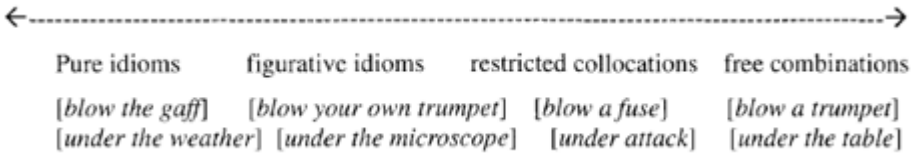


Figure 1. Howarth's model of continuum (Howarth 1998, p. 28)

Another way to categorise collocations is by using a continuum of free collocation, restricted collocation and idioms (including figurative and pure idioms) in the figure 1. This continuum is derived from criteria such as restricted collocability, semantic specialisation and idiomaticity. Free collocations, also known as open collocations or free combinations, are employed in their literal sense. The constituent words do not necessarily have to be used

together; they can coexist freely with numerous other lexical items that share their semantics (for instance, purchase a book, TV, piano, etc.).

At the other end of the continuum, figurative and pure idioms are the most ambiguous and fixed. Figurative expressions have a present literal interpretation of their overall metaphorical meaning (e.g., “blow your own trumpet”), whereas pure idioms do not have a logical literal meaning (such as “having a chip on one’s shoulder”), and their meaning cannot be inferred from the meanings of their constituent parts.

Restricted collocations, located in the middle of the continuum, typically involve one element, referred to as a node word (Stubbs, 1995). These collocations are used in specific contexts and are accompanied only by specific types of collocations. There are three main categories of restricted collocations. First, some words are almost exclusively used in conjunction with just one or two other phrases or a small group of words due to their incredibly limited and specialised meanings (e.g., white hair). Second, collocates may not always encompass all lexical items or sets semantically compatible with a word’s primary meaning, as certain words are used metaphorically. Third, when a word is used in its delexicalised form, the range of its possible collocates seems to be limited for no apparent semantic reason.

Verb-noun lexical collocations have been determined to be the most difficult for learners to learn among the many forms of collocations (Chan & Liou, 2005; Huo, 2014). Through

lexical semantic analysis, Liu (2002) looked at verb-noun miscollations in Chinese learners' essays and found that 87% of them were caused by verb-noun miscollations, with 93% of them coming from improper usage of verb collocates.

In most cases, investigations into L2 collocations have been focused on specific structures (Fan, 2009). The current study focuses mainly on lexical collocations, especially verb collocations. As a result, lexical collocations are considered in this study.

2.2 Corpus-based Approaches to Language Learning

According to limited meta-analyses of corpus use in language learning (Mizumoto & Chujo, 2015; Boulton & Cobb, 2017), target language skills (e.g., listening, reading, speaking and writing) and language aspects (e.g., vocabulary knowledge) have attracted more interest in the field. Specifically, writing skills have been extensively explored, while lexico-grammar knowledge and vocabulary knowledge have been frequently researched in corpus-based language learning.

2.2.1 What is a Corpus?

According to McCarthy and O'Keeffe (2010), a corpus is a collection of real linguistic data that dates back to the 13th century. The creation of a corpus originated from the desire of biblical researchers to understand linguistic phenomena in important texts or groups of less important texts. They were the first to achieve this goal by manually indexing lines of words.

A thorough way of finding terms with citations of their passages' locations was by screening concordance lines and indexing. Modern corpora, which involve gathering actual data for linguistic studies, were first developed by American structuralists such as Harris, Fries and Hill (McCarthy and O'Keeffe, 2010, p. 33), who committed themselves to making exact language data the focal point of their studies. The advancement of computer technology led to the creation of electronic corpora, which played a crucial role in compiling dictionaries. In the late 1950s, the first concordances produced by computers began to appear. The initial corpus was used for English language instruction at Aston University (Birmingham, UK) late in the 1960s. Early in the 1970s, it was adopted in English for Special Purposes courses at Nottingham University (Nottingham, UK). The idea of using KWIC to replace catalogue-indexing cards and automate subject analysis caught the attention of library and information scientists. Throughout the 1980s and 1990s, corpora developed into various scales and became valuable tools for linguists and applied linguists. However, corpora can be enormous, such as COCA, which contained 1.1 billion words between 1990 and 2010, or tiny and designed for specific uses, such as a learner corpus at a particular institution aiming to replicate the usage of the local tongue. The language found in corpora typically comprises spoken and/or written material that occurs organically when speakers are engaged in activities other than simply explaining how a device operates. Corpora aim to represent all or a portion of a language; they are not merely collections of information.

The study of language data on a massive scale is known as corpus linguistics, and it typically involves computer-assisted analysis of vast collections of written or spoken material

(McEnery & Hardie, 2012). Corpus linguistics employs a range of techniques and approaches for analysing languages. Techniques such as concordance lines can completely reshape the way we approach language study. Corpus linguistics presents challenges, as it deals with machine-readable texts that provide a suitable foundation for researching particular linguistic queries. Concordances and frequency statistics derived from corpora enable both qualitative and quantitative research. Frequency data can be generated for quantitative studies, such as identifying the most common words in the target corpus or the prevalence of specific phrases (e.g., two- and three-word phrases). In addition, it makes it simple to locate numerous examples of real language used by actual speakers. Corpora also facilitate qualitative research, allowing for diachronic syntactic comparisons to examine how sentences are combined and used in different contexts. As a result, it will be simple for you to analyse all the terms required to bring your investigation to a close.

2.2.2 Corpus-based Language Studies for EFL Learners

Corpus-based studies frequently utilise corpus data to investigate theories or hypotheses. In contrast, corpus-driven linguistics asserts that assumptions about a language should be derived from the corpus itself (McEnery & Hardie, 2012, p. 6).

Johns (1991) provided two groundbreaking examples, “persuade vs convince” and “varieties of should”, marking the beginning of the transition in English language education and learning towards data-driven learning (DDL). Since then, several scholars have investigated

the usage of corpora, including the use of concordance lines as writing-process feedback (Luo & Liao, 2015). However, because of things like the lack of corpus learning in teacher training, teachers' associations of corpus linguistics with research activities, and the perception that using corpus technology is difficult, the teaching community has remained mainly unaware of the corpus-based linguistic approach. Some innovators (Gaskell & Cobb, 2004; Yoon, 2008) have integrated corpus data into regular English classes. In research by Gaskell and Cobb (2004), 20 adult Chinese EFL learners—the majority of whom had bachelor degrees—participated. Although they were considered lower intermediate English learners, who enrolled in a three-class-per-week English writing course. The participants completed 10 assignments over a 15-week semester. The assignments were submitted as a first draft in the first week (with accompanying peer input), and modifications were due in the second week. The instructor provided feedback using online concordances from the lextutor website. The researchers observed a substantial decrease in word order, capitalisation/punctuation and pronouns when analysing the pre- and post-test writing samples. However, they recommended including a control group and providing longer training sessions to further validate the findings. In a different approach, Yoon (2008) conducted a qualitative case study over an extended period with six L2 EFL postgraduate writers (masters or PhDs) enrolled in an EAP writing course. Each week, the participants were instructed to email search results from their Collins COBUILD Corpus searches related to their writing issues. Before the in-class sessions, the instructor provided feedback on any writing errors, incorporated these results into handouts and advised the students on how to perform corpus research. It was determined that corpora use improved language awareness

and helped students address their language issues, fostering independence and confidence in writing.

In Karras' (2016) study, 100 participants received instructions on how to use DDL for vocabulary acquisition. The experimental group was taught how to use the English language concordance interface of the Compleat Lexical Tutor in the ICT (Information and Communications Technology) lab and completed weekly DDL vocabulary handouts using the concordance generated from each vocabulary item. The control group, on the other hand, did not get DDL instruction or make use of concordance as part of their weekly vocabulary development. To assess the effectiveness of the two vocabulary learning strategies, weekly vocabulary tests were administered.

Seidlhofer (2002) popularised the phrase "learning-driven data" following the creation of learner corpora, which are digitised collections of authentic texts written by language learners. Learner corpora have become increasingly common in language teaching and learning (Cotos, 2014). Cotos (2014) conducted a study involving 31 overseas graduate L2 students with TOEFL scores ranging from 83 to 107. The study compared the use of a native speaker's corpus alone and when combined with a learner corpus. The analysis aimed to examine the potential of local learner corpora in language-learning contexts. Writings from the course assignments of the participants were gathered to create a local learner corpus. Data were collected before, shortly after and four weeks after the experiment to track changes in language use. For each individual, the results on the use of adverbs were favourable.

Using a corpus-based strategy offers several benefits (Bowker, 2018). Corpus data can provide objective information that is free from the influence of prior beliefs. It also enables substantial data to be collected more effectively than manual methods. Researchers can analyse concordance lines to identify patterns, such as the frequency of specific words in particular contexts or themes. However, Bowker (2018) noted that analysing pragmatic devices can be challenging, and corpora represent what has been spoken or written rather than the full range of linguistic possibilities. Additionally, Charles (2012) suggested that creating a corpus can help students learn more useful collocations. In the study, 50 advanced-level EAP students were instructed to create a specialised corpus comprising 10–15 research articles. The participants reported benefitting from this corpus creation process, as it exposed them to appropriate vocabulary and collocations relevant to their field of interest.

Unlike the above international studies, most corpus studies in China have been inclined towards corpus-driven research, where assumptions about language are derived from the corpus data itself (Wang & Zhou, 2020). However, there have been fewer corpus-based research studies (Luo & Liao, 2015). China's interest in corpus linguistics has significantly increased over the past 20 years, which has led to a focus on language description in corpus-based studies. Lexis, syntax and discourse are the primary facets of language examined in this field (Liang et al., 2010). Lexically speaking, research begins by looking up words and collocations in concordance lines and recurring patterns are summarised. By examining the semantic tendencies of a language's collocations, the meaning can be translated after learning

the features and laws of language production. For instance, Wang and Zhou (2020) collected over 18,000 words from the verbal productions of 73 senior English majors to study verb–noun collocations in spoken English. They concluded that approximately eight verb–noun collocations, mainly verb + objects, were accepted per 100 words, with around 30% of them being incorrect. The CLEC was manually error-tagged into 61 types of errors, six of which were collocation errors: "noun–noun, noun–verb, verb–noun, adjective–noun, verb–adverb and adverb–adjective" collocation errors (Yang et al., 2005). Sun (2017) analysed the productive English use of non-English majors at a prestigious university in China and found that some outstanding students reached Level B2 (CEFR) in productive vocabulary, while most students were at a level comparable to Level B1. Zou (2019) found that verb–noun collocation errors accounted for the largest proportion (that is, 57.14%) of the total collocation errors.

In summary, international corpus-based research has primarily focused on advanced EFL learners (Gaskell & Cobb, 2004; Yoon, 2008; Cotos, 2014), while corpus research in China has leaned more towards corpus-driven approaches to uncover language features and phenomena, such as verb–noun phrases (Wang & Zhou, 2020). Only a few studies in China can be categorised as corpus-based language research. For example, Luo and Liao (2015) investigated the impact of corpora on rewriting essays written in English. The study included training on hands-off to hands-on DDL, followed by writing practice, after which the participants completed a 6-point Likert scale questionnaire. The second part involved composing an essay, guided feedback from teachers, learners' independent error repair, and

teachers' evaluation and comments because the research was focused on editing essays. Luo (2016) conducted a subsequent empirical study to compare the effectiveness of the BNC website and the Baidu search engine in supporting DDL. The study aimed to evaluate the benefits of direct referencing tools commonly used by Chinese EFL learners. Pre-treatment, treatment, and post-treatment stages made up the data collection process. Both the experimental group and the control group finished a pre-writing test during the pre-treatment stage. The participants in both groups were given instructions to create compositions and make changes using a variety of referencing tools depending on the regions the researcher had highlighted during the treatment stage. Finally, in the post-treatment phase, all participants were given a post-writing test in class with the same time constraints and subject matter as the pre-writing test. Collocations, especially those involving verbs, have received more attention in recent years (Daskalovska, 2015; Vyatkina, 2016).

2.2.3 Corpus-based Collocation Learning

Collocations have been the subject of study in corpus-based activities (Lewis, 2000, p. 48). Traditional vocabulary instruction often focuses on individual word learning without considering the context. For instance, learners may rely on paraphrasing for common word suffixes while translating the lexicon into their first language, which emphasises a high-information word (Lewis, 2000, p. 33). However, linguistic input suggests that students need to modify their existing knowledge. Learning involves reorganising prior examples of interlanguage and is not simply an additive process. Instead of correcting students' grammar

or introducing many new words, some of which may be uncommon or difficult to understand, Lewis (2020) suggested that teachers provide practical collocations to students to help them remember these structures.

In a subsequent study, Zou (2019) analysed the data from CLEC (Yang et al., 2005) and found that verb–noun collocation errors accounted for the largest proportion (that is, 57.14%) of the total collocation errors. Building on Nesselhauf's (2005) framework, Wang and Zhou (2020) further categorised verb and noun collocations into six types: *verb + object*, e.g., achieve success; *verb + object + complement*, e.g., make life meaningful; *verb + indirect object + direct object*, e.g., give you an example; *verb + preposition + object*, e.g., participate in the talent show; *verb + object + preposition + object*, e.g., give a chance to people; *verb + object + infinitive*, e.g., encourage people to.

Boers et al. (2014) found that intermediate-level L2 learners face particular challenges in learning verb–noun collocations, such as "make a mistake". For instance, the productive knowledge of verb-noun collocations among lower and upper intermediate groups of EFL learners was not significantly different, according to Laufer and Waldman's (2011) research. Learners often substitute the verb when equivalent first-language nouns clash with a different verb, leading to unexpected choices (*do a mistake). Such substitutions are likely influenced by interference from the mother tongue (Nesselhauf, 2005; Yamashita & Jiang, 2010).

Miscollocations in verb–noun combinations can be attributed to three main reasons: L1

interference (Chan & Liou, 2005; Nesselhauf, 2005; Yamashita & Jiang, 2010; Zou, 2019), misuse of delexicalised verbs (Chan & Liou, 2005) and lack of awareness of collocational constraints in synonyms and hypernyms, which are semantically similar lexemes (Chan & Liou, 2005). When collocations contain words that learners are already known (e.g., have + dreams), they may pay attention to the word combination itself. Given the frequent occurrence of verbs included in numerous collocations, learners are likely already familiar with these verbs and may not pay much attention to them. Additionally, in verb-noun collocations (such as make and mistake), the noun often carries most of the semantic weight because the verb is frequently a multipurpose, semantically ambiguous word. This reduces the need to focus on the verb to interpret the statement. The limited contribution of the verb to the semantics of some collocations is evident in near-equivalent pairs, such as *you are lying* and *you are telling lies.*

The lack of distinctiveness among verbs is another reason why learners mistakenly substitute them (Boers et al., 2014). For example, the verbs create and do in “make a mess” and “do damage” and tell and say in “tell the truth” and “say a prayer” are examples of collocations that learners can employ as synonyms.

For these reasons, learners likely need multiple encounters with verb–noun collocations to establish a strong association between the verb and the noun (Boer et al., 2014). Learners are needed to correctly match sets of verbs and nouns in the normal practice styles found in textbooks and other educational resources. However, even with corrected feedback, there is

still a chance that learners may make incorrect connections during matching tasks, which can negatively affect their memory. In four small-scale trials (total n = 135), the learning gains from verb-noun matching activities are compared to those from a format in which learners are given with the target collocations as full wholes. Small pre-test to post-test gains were always the consequence of learners replacing originally correct choices with distractions from the activities. As opposed to tasks where learners engaged with collocations as complete units, matching exercises were where this negative impact was most frequently seen.

In a comparative research study, Daskalovska (2015) investigated verb–adverb collocations. Two groups of 46 first-year English majors were formed, with the control group (25 students) receiving conventional exercises and the experimental group (21 students) using online concordance tools that provided unrestricted access to the BNC. Pre- and post-test analyses revealed that the corpus-based activities for learning collocations had some advantages over conventional methods after four activities. Additionally, 11 North American undergraduate students with intermediate foreign language proficiency were examined by Vyatkina (2016) and their reactions to the use of DDL with German verb–preposition collocations (B1, according to CEFR). The participants engaged in one computer-based hands-on lesson and one worksheet-based hands-off activity. It was found that all learners benefited equally from both settings. Furthermore, the participants' overall language proficiency improved, and they expressed a desire to continue using this strategy.

However, traditional activities for learning collocations typically involve matching, gap-

filling or multiple-choice exercises, which are often presented as isolated tasks without further practice of using these collocations in meaningful contexts within language courses. The treatment of collocations does not guarantee that students will learn and remember the collocations effectively because these exercises require little time and effort to complete and lack depth of processing, which is necessary for successful learning and retention of language input (Daskalovska, 2015).

2.2.4 Corpus Use and Learner Autonomy

Autonomy has become an essential goal of education in the current learning environment, and it supports lifelong learning and the increasing demand for distance learning (Spratt et al., 2002). During the recent COVID-19 pandemic, when students stayed at home and watched teachers' live broadcasts without adult supervision, this aim for learner autonomy received more attention than ever. Students were expected to complete their schoolwork independently.

Learning autonomy refers to the capacity that students have or exhibit in various situations (Benson, 2001). While experts agree that autonomy cannot be taught or acquired, educational programmes can foster and promote learners' autonomy growth (Benson, 2001). The Modern Languages Project of the European Council played a significant role in shaping Holec's definition of autonomy as "the capacity to take charge of one's learning" (Benson, 2007).

Given the importance of learner autonomy, especially in the current situation (Spratt et al., 2002), Benson (2001) and Dörnyei (2001) provided their own insights into facilitating the development of learner autonomy. Benson (2001) proposed six alternative strategies for nurturing and growing learner autonomy. He believed that learners already possess and exhibit autonomy to some extent in various circumstances. While it is widely acknowledged that learner autonomy cannot be taught or learned, developing autonomy is considered an educational endeavour that fosters the growth of learners' autonomy. These programs can be broadly divided into five categories: learner-based, classroom-based, curriculum-based, technology-based and teacher-based approaches. These approaches are closely related to DDL or corpus-based language learning. Although Dornyei (2001) expressed doubts about the theory of learning autonomy, he also provided strategies for fostering autonomy. He argued that educators often find it challenging to implement the desired changes in educational institutions. Therefore, there is growing focus on preparing students to succeed regardless of the quality of their education. Meanwhile, students who can learn autonomously may become more proficient. Dörnyei (2001) concluded that altering the teacher's role and enhancing learner involvement in the organisation of the learning process are crucial. Students should have a voice in various aspects of language learning and be given significant control, such as assigning course tasks to elected student councils in class, as choice is central to responsibility since it allows students to have a sense of ownership over their learning experience. Peer teaching and project work are also language learning activities that can increase the proportion of student teachers. Additionally, self-assessment methods can help learners become more aware of their learning successes and failures and offer them a real

sense of involvement. Second, the transformation of the teacher's role is crucial. Teachers need to shift from being mere instructors to becoming guides and instructional designers who help learners find and develop their own understanding of the world while enhancing their involvement in the learning process.

All the strategies proposed above for learner autonomy development have a connection to the use of corpora and corpus-based studies. With concordance tools, even young learners can achieve learner autonomy as they can search any electronic text bank to create texts with more effective collocation searches tailored to their needs (Lewis, 2000, p. 42). Since Johns first suggested the idea of DDL, Mizumoto and Chujo (2015) observed that practitioners and researchers have become interested in it because it is an excellent tool for empowering language learners and assisting them in becoming autonomous learners.

Although evaluating autonomy has received increased attention, creating a scale or questionnaire to assess learner autonomy can be challenging as administering such tests may inadvertently demotivate students. Autonomy manifests differently in different individuals, as some learners excel at creating effective plans while others maximise their chances to converse with language experts and practice their communication skills. Furthermore, linguistic awareness and learning motivation are directly linked to autonomy. While it may sound challenging, Benson (2001) proposed a temporary solution, as academics often categorize the idea of learner autonomy into various groups based on other theories. These categories include choosing methods and procedures, establishing objectives, specifying

content and progression, monitoring the acquisition process and assessing learning outcomes (Spratt et al., 2002). Other aspects involve recognising learners' responsibilities for past, present and future learning, both inside and outside the classroom (Sakai & Takagi, 2009), and students' opinions of their participation in the learning process, their capacity for autonomous action, their use of metacognitive strategies, their level of motivation and their actual autonomous practices (Karabiyik, 2008). Learner autonomy also intersects with other ideas, such as learning strategy theory (Macaskill & Taylor, 2010) and preparedness for self-directed learning (Dafei, 2007; Zhang & Li, 2004).

In the early 2000s, autonomy gained increased attention (Benson, 2001). In China, the study of learner autonomy has been of longstanding interest due to the differences in the Chinese learning environment compared to Europe, where the notion of autonomy was first developed. Mandarin Chinese is the only official language in China, and other languages are typically considered dialects. As a result, the country is neither multiracial nor multilingual, and English is not used in daily life but rather taught as a foreign language in schools. Additionally, due to the influence of Confucian culture, classroom activities tend to be more teacher centred. Moreover, the highly competitive College Entrance Examination has made English study exam oriented. To fulfil the deadlines for passing the required tests and acquire a substantial vocabulary for reading and writing, more effective learning methods are needed. However, due to time constraints, learning new language is frequently not exposed to enough people for this to happen naturally (Cobb & Boulton, 2015).

Zhang and Li (2004) compared Chinese and Western European students' autonomy using questionnaires and found that Chinese students had less autonomy compared to their Western European counterparts. Dafei (2007) conducted surveys and interviews among 129 college students to study the relationship between learner autonomy and English proficiency and found a strong positive relationship between the two variables. Guan (2013) explored the theoretical foundations of DDL and its potential applications in the classroom. Based on his analysis, it was concluded that such applications could significantly enhance student learning autonomy and transform the traditional teacher-centred environment. Although research has shown a positive association between language proficiency and learner autonomy, there are still significant gaps in the existing literature. For example, in Deng's (2007) study, 129 non-English majors participated in a questionnaire about learning autonomy, and the results were compared to their Practical English Test for Colleges (PRETCO) results, of which 496 students passed and 254 failed. Nevertheless, only data from 129 non-English majors were used for the analysis. Another study by Sakai and Akiko (2009) involved approximately 721 Japanese university students from various majors, but their proficiency was assessed using a vocabulary levels test.

By fostering noticing and consciousness-raising, which enhances autonomy, DDL—one strategy for incorporating corpora into English language teaching—helps students become better language learners outside the classroom (Boulton, 2010). Corpus-based learning is a constructivist and inductive method that promotes cognitive and metacognitive growth, critical thinking, noticing abilities, linguistic awareness and sensitivity to actual texts,

autonomy and lifelong learning (Cobb & Boulton, 2015).

In summary, the strategies proposed by Benson (2001) and Dörnyei (2001) for learner autonomy development are closely related to corpus use and corpus-based studies. Many scholars (e.g., Spratt et al., 2002; Karabiyik, 2008; Sakai & Takagi, 2009) have attempted to develop rigorous questionnaires to measure learner autonomy and conducted correlational research on other learner factors such as English proficiency (Deng, 2007) and learning strategies (Macaskill & Taylor, 2010). However, empirical studies on the effects of corpus-based studies on learner autonomy in the English classroom remain relatively scarce based on the researcher's knowledge in the present study.

2.2.5 Corpus-based Language Pedagogy

The incorporation of corpora into classroom pedagogy has been extensively researched, highlighting its benefits and drawbacks (Cobb & Boulton, 2015; Gilquin & Granger, 2010; McCarthy & O'Keeffe, 2010; Timmis, 2015). One major advantage is the authenticity that corpora provide, exposing students to real language instances and thousands of examples of specific language phenomena. By comparing instances produced by native speakers in corpora with learner corpora (including error-tagged learner corpora), students better understand interlanguage aspects and improve their language production. The use of corpora in the DDL technique makes the learning process more engaging by allowing students to investigate natural structures. This trial-and-error process enables language learners to

observe, analyse, hypothesise, create rules, gain insights and deepen their understanding. Moreover, including corpora can significantly increase the impact on language-learning processes (Johns, 1991) because data is the main source and inductive language education is the method. However, the instructor's knowledge may be challenged because they cannot predict the rules or patterns that language learners will encounter while exploring various corpora. This aspect of uncertainty draws a lot of attention. Furthermore, when students take the initiative to generate new questions after consulting corpora, the teacher's position tends to become more of a director or coordinator (Johns, 1991). The teacher starts questioning what is irrelevant to the curriculum or texts, and the impact of grammar becomes questionable as a result of this change. This is because corpora compel students to explore and inquire, which is the beginning of learning, whereas traditional grammar-based techniques prescribe how and what should be learned. Additionally, corpora can support long-term metalinguistic and metacognitive awareness (Timmis, 2015). The information in corpora exposes students to words in various situations, which expands their understanding of vocabulary through collocations and registers.

However, according to Cobb and Boulton (2015), the use of corpora in reference activities can be complex and open-ended because students work as explorers to elicit their goal words or meanings. As a result, learners need to receive practical training and practice. Moreover, specific corpora require payment for access, and the concordance lines provided in corpora are often shorter than complete passages or paragraphs, which can occasionally lead to decontextualisation (McCarthy & O'Keeffe, 2010). Selecting useful concordances from these

shorter lines can prolong the time required for classroom activities.

To facilitate the use of corpora in the classroom, researchers have developed corpus-based language pedagogy (CBLP). Moon and Oh (2017) drew on Flowerdew's (2009) paradigm of corpus-based activities, known as the "4 Is"—illustration, interaction, intervention and induction—which served as the foundation for the instructional approach of DDL. Moon and Oh (2017) reported the cognitive and affective benefits of DDL for secondary-level Korean EFL learners. Their instructional approach involved helping students notice and unlearn their tendency to overgenerate the word *be* (e.g., "She is go to university.") by comparing native English-speaker and learner corpora through guided induction. The instruction comprised several steps. In the first step, students examined hand-picked paper-based data selected by the professors, focusing on concordance line patterns. During the second step, they discussed their perspectives and shared their observations in pairs and groups based on the prompt questions on their worksheets. In the third step, the teacher provided broader induction cues where appropriate instead of explicitly explaining regulations, assisting lower-level students with any unfamiliar words and directing their attention to the target patterns. This provided a more decisive direction for discovering the discrepancy between learner data and native-speaker data in the fourth step. Individually, the students developed their hypotheses, presented their findings to the class and, if required, updated them based on the group's comments. Finally, they applied the learned guidelines in practical exercises by rewriting incorrect sentences.

More recently, using Shulman's concept of pedagogical content knowledge, Ma et al. (2022) investigated how TESOL (Teaching English to Speakers of Other Languages) teacher trainees acquired their corpus literacy and corpus-based language pedagogy (CBLP). The study team created a four-step corpus-based lesson plan based on Gass' (Gass & Selinker, 2001) L2 acquisition model, including testing student understanding (e.g., looking for grammatical errors), hands-on corpus search by students (e.g., looking for language patterns), inductive discovery by students (e.g., summarising language patterns), and output activities (e.g., using newly learned terms actively) are some of the learning strategies used. This design strategy combined theoretical understanding about teaching and learning languages with real-world corpus implementations in the classroom. The study was conducted with secondary school students in Hong Kong and mainland China, and the feedback was encouraging.

Based on the literature above, CBLP has been found to be applicable at the secondary school level, as demonstrated in studies conducted in Korea (Moon & Oh, 2017), Hong Kong and Mainland China (Ma et al., 2022). However, a question arises regarding its suitability for higher vocational institutes. Considering the specific needs of students in these institutes and the increased attention they have received in recent years due to the development of China's economy and society; it becomes important to explore the potential application of CBLP in this context.

2.3 DIY Corpus

2.3.1 What is DIY Corpus?

In contrast to large general corpora such as BNC or COCA, some recently collected corpora are compact and DIY (Charles, 2018, 2019; Smith, 2020; Zhang et al., 2017). DIY corpora are a small-scale collection of electronic texts created by professors or students. They are also referred to as local corpora, disposable corpora or personal corpora (Charles, 2018, 2019). Existing research (such as Charles, 2018, 2019) shows that a DIY corpus is typically compiled by high proficiency students who have the motivation and skills for self-learning.

The process of creating a DIY corpus has been outlined in the existing literature (e.g., Charles, 2018, 2019). The first stage is selection, where goals, audience and type of material to be retrieved need to be considered before collection. For example, a collection of academic publications in a relevant field would be beneficial for seminars on advanced academic writing, while online sources or student textbooks could be used for lower-level pupils. The size of the corpus and the language standard are also important factors to consider in the selection stage. A corpus with 50,000 to 250,000 words is generally considered sufficient for specific purposes and ensuring that academic articles are written by native English speakers is crucial for quality. Once selected, the articles are converted into plain text format for the corpus programme to read (such as *AntConc* (Anthony, 2020)), and the files are checked to ensure that the conversion is successful. The ".txt" folder is then renamed to reflect the name of the corpus, and cleaning the corpus by removing irrelevant content is an optional step that involves manual review.

2.3.2 Empirical Studies on DIY Corpus

DIY corpora offer several benefits over large general corpora, including the ability to overcome challenges posed by irrelevant information and provide more tailored resources. Students using DIY corpora can take more ownership of their writing and adjust the corpus to their specific needs, making it suitable for multidisciplinary learners (Charles, 2012). When students study English for a specific reason not met by the broad corpora already built, teachers must be able to respond to their demands (Charles, 2018). However, creating corpora can be incredibly time-consuming, and specific, exceedingly rigorous or precise queries might yield few (or no) examples (Charles, 2012).

Despite the challenges involved, DIY corpora have demonstrated promising results in academic language instruction. In a 6-week academic writing course, Charles (2012) introduced the idea of the DIY corpus by educating advanced-level graduates to create and analyse their own discipline-specific corpora. Positive reviews were indicated by the students' feedback. Additionally, 40 international graduate EAP students used their own DIY corpora over an extended period of time, according to Charles (2014). Using their own corpus to check their grammar and lexis while producing and revising, 70% of participants who responded to an email survey sent a year after the course felt their academic writing had improved. Zhang (2017) also included a corpus-based project on academic writing in the Academic English course, in which 35 postgraduate medical students in their first year were taught how to gather and utilise their own corpora of research papers for academic writing. The participants' DIY corpora, which averaged 309,413 words, and research papers, which

averaged 6,223 words each, were from 13 different disciplines. According to a questionnaire survey, 94.2% of the participants had mastered and applied their self-compiled corpora both inside and outside of the classroom. Most participants used their corpora to verify lexical collocations or grammatical usage when drafting and rewriting their papers. As a result, 85.7% had active and favourable attitudes towards using corpora, and 95% found them helpful. Furthermore, it was hoped that using corpora would contribute to their improved academic writing. Nevertheless, a delayed survey conducted six months after the project's completion revealed that only 5.7% of the respondents were still frequent users. Charles (2012, 2014) also identified three issues in their studies. The participants reported difficulties in using the corpora software, thus requiring training and instructions. Second, DIY corpora's unquestionably modest size (compared to huge general corpora) may have left them unrepresentative and unable to solve specific issues. Furthermore, several publications prioritised ideas or academic importance over linguistic excellence due to their small size, which created reliability problems.

DIY corpora can also be utilised for vocabulary research. Six accounting and finance for international business students took part in an initial inquiry led by Smith (2020). In an EAP curriculum at a public university in the UK, the students created DIY corpora, and it was discovered that they needed greater motivation when using their DIY corpora. The study involved 94 students (mainly from China) during an 11-week second semester. Four class groups, two for hands-on DDL and two for hands-off DDL, were formed from the participants. The hands-on groups built their corpora, used them to construct lists of

accounting and financial terms, and then added those terms to the students' vocabulary portfolios. Performing this task in the computer labs took approximately 20 minutes per week. In contrast, the hands-off groups received weekly lists of specialised financial and accounting language generated from corpora created by the author, similar to those used by the hands-on groups. The lecturer's PowerPoint presentations, the seminar participants' notes from the discussion, old test questions and other materials were used to create the participant corpora. Comparing the pre-and post-test results revealed that the hands-on groups showed significantly greater improvements in their domain vocabulary knowledge compared to the hands-off groups. The questionnaires also indicated that the students found vocabulary learning methods helpful.

In summary, the empirical research on the incorporation of DIY corpus has mainly focused on students in academic courses, such as EAP (Charles, 2012; 2014; Zhang, 2017; Smith, 2020). As advanced students can develop their own corpora and become less dependent on native speaker professors, proofreaders, and editors to improve their texts, the participants in these studies were predominantly high-level English users (Charles, 2019). However, Charles (2012) stated that small groups of resourceful and technically adept students do not necessarily need to be the only ones who can develop DIY corpus, suggesting that students in higher vocational education systems could also benefit from this approach. Moreover, DIY corpora can provide access to discipline-specific texts, making it simpler to research discourse and provide suitable resources for particular classes (Charles, 2018).

2.3.3 DIY Corpus and Learner Autonomy

Learner autonomy refers to the ability to take responsibility for one's own learning (Benson, 2007). It has become a goal of education, especially in the current situation where there is an increasing need for lifelong learning and distance learning (Spratt et al., 2002). The strategies proposed by Benson (2001) and Dörnyei (2001) for the development of learner autonomy are closely related to corpus-based language learning, as the corpus is introduced as an inductive language learning method that promotes critical thinking, linguistic awareness, autonomy and lifelong learning (Cobb & Boulton, 2015). However, large general corpora have been criticised for overwhelming students with excessive data (Charles, 2012), while concordance lines can sometimes be decontextualised due to their limited length (McCarthy & O'Keeffe, 2010). In comparison, tailor-made discipline-specific corpora can provide a viable alternative to large general corpora (Charles, 2012).

The promotion of learner autonomy is a crucial factor in the use of DIY corpora (Charles, 2012). Students have the autonomy to choose which linguistic information is included in their corpus, and consulting their corpus can help them become less dependent on teachers to achieve their writing objectives. DIY corpora are resources that can be accessed without an internet connection and are always available wherever and whenever needed. Yoon (2008) conducted a qualitative study with six EAP writers and found that using a corpus in connection with the writing process enhanced their sense of responsibility, increased their writing autonomy and improved their writing assurance. Charles (2012) also conducted qualitative research over six two-hour weekly sessions, which included compiling

information from the final questionnaires to determine how the students received the corpus work. The 50 participants showed generally optimistic attitudes and were able to comprehend the disciplinary culture through suitable language, collocations and subject-specific terminology. This strategy was deemed appropriate for regular EAP classrooms as well. However, it should be noted that the literature on autonomy cited is based on qualitative research, which means that the responses may have been influenced by respect or courtesy towards the instructor.

Charles (2019) stated more recently that advanced students can build their own corpora and frequently demonstrate enthusiasm for using corpus data, valuing the autonomy that comes with having access to their own corpora.

However, while the relationship between corpus use and learner autonomy is acknowledged, the empirical evidence for the profound effect of DIY corpora on autonomy is limited, except for the qualitative feedback of the participants (Yoon, 2008; Charles, 2012).

2.4 Research Questions

All input-based approaches to vocabulary learning have their disadvantages. Direct learning through word cards or word lists can be too decontextualised (Oxford & Crookall, 1990, pp. 9–10), reading can be slow and prone to errors (Peters et al., 2009), listening comprehension can be as complex as vocabulary learning (Vandergrift, 2013) and watching video materials

places more pressure on educators for selection and implementation (Peters & Webb, 2018). In contrast, output-based approaches align with the output hypotheses (Swain, 1995) and have relatively higher indices according to ILH (Hulstijn & Laufer, 2001) and TFA (Nation & Webb, 2011), with higher indices indicating better learning outcomes. Certain output activities, such as the pushed activity definitions (Bao, 2018), have been found to be more effective than input-oriented activities in terms of EFL vocabulary knowledge. However, writing, as a prominent output activity, has not been empirically analysed for vocabulary acquisition. Collocation use is a significant predictor of writing quality (Chang et al., 2008; Daskalovska, 2015; Siyanova-Chanturia, 2015; Zou, 2019). Words are part of an interconnected linguistic system (Nation, 2022), and multiword speech accounts for over half of written discourse (Siyanova-Chanturia, 2015), with the use of collocations seen as a hallmark of near-native language proficiency (Chang et al., 2008). Among the different types of collocation errors, verb collocation (especially verb–noun collocation) errors are prominent among Chinese EFL learners (Zou, 2019; Wang & Zhou, 2020). However, traditional activities for learning collocations in course books (Daskalovska, 2015) often require less time and effort and do not trigger depth of processing for successful learning. Thus, the first research question is:

To what extent does incorporating a DIY corpus improve writing quality and the use of verb collocations?

Furthermore, vocabulary is the building block of language and is crucial for language ability (Webb & Nation, 2017). Empirical corpus-based studies have highlighted many benefits

(Bowker, 2018), including vocabulary acquisition (Karras, 2015). However, corpus-based research abroad has predominantly focused on advanced EFL learners (e.g., Gaskell & Cobb, 2004; Yoon, 2008; Cotos, 2014), while Chinese researchers have devoted more efforts to describing language use as seen in CLEC (Yang et al., 2005), productive vocabulary use (Sun, 2017) and verb–noun collocation use (Wang & Zhou, 2020). Except for a few corpus-based studies conducted by Luo and Liao (2015) and Luo (2016), which primarily focused on writing quality instead of vocabulary acquisition, empirical studies on vocabulary acquisition through corpus-based approaches are limited. Hence, the second research question for this study is:

To what extent do participants improve their knowledge of target vocabulary within writing tasks after incorporating a DIY corpus?

Meanwhile, several scholars (Spratt et al., 2002; Sakai & Takagi, 2009) have attempted to develop reliable questionnaires to measure learner autonomy, as it has become increasingly important for lifelong learning and distance learning (Spratt et al., 2002) when face-to-face teaching is not available. Corpus-based language learning aligns with the strategies proposed by Benson (2001) and Dornyei (2001) for fostering learner autonomy. However, empirical evidence for the impact of corpus use on learner autonomy is limited, except for a few studies providing qualitative feedback from participants (Yoon, 2008; Charles, 2012). Although CBLP was proposed by Ma et al. (2021), its application in the context of higher vocational institutes is rarely explored. Therefore, the third research question is:

To what extent does the use of a corpus facilitate participants' learner autonomy?

Finally, one of the most recent advancements in corpus linguistics, the DIY corpus, has overcome the drawbacks of large general corpora, such as overwhelming students with unnecessary data, by providing learners with more discipline-specific data (Charles, 2012). Charles (2012) suggested that small groups of financially well-off and technically proficient students do not necessarily have to participate in DIY corpus-building. Thus, the fourth research question for the present research is:

What are the participants' attitudes and perceptions towards using corpus data in English writing?

Chapter 3:

A Pilot Study on Corpus-based Approaches

Although successful learners are very similar, every unsuccessful learner will fail in their own unique way (Sakai & Akiko, 2009). A higher vocational institute (HVI) admitted students from Guangdong Province whose scores on the College Entrance Examination (CEE), also known as Gaokao, were between 180 and 445 out of 750 in 2022. Notably, HVI freshmen typically have low English scores. Accurate and appropriate collocation contributes to fluency and idiomaticity in second language acquisition (Wang & Zhou, 2020); however, the percentage of verb collocation errors in the Chinese Learner English Corpus (CLEC) is approximately 11.61% (Yang et al., 2005, p. 15). Following a deep analysis of 60 doctoral English augmentative essays, it was found that *verb + object* was the most frequently used syntactic structure, with only approximately 25% of these constructions being incorrect on account of verb and noun misuse (Wang & Li, 2018). Furthermore, among 72 undergraduates majoring in English, *verb + object* and *verb + prepositions* were the most frequently used syntactic structures, with approximately 30% of these being incorrectly constructed (Wang & Zhou, 2020). For this dissertation, more than 300 answer sheets for the written parts of the final English examinations for two consecutive school years at an HVI were carefully proofread, and the findings are congruent with the literature. Although it has been proposed that teachers should attach more importance to teaching collocation (Wang & Li, 2018; Wang & Zhou, 2020), a pilot study is required to substantiate this premise.

A quasi-experimental study was carried out to evaluate the efficacy of incorporating corpora into writing instruction for reducing verb collocation errors in student essays, while also piloting research instruments. In addition, vocabulary tests were administered to students to assess their vocabulary knowledge, and a questionnaire was adapted to measure learner autonomy. The tests were taken by two classes of students before and after they were instructed to write an essay. In this way, the research attempts to answer the following questions:

(1) Does corpus-based writing impact the quality of verb collocation use?

(2) Do participants improve the target vocabulary knowledge embedded in the writing task?

(3) Does corpus use facilitate the learner autonomy of the participants?

3.1 Participants

Generally, all HVIs are registered at the Ministry of Education and managed by the local municipal people's government. These institutes are primarily open to students from the province in which the institute is located. The majority of applicants to HVIs are high school graduates who have passed the CEE (Gaokao) or independent admission tests (especially for those who studied at middle vocational schools) in each province. However, the Targeted Poverty Reduction Policy broadened the scope of admission, requiring HVIs to admit beneficiaries of the policy, including students from mountainous areas such as Yunnan and Guangxi Provinces and veterans. Notwithstanding, most enrollees take an English test as part of their Gaokao. For the pilot study of this research, students from an HVI in Shenzhen,

Guangdong Province, were recruited as participants. The pilot study had 84 pupils in total, including 44 in the experimental group and 40 in the control group. Demographic data on the subjects was gathered using a personal information questionnaire (Table 8).

Table 8. Demographic information on participants in the pilot study

Group	Age	No.	Years of English learning	Average score on Gaokao English (total = 150)
Experimental	18	44	9	80
Control	18	40	9	79

3.2 Instruments

The instruments used in the pilot study included two large general corpora, one writing task, vocabulary tests, and a questionnaire on learner autonomy.

3.2.1 Corpora

The two general corpora used in the pilot study are the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC). The COCA comprises six genres: TV/movies, spoken, fiction, magazines, newspapers, and academic journals. In addition to these registers, the COCA also contains over 125 million words from blogs and about 130 million words from websites, making it indeed a corpus composed of contemporary American English. It can be accessed at the *English-Corpora.org* website

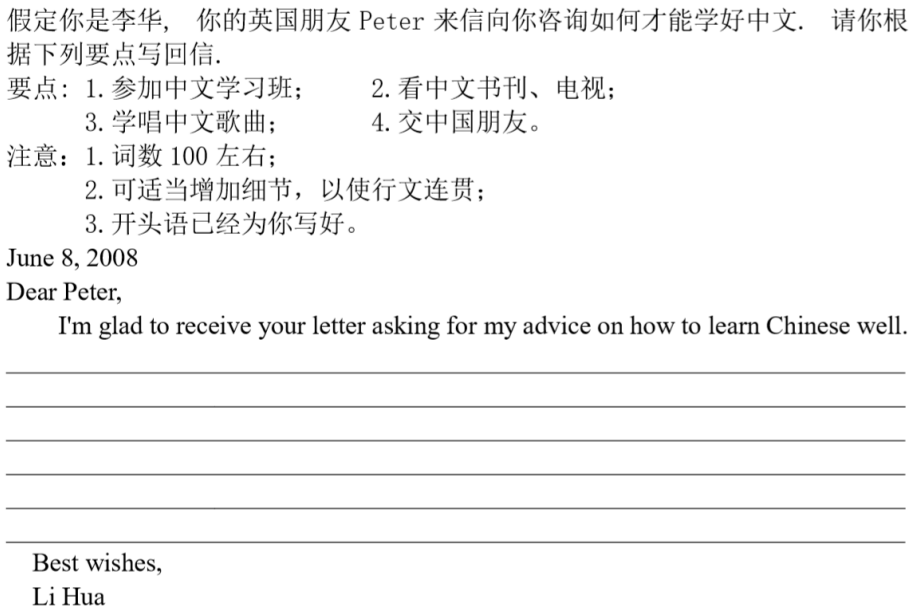
(https://www.english-corpora.org/coca/). The COCA is sizable enough to illustrate learners the low-frequency use or meaning of target words and is representative of American English (Lee & Lin, 2019). The other corpus used in the pilot study is the BNC, which can be accessed free of charge at the *English-Corpora.org* website (https://www.english-corpora.org/bnc/). The BNC is a corpus of spoken and written English samples in late 20th-century British English that totals 100 million words. Ninety percent of the samples included in the BNC were taken from local and national newspapers, academic journals and periodicals, novels (including fiction and non-fiction), other published and unpublished materials, and research studies. Transcriptions of recordings taken during particular meetings and events make up the majority of the remaining 10% of the data.

3.2.2 Writing Task

Participants in the pilot study had previously taken the CEE English test or had practiced similar topics to those tested in the exam. Because the level of English proficiency possessed by vocational college students is generally low, it is not suitable to test them on particularly difficult topics. The subject of the writing assignment that was part of the pilot research is also closely tied to the subject of the first unit of the English textbooks in the institute. After studying this unit, students will know enough about format preparation and content preparation for the writing task. This would make the participants feel that the writing task is not particularly difficult. The writing task was adapted from a Gaokao English test used nationwide in 2008. Although the original test was used years ago, it matches the topic of the

textbook of the institute, in which Unit 1 is *Learning English: A New Start*, containing topics such as *Lesson 3 What You Can Do to Help Your Language Learning*, and *Lesson 4 Developing as a Language Learner*'.

The writing task instruction was 'Suppose you were Lihua, and your English friend, Peter, wrote to you asking how to learn Chinese effectively, please write back addressing the following four key points: take a Chinese course; watch TV and read books, newspapers, and magazines in Chinese; learn and sing Chinese songs; make more Chinese friends'.



假定你是李华， 你的英国朋友 Peter 来信向你咨询如何才能学好中文． 请你根据下列要点写回信.
要点：1. 参加中文学习班；　　2. 看中文书刊、电视；
　　3. 学唱中文歌曲；　　4. 交中国朋友。
注意：1. 词数 100 左右；
　　2. 可适当增加细节，以使行文连贯；
　　3. 开头语已经为你写好。
June 8, 2008
Dear Peter,
I'm glad to receive your letter asking for my advice on how to learn Chinese well.
Best wishes,
Li Hua

Figure 2. Sample from the first writing task

3.2.3 Target Verbs and Vocabulary Tests

In the pilot study, the essential target vocabulary was chosen randomly based on reference articles and teaching experience. The instructor analysed the reference articles to determine the core vocabulary and patterns for every clause, and then deliberately chose fallible

structures, mostly verbs, e.g., *participate in (doing) sth*.

Table 9. Required vocabulary used in writing tasks

Participate	Read	Watch	Listen	Sing	Make (friends)

Three types of vocabulary tests were adopted: pretest, posttest, and delayed posttest. The tests were administered in two rounds of testing, with the first two tests (the pretest and the posttest) administered using the Vocabulary Knowledge Scale (VKS) before and after the writing task. The information processing framework for second language (L2) knowledge acquisition served as the basis for the design of the VKS, which was modified from research by Paribakht and Wesche (1996). According to certain classification methods, it is further divided into subcategories for "selective attention, recognition, manipulation, interpretation, and production" (Paribakht & Wesche, 1996). A five-point scale from utter unfamiliarity to utilising the target word in a sentence is used by the VKS to measure students' word knowledge at various levels and track their vocabulary growth. The VKS is useful since the amount of target words determines how long it takes to complete. Additionally, it enables participants to show that they have some understanding of each issue, which reduces their worry. In addition, the VKS is commonly adapted by researchers—for example, Lee and Lin (2018)—indicating that it is a tool with high reliability. The research participants were provided with clear instructions in Chinese before the posttest, and timely help was offered while they filled in the scale.

1. I do not remember having seen this word before
我从未见过该词
2. I have seen this word before, but I do not know what it means
我见过该词但不知道其意义
3. I have seen this word before, but I think it means _____ (synonym or translation)
我曾见过该词，我觉得它的意思是（翻译或同义词）
4. I know this word. It means _____ (synonym or translation)
我确认我认识该词，且它的意思是（翻译或同义词）
5. I can use this word in a sentence, e.g.: _____ (if you do this section, do section 4)
我能在句子中使用该词（如果你做这个也要完成第4题）

Figure 3. Sample from the Vocabulary Knowledge Scale used in the pretest and posttest

3.2.4 Questionnaire on Learner Autonomy

Autonomy is a characteristic possessed by learners and is exhibited to a certain extent in different contexts (Benson, 2007). Learners can freely explore and experiment with new language inputs thanks to their direct or indirect contacts with resources. A corpus is a collection of real language use in context that enables learners to pinpoint their own grammatical or vocabulary weaknesses.

Benson (2001) noted that researchers frequently divide the idea of learning autonomy into various constructs, such as setting goals, defining content and progressions, choosing methods and techniques to employ, keeping an eye on the acquisition process, and assessing the results (Spratt et al., 2002). They developed a questionnaire (Figure 4) to examine changes in students' perceptions of their responsibilities towards English learning—as a measure of their learning autonomy. I adapted this questionnaire and administered it to the

study participants before and after the intervention task to assess changes in their learner autonomy. Because this was a pilot study, the two parts of the questionnaire, comprising two sections and 22 questions with responses on a 5-point Likert scale, were utilised. Part one of the questionnaire assesses students' perceptions of their responsibility towards English learning. It includes questions such as 'How much responsibility should you assume when you are enrolled in classes with respect to the following things: …' Part two focuses on students' English learning activities outside of the classroom and asks questions such as, 'How often have you voluntarily engaged in the following English learning activities: ...?' The purpose of part two of the questionnaire is to determine whether students are likely to seek out similar resources to learn English in their free time. Because the COCA and the BNC include genres such as movies and lyrics, they could also engender participants' interest in more English learning activities after class.

Part One. Your Perception of Responsibility toward English Learning 你对英语学习责任的认识
When you are taking classes, how much responsibility should you take concerning the following items?
当你在上课时，你认为对下列事项的应承担多少责任？
1- Never 从不 2- Seldom 几乎不 3-Sometimes 有时 4-Often 经常 5-Usually 通常

1) To decide your goal of study in one semester 去决定你这个学期的目标
2) To check how much progress you make 去检查你取得多少进步
3) To decide the learning materials you use in class 去决定你在课堂上使用的学习资料
4) To decide topics and activities you learn in class 去决定你上课的话题和活动
…
Part Two. English learning activities outside the class 课外英语学习的活动
How often have you done the following English learning activities voluntarily before?
2- Never 从不 2- Seldom 几乎不 3-Sometimes 有时 4-Often 经常 5-Usually 通常

1). I have used the Internet in English 使用英语网页
2). I have listened to English songs 听英语歌曲
3). I have watched English movies 看英语电影
4). I have practiced speaking English with your friends 与朋友一起练习英语
5). I have learned English grammar 学习英语语法
6). I have prepared for proficiency tests such as PRETCO 准备英语水平测验比如英语 B 级
…

Figure 4. Questionnaire adapted from Spratt et. al. (2002)

3.3 Data Collection

The pilot study's quantitative data was gathered in two stages. The first phase of the study comprised teaching the participants how to use the COCA and BNC websites' features, and the second phase concentrated on the participants' corpus exploration and composition development. The remarks and feedback made by the students during and after class were used to compile qualitative data.

Training before the experiment is helpful for students to become familiar with the corpus websites (Gaskell & Cobb, 2004; Zhang et al., 2017). The training was conducted using hands-off and hands-on formats. During the hands-off training phase, the teacher demonstrated how to conduct searches in corpora and presented PowerPoint presentations outlining the functions of the corpus websites. During the teacher demonstrations, the students did not have access to the corpora. Usually, the teacher chooses one word from each task on the students' worksheet to demonstrate the detailed procedures for consulting the target corpus. To facilitate students' knowledge acquisition on the corpus, the teacher carefully predesigned and videotaped the steps before the class and uploaded the video online for their reference.

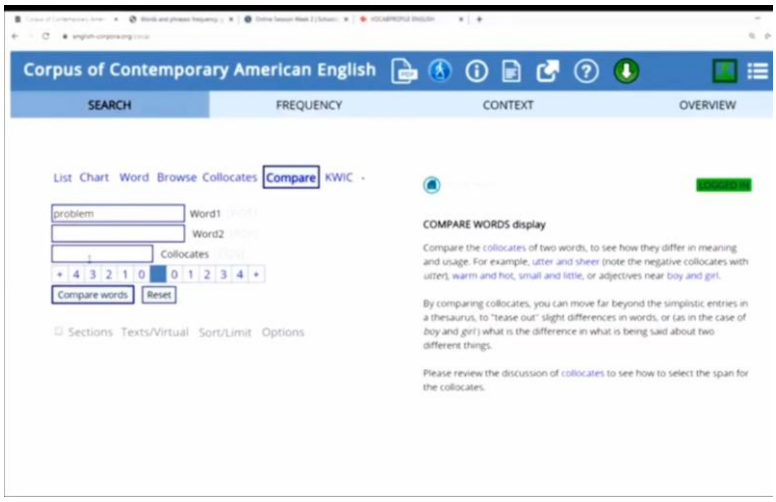
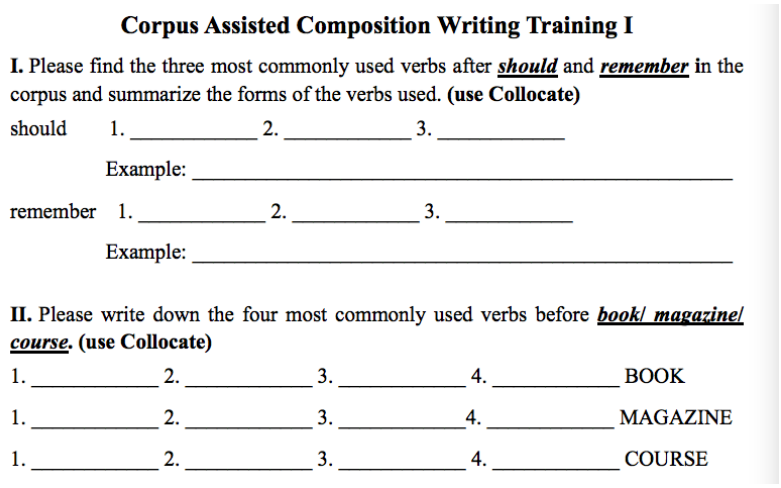


Figure 5. Screenshot from video clip of corpus website demonstration.

The demonstration was conducted in a common classroom, and the appliances included only a desktop and a projector. A computer room was unavailable at the time because they had all been booked for courses related to the students' majors. During the hands-on phase, the students tried different search functions on the COCA website (Figure 5), including List, Collocate, and Compare, to accomplish the tasks on the worksheet. All the functions and techniques that the students would need in the hands-on phase were demonstrated in class during the hands-off phase, and relevant video clips were also provided for the students' reference. Whenever the students encountered challenges consulting the corpus, they could refer to the videos or text the instructor. The worksheet (Figure 6) was designed to draw the student's attention to the target vocabulary and the structures in use. The preferred functions for completing the research were provided in parentheses. A Chinese version of the worksheets was also provided, just in case.



Corpus Assisted Composition Writing Training I

I. Please find the three most commonly used verbs after ***should*** and ***remember*** in the corpus and summarize the forms of the verbs used. **(use Collocate)**

should 1. _____ 2. _____ 3. _____

Example: _____

remember 1. _____ 2. _____ 3. _____

Example: _____

II. Please write down the four most commonly used verbs before ***book/ magazine/ course*****. (use Collocate)**

1. _____ 2. _____ 3. _____ 4. _____ BOOK

1. _____ 2. _____ 3. _____ 4. _____ MAGAZINE

1. _____ 2. _____ 3. _____ 4. _____ COURSE

Figure 6. Sample of worksheet for corpus-assisted composition writing

For the second part of the data collection, a corpus website was used to write compositions. A five-step approach (Table 10) was designed to ensure that all the required steps were implemented.

Table 10. Student and teacher activities during the five steps of the experiment.

Step	**Students' activities**	**Teacher's duties**
1	Fill in questionnaire on learner autonomy, take vocabulary test, and read the composition instructions.	Hand out test papers, supervise students, explain the instructions, and invigilate the writing task.
2	Read the worksheet paper and understand how to complete the task.	Demonstrate (in class) the functions required for each exercise.
Gap Weekends (A weekend and two days before the next lesson)		
3	Confirm the answers in their worksheet and ask questions.	Present the reference answer and some critical functions in class.

4	Write and revise composition, as required.	Walk around to ensure the experimental conditions and offer help.
5	Hand in composition and take a vocabulary test(s).	Confirm submission of student documents, hand out test papers, and supervise test(s).

During gap weekends, the students had four days to explore and complete the worksheet. The students were free to contact the teacher with questions concerning the worksheet. Step 1 and Step 2 were conducted in one session, and the worksheet was handed out as homework, requiring the students to use their laptops or mobile phones to fill out the worksheet. Steps 3–5 were implemented in two separate lessons conducted in a single session.

In Step 1, the students took the VKS test without prior notice. Students were then told to read the instructions for the writing assignment and include the target vocabulary into their work. In Step 2, the students had their copies of the target vocabulary and were required to complete the worksheet by referring to the uploaded video clips and the recommended functions. They were given four to five days to explore the corpora. The students in the control group, in contrast to the experimental group, were not given the worksheet; rather, they were given the assignment of looking up the target language in a dictionary. After a gap of four to five days (including the weekend), the instructor shared the findings with the students (Step 3) and demonstrated one sample for each worksheet. After the students confirmed their answers on

the worksheet, they were given most of the class time to draft and revise their compositions before handing in the final version (Step 4). In Step 4, the students in the experimental group were instructed to refer to the worksheet or the corpus websites. In contrast, the control group could use any tools (excluding corpora) they preferred. In the last step, a reordered VKS test was administered to the students to test for improvement in their vocabulary knowledge.

As part of the pilot study, feedback from the students (i.e., the participants) regarding the pilot research was collected through observation, face-to-face conversation, and online chat. Observation of the study participants was a component of every step of the data collection. Face-to-face interactions typically took place during breaks and during the writing assignment, while online dialogues took place after the students had finished the worksheets and answered the questions over the weekend. Table 11 presented major excerpts from the participants' oral feedback, WeChat screenshots, messages at the bottom of worksheets, and class observation notes made during class. To provide anonymity and honour the human rights of the participants, the students were referred to as A, B, C, D, E, F, and their actual words were translated with a focus on their intended meaning.

Table 11. Excerpts from students' feedback on corpus-based writing

Student	Feedback
A	I do not have a laptop, and accessing the material via smartphone was not easy.
B	I disliked carrying my computer from one classroom to another, as I had

	to carry it to the next classroom for 20 minutes in the burning sun.
C	The COCA website repeatedly asked me to register and even pay for it. Writing compositions after class was a mission impossible.
D	I am not interested in the corpora, but I am worried about passing my (English) exams.
E	I felt it was too hard to write in English; even reading the (concordance) lines was challenging.
F	English is hard for me, and I am reluctant to do English homework after class. There are more new words than the target verbs in (concordance) lines.

3.4 Data Analysis

After the data collection, the instructor proofread the compositions and asked another teacher to check and confirm mistakes pertaining to the target verbs. The total number of verb collocation errors was divided by the group's total number of pupils to arrive at the average. Subsequently, a t-test was used to compare the average for the control group against that for the experimental group.

Table 12. Number of verb collocation errors made in the writing task.

Group	N	Total no. of verb collocation errors	Average	SD	Sig. (2-tailed)

Experimental	39	42	1.08	1.06	<.000
Control	35	78	2.22	1.54	

The students' compositions were graded by the researcher and then crosschecked by a colleague who is also an English teacher — with over seven years of English teaching experience. Afterwards, the data were inputted into the Statistical Package for the Social Sciences (SPSS) version 27.0, and an independent sample t-test was conducted to compare the number of verb collocation errors made on the writing task. As shown in Table 12, participants in the experimental group (N = 39) made significantly fewer verb errors ($p < .000$) than those in the control group (N = 35).

Table 13. Between-group comparison of pre-test and post-test vocabulary assessments (VKS)

Test	Group	N	Average (total = 30)	SD	Sig. (2-tailed)
Pretest	Experimental	43	21.48	4.4	.85
	Control	36	21.75	7.2	
Posttest	Experimental	10	24.1	2	.85
	Control	25	23.84	6	

There was a small change in the students' knowledge of the target verbs (Table 13), as tested using the VKS. Although there were no significant differences in the results of the pretest and posttest, it should be noted that the number of test-takers decreased on the posttest. The test was administered to the study participants after they completed the composition exercise. The

students were expected to self-report the VKS after class; however, some students were occupied with tasks related to their majors and could not spare sufficient time to complete the test.

Table 14. Independent sample t-test for within-group comparison of pre-test and post-test vocabulary assessments (VKS)

Group	Tests	N	Average (total = 30)	SD	Sig. (2-tailed)
Experimental	Pretest	43	21.48	4.4	.008
	Posttest	10	24.1	2	
Control	Pretest	36	21.75	7.2	.250
	Posttest	25	23.84	6	

Table 14 presents the t-test for the within-group comparison of the pre- and post-test vocabulary assessments (VKS). The experimental group, which used a corpus-based approach to writing, improved significantly on the target vocabulary knowledge ($p = .008$, 2-tailed), while the control group did not outperform themselves on the vocabulary assessment ($p = .250$, 2-tailed). The average score improved by about three in the experimental group, while the students in the control group improved by only about two. Although the number of test-takers decreased in the post-test assessment, it still captured these improvements.

Table 15. Independent sample t-test of learner autonomy using a scale adapted from Spratt et

al. (2002).

Section	Pretest/Posttest	N	Group	Average	SD	Sig. (2-tailed)
English learning responsibilities	Pretest	44	Experimental	33.8	5.76	.045*
		38	Control	36.7	7	
	Posttest	40	Experimental	34.8	7.8	.878
		38	Control	34.6	6.2	
Frequency of English learning activities outside of the classroom	Pretest	44	Experimental	33.2	7	.428
		38	Control	31.9	7	
	Posttest	40	Experimental	36.2	7.8	.700
		38	Control	35.6	7.4	

Note. *$P < .05$.

The learner autonomy questionnaire originally comprised four sections, but only two sections were selected for use in this pilot study: Part A and Part B. Part A is on students' perception of responsibility towards English learning, with scale items asking — for example — how they decide their study goal for a semester. Part B focuses on the frequency of students' English learning activities outside of the classroom, with scale items asking — for example — how often they have searched the English dictionary after a class. Because the corpus worksheet was given out as homework during gap days, it was believed that it could somehow influence the learner autonomy of the students. Interestingly, the students in the experimental group had significantly lower pre-test scores (p = .045) on perceptions of responsibilities towards

English learning than students in the control group. However, after the corpus-based writing exercise, their performance, as assessed using the VKS questionnaire, showed no significant changes, as the p-value came to .878 (2-tailed). Regarding the frequency of English learning activities outside of the classroom, both groups recorded an increase, but the detected differences were not significant. Because this was a one-time experiment, and the students did not provide much positive feedback on the large general corpora, it can be concluded that corpus-based activities can affect their learner autonomy. However, longitudinal research and more experiments are still needed to confirm the results of this pilot study.

From the students' feedback, it was found that the students were reluctant to bring their laptops or tablets to class because most of them are habitual mobile phone users and 'do not have a laptop'. Laptops and desktops are perceived as recreation tools instead of learning tools, and students do not want to bring them along to class for learning purposes, even when they have one, as seen in this feedback: 'I disliked carrying my computers from one classroom to another because I had to carry it to the next classroom for 20 minutes in the burning sun'. Accessing the COCA or the BNC websites was not convenient for mobile phone users, and the experience was unfriendly for those who did not register, as captured in this feedback: 'The COCA website repeatedly asked me to register and even pay for it'.

The English proficiency of some students was unexpectedly low, and they might need more help before being able to write or read complex concordance lines. As some students said, 'I felt it was too hard to write in English; even reading the (concordance) lines was

challenging', and 'English is hard for me'. In addition, taking the assessment tests, writing compositions after class, or doing homework may not have had the desired outcome because, as some students said, 'I am reluctant to do English homework after class', and 'Writing compositions after class was a mission impossible'. Most importantly, some students told the instructor that 'There are more new words than the target verbs in the (concordance) lines' and 'I am worried about passing (English) exams'. Thus, the concordance lines would effectively hold their attention if the lines were related to their English exams, such as the College English Test Band 4 (CET 4).

3.5 Findings

Regarding the research questions for the pilot study, it was found that the corpus-based writing significantly decreased the errors made in verb collocation use by students in the experimental group when compared with their counterparts in the control group ($p < .000$, 2-tailed). The experimental group, which used a corpus-based approach to writing, significantly improved on the target vocabulary knowledge ($p = .008$, 2-tailed) compared with their performance on the pretest. In addition, as seen in Table 15, participants in the experimental group improved their perception of their responsibilities towards learning the English language. However, the experiment did not impact their English learning activities outside of the classroom.

This pilot study also revealed some of the cons of using large general corpora, which are

embedded with stances of irrelevant chunks and clauses (McCarthy & O'Keeffe, 2010). Furthermore, students are overloaded by the concordance lines they discover (Cobb & Boulton, 2015) within the corpora. In this pilot study, limited vocabulary, low language proficiency, and low learning motivation impeded the participants from fully exploring the corpus websites selected for use. In other words, selecting the appropriate concordance lines for students — in particular, concordance lines relevant to 'passing (English) exams' — and refining the target vocabulary or structures where necessary, is critical for low-level students because 'reading the (concordance) lines was challenging'.

Hardware and software challenges also impacted the experimental conditions. The corpus servers for the COCA and the BNC websites are based in the oversea areas. It took the students more time to index their target vocabulary than it would have if it were a local website. The experiment was conducted in the context of a public English course. The facilities in the classrooms included only one desktop computer and projector(s) for the teacher's use. The students could not access computer rooms, except for a course that was a major or practical in nature. Worse still, the students' smartphones cannot display all the functions of the corpus websites, and there were unprecedented Internet delays when about 40 students browsed the website simultaneously. Printing out do-it-yourself (DIY) corpus materials resonated with the needs and interests of the students.

Students are reluctant to bring their laptops to class; they are also reluctant to write English compositions after class. Public English courses are scheduled between other general elective

or foundation courses; thus, students have to carry their laptops around — from one classroom to another — in the heat. Furthermore, most students do not have a laptop, tablet, or device other than a smartphone. They are heavily burdened with tasks related to their major and had a tight schedule due to public holidays and practical training courses. Because of their limited English knowledge and lack of intrinsic motivation, they were reluctant to complete English homework (a research procedure) after class, which required them to use a laptop or desktop computer. The computer rooms on campus were fully booked and overloaded for courses related to the major. In contrast, exam-related materials can grab students' attention and motivate them to complete a task, as observed with the students in the control group.

3.6 Implications

Considering the students' reluctance, the researcher may alter the final score grading criteria. When students complete the English writing tasks, they can be awarded additional class participation points. In addition, having students take the VKS and other tests during class hours and not after class can reduce their homework burden.

The homogeneity of the participants recruited for the study needs to be considered very seriously. Participants in the pilot study comprised two groups of students. The English learning backgrounds of the participants were different, making the analysis of learner autonomy and English proficiency uncertain. Selecting students with similar demographics,

especially high school education backgrounds, can also be a critical factor for this research.

A different type of corpus (i.e., a DIY corpus) should be used in the main study. The participants in the pilot study exhibited evident interest in utilising a new approach to English writing and learning. However, they lost interest immediately they encountered difficulties posed by the large general corpora, such as excessive data (Charles, 2012) and decontextualisation due to limited length (McCarthy & O'Keeffe, 2010). Worse still, the lower English proficiency of the target participants, the unauthorised access to overseas internet resources, and the shortage of personal laptops or desktops all hindered the incorporation of the large general corpora, i.e., the BNC or the COCA in this case. However, a tailor-made corpus can provide a viable alternative to large general corpora (Charles, 2012). In this regard, a DIY corpus with authentic data from past exams, as desired by the participants, could tackle the problem of irrelevant data posed by large general corpora. The problem of decontextualisation can also be overcome via the careful selection of appropriate concordance lines. Furthermore, the challenge of only a limited number of students owning the required hardware can also be addressed by printing out concordance lines for students. In sum, printing out the DIY corpus for students with low English proficiency at an HVI can sustain participation and maintain their interest.

Chapter 4: Research Methodology

In light of the results of the pilot study, incorporating DIY corpus material can help improve vocabulary knowledge and certain aspects of learner autonomy. In addition, more writing tasks were adopted in the main study to investigate the impact of the volume of writing tasks on learner autonomy and writing quality produced by students. The data collection methods used in this study are described in this chapter.

The following elements are addressed in this chapter: research design and participants, quantitative methods, qualitative methods, and ethical concerns. The instruments used, the data collection procedure, and the data analysis procedure are discussed for both the quantitative and qualitative method components of the study. Regarding the quantitative data, the quality of the writing was assessed using grading schemes, vocabulary knowledge was measured via pre/immediate and post/delayed post-test assessments using the VKS, changes in learner autonomy were captured through pre- and post-test administration of a questionnaire, and the participants' perceived reactions to using DIY corpus material in their writing were captured using another questionnaire. For the qualitative data, a semi-focused group interview was used to capture relevant data, as well as to compensate for and enrich the corresponding quantitative data.

4.1 Research Design

As outlined earlier, a mixed methods approach was adopted for this study. Specifically, a quasi-experimental approach was adopted for the quantitative aspect of the study and a semi-focused group interview were used for the qualitative aspect.

The mixed methods approach was adopted to achieve 'hybrid vigor' (Dörnyei, 2007, p. 42) and "additional benefits for an understanding of the phenomenon and question" (Dörnyei, 2007, p. 47). According to Dörnyei (2007), several arguments have been made regarding the value of a mixed methods approach. First, it is known for "adding advantages while eliminating disadvantages" (Dörnyei, 2007, p. 45) because the advantages of one method can be used to offset the disadvantages of another method used in the same research. Second, a mixed methods research approach is particularly suitable for 'multilevel analysis of complex problems' (p. 45) because it allows researchers to obtain data on individuals and especially broad social backgrounds. Third, mixed methods can 'improve the effectiveness of research', as 'the corresponding evidence obtained through a variety of methods can also improve the universality of the results, that is, external effectiveness' (p. 45). Finally, a popular benefit of mixed methods is that 'the end result is usually more acceptable to a wider audience than the results of a single method study' (p. 46). Creswell (2014) pointed out that a mixed methods design provides comprehensive answers to each research question in a study and believed that a research design integrating different methods may produce superior results in terms of quality and scope. By mixing data sets, researchers can 'understand research problems better than when using any method alone' (p. 552). However, the mixed methods approach can be

significantly challenging to a researcher because ‘researchers are not adequately trained in both methods’ (Dörnyei, 2001, p. 46). Having had the experience of a master’s dissertation, and with help from supervisors, this challenge has made this study a rewarding research experience.

Regarding the options for coordinating the order of quantitative and qualitative methods, an explanatory sequential mixed methods design that is appealing to those with strong quantitative backgrounds or from sectors that are more recent adopters of qualitative approaches was outlined by Creswell (2014): “A two-phase project” (p. 330). In this study, quantitative data were collected in the first phase, and the results were analysed and then used to plan the second (qualitative) phase. The types of individuals chosen for the qualitative phase and the kinds of questions posed were informed by the quantitative results. In the explanatory design, a broad picture of the research problem (learner autonomy, perceived reactions to using the DIY corpus) was obtained using quantitative tools such as questionnaires and vocabulary knowledge tests. Sequential qualitative studies in the form of interviews were then used to generate insights to explain the initial quantitative results. As predicted by Creswell (2014), there were two challenges to implementing a mixed methods design in this study. The first step was properly planning which quantitative findings to investigate further and which subjects to interview for qualitative information during the second phase. The second difficulty was that because the qualitative study’s goal was to follow up on and thoroughly examine the quantitative study’s conclusions, samples for it had to be taken from participants in the quantitative study.

Following Creswell (2014), an explanatory sequential mixed methods design was employed to investigate the outcomes and perceptions of Chinese students at a higher vocational institute (HVI) after incorporating DIY corpus tools into an English writing task. This involved using vocabulary tests, grading writing quality, and semi-focused group interviews to achieve a comprehensive understanding of the impact of the corpus tool. Therefore, a vocabulary knowledge scale, a questionnaire on learner autonomy, a writing quality (geared towards verb collocation errors and overall grading), and language proficiency tests were adopted primarily to quantitatively understand the effects of the DIY corpus tool used while writing. Then, a questionnaire and a semi-focused group interview were used to gain further insight into the mental world of the participants to further explore the impact of the corpus tool.

4.2 Research Context and Participants

Mandarin Chinese is the national language of China and the first language of the Chinese Mainland. It is also the language of instruction in most provinces and cities, from primary schools to universities, while English is taught as a foreign language. According to the *Curriculum Standards for College English in Higher Vocational Education* (Ministry of Education [MOE], 2021), the basic module of the English course is a compulsory or limited elective course for first-year students, and the total class hours are 128–144 over two semesters. Despite their relatively poor academic performance on the College Entrance Examination (CEE), also known as Gaokao, students at the HVI in this study still needed to

master about 500 new words, building up to a total vocabulary size of 2,300–2,600 in just two semesters.

The Shenzhen Institute of Information Technology (SZIIT) is a full-fledged public institution that provides higher education in Shenzhen City. Approved by the People's Government of Guangdong Province, SZIIT is registered under the MOE and administered by the Shenzhen Municipal People's Government. Thus, SZIIT is open primarily to students from Guangdong Province. All study participants admitted to SZIIT were expected to have taken the CEE (Gaokao); hence, they took the same Gaokao English as all other students at the vocational institute.

A total of 86 participants participated in this study. In contrast to the pilot study, all the participants in the main study hail from Guangdong Province. This selection criterion was adopted to facilitate a quasi-experimental research design by minimising the differences between the experimental and control groups: if all the participants took Gaokao in Guangdong Province, they would all have taken the same Gaokao English test. Before the experiments, the study participants filled out a bilingual personal information questionnaire, which was administered during class.

Table 16. Demographic information collected via a personal information questionnaire (Appendix 1).

Group	Age	No.	Gender	Years of	Gaokao English

				English learning	**average score (total = 150)**
Experiment	18	46	4 Females, 42 Males	9	78
Control	18	40	15 Females, 25 Males	9	79

The experimental group comprised 46 participants (4 females and 42 males) aged 17–18 years, while the control group comprised 40 participants (15 females and 25 males), also aged 17–18 years. These students generally started learning English as a foreign language from Grade 3 in elementary school. Most of the participants took the same Gaokao English test simultaneously, and their average scores were 78 for the experimental group and 79 for the control group, against a total possible score of 150. In calculating the average score, approximately five participants in each group who did not take the Gaokao exam were excluded from the calculation. These individuals were admitted through an independent admission exercise conducted by SZIIT in response to the broadening of the scope of admissions mentioned in the introduction section of this paper.

4.3 Quantitative Method

The first two research questions were designed to assess the effectiveness of using DIY corpus material in improving student learning outcomes with respect to vocabulary

acquisition and collocation in writing. Question 1: To what extent does incorporating a DIY corpus improve writing quality and the use of verb collocations? Question 2: To what extent do participants improve their knowledge of target vocabulary within writing tasks after incorporating a DIY corpus? A typical experimental setup would be an intervention study comprising at least two groups: an experimental group that receives the intervention treatment (i.e., the DIY corpus material in this study), and a control group to provide a baseline for comparison. However, a true experimental design with random assignment to the experimental groups is rarely feasible; therefore, the common method applied employs intact class groups (Dörnyei, 2001, p. 117). A quasi-experimental design was used to collect the quantitative data in this study.

In the current educational context, this research is an experimental study. The study participants in one class (the experimental group) majored in Telecommunications, and their English proficiency was considered one of the lowest among all the SZIIT departments. To minimise the differences between participants on the pretests, the participants in the other class (the control group) majored in Intelligent Manufacturing. Hence, the students had similar majors (i.e., their majors were in the same field, the sciences). More importantly, there was no significant difference in their English proficiency ($p = 0.84$, 2-tailed), as indicated by a t-test comparing their scores on the Gaokao English test. In addition, the two classes were taught by the same teacher. The teacher, as the researcher conducting this study, was in an ideal position to understand the needs of the participants and create a free and empowering environment in a way that an external researcher could not. From the perspective of ethics,

teachers and researchers can be more flexible when integrating research into compulsory courses, such that the investigation is beneficial to students' learning and does not merely increase their workload, which may negatively affect students' learning and academic performance.

As illustrated in Figure 7, the independent variable in this research is the use of DIY corpora material, while the dependent variable is student learning outcomes, which is further analysed in terms of target vocabulary knowledge and writing quality. There were five writing tasks (four administered in daily teaching and one in the final exam) over the course of a semester.

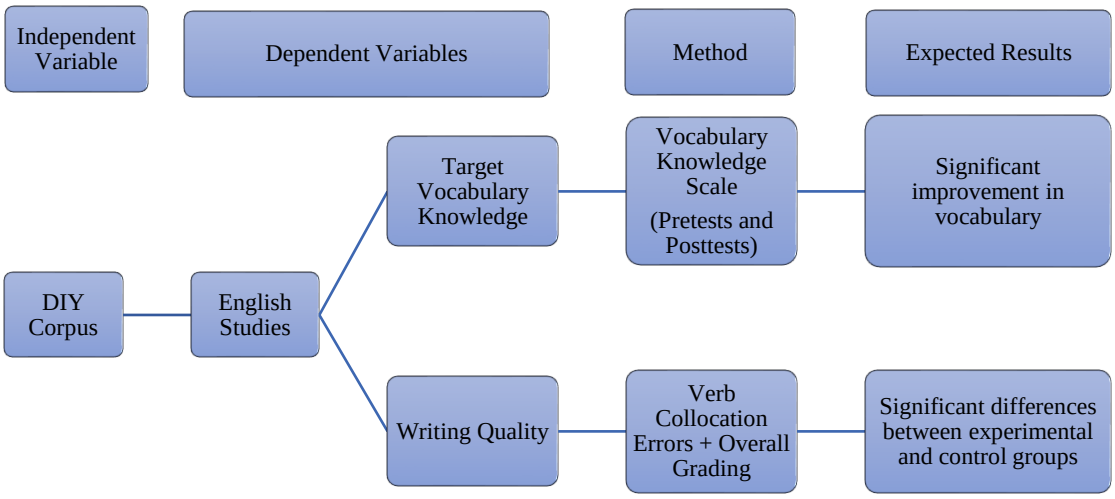


Figure 7. Broad research design

The study lasted for one complete semester, and the writing tasks were given at the end of each unit of the English course (with four units taught in a semester). While the students were drafting and revising their essays, they were allowed to consult the corpus-based handouts occasionally — instead of mobile devices or laptops, hard copy DIY corpus handouts were

adopted in the main research. Based on the pilot study, the corpus tool proved more user friendly than laptops to the students. Moreover, the participants in the pilot study were reluctant to carry laptops to class, and most of them did not even have laptops. In contrast, mobile devices were more common and easy to incorporate into daily campus life. The research plan was aligned with the teaching plan designed for each semester, fitting into the exercise and practice sessions for college English classes. The experiment was conducted at the end of each unit, comprising four lessons; four units of the English course were required in one semester. Students typically took four class hours of the English course every week in one semester (a semester is approximately 13 weeks). Each intervention task was incorporated into a practice session (two consecutive 40-minute periods) after the completion of a unit of the course.

4.3.1 Quantitative Instruments

Because this is a quasi-experimental design for generating quantitative data, two specific ways to improve the design — proposed by Dörnyei (2007, p. 117) — were adopted: preventing students from self-selecting and minimising pre-test differences between the experimental group and the control group. A questionnaire collecting personal information and assessing knowledge of human rights can be designed to serve this purpose. Additionally, the results of the pilot study show that large general corpora do not help students with low English proficiency. Because a suitable corpus for the students in the pilot study was not available, a DIY corpus tool was needed. Furthermore, the corpus application software

AntConc (Anthony, 2020) was adopted to access the DIY corpora.

To answer the research questions concerning verb collocation errors, essay writing tasks adapted from official English tests were utilised as the intervention or treatment for the experimental group. The other quantitative-oriented question concerning vocabulary knowledge was adequately addressed using the Vocabulary Knowledge Scale (VKS) (Paribakht & Wesche, 1996) to assess acquisition of the target vocabulary embedded in the writings.

4.3.1.1 Compiling a DIY Corpus

Compilation of DIY corpora was an appropriate task for the target students in this study. First, the DIY corpora were referenced as small-scale databases for teachers' use (Charles, 2012, 2015, 2017). As expected, the data collected meets student needs and is suitable for teaching purposes, as the study participants were anxious about passing their exams in order to meet the graduation threshold for obtaining a degree. The corpora were composed of authentic language data from published books or past tests, such as reading passages (excluding any exercises). Over 240,000 words in all make up the DIY corpus used in this investigation. Despite the fact that just a few students (5 out of 46 in the experimental group) had any interest in learning English, all of them were eager to pass the English test required for graduation. Accordingly, material related to the English requirement test caught their interest and motivated them to use the corpus. Furthermore, the material matched their English proficiency; CET 4 was slightly above their level of English proficiency, and the

Gaokao English materials were slightly below their level.

Table 17. Basic information on the DIY corpus data.

Exam	**Corpus name**	**Size (242,800 words)**	**Source**
CET 4 material	1. Listening scripts	72,300	Tests taken over the last ten years
	2. Reading passages	117,800	
	3. Writing Part 1	21,900	Published best-seller reference books
Gaokao English materials	4. 40 articles	7,800	
	5. Writing Part 2	3,000	
	6. CELST scripts	20,000	Tests taken over the last ten years

Note. CELST = Computer-based English listening and speaking test

The corpus data primarily comprised reading comprehension passages, scripts from the listening sections of tests, and writing adapted from published reference books (Table 17) in the CET 4.

An excerpt from the Writing Part (1) corpus (Table 17) is presented below. It is from an article titled *Should the University Campus Be Open to Tourism?*

'Should the University Campus be Open to tourism?

Nowadays, many famous university campuses have become popular tourist attractions. It has been shown on TV and on the radio that every year thousands and thousands of middle school students visit Tsing Hua University, Peking University, and other famous universities in China. In places far away from the capital city, local students also visit universities that are famous in their provinces.

As the present situation is concerned, is it a good or bad thing to open the university campus to tourists? Different people have different opinions. On one hand, some people argue that it is a good thing for the students to visit a famous university campus in that it can enable middle students to get more information about the university, and they can have enough time and opportunity to prepare themselves for a chance to get into the university. On the other hand, some people hold a negative view of this phenomenon. In their opinion, public tourism will have a negative effect on the universities because it will not only do harm to the environment but also to the intellectual atmosphere.

In my opinion, tourism at universities is not a good thing. The campus is mainly a place for study. Increasing tourism on the campus will ruin the spiritual atmosphere in this learning field.'

A significant proportion of the DIY corpus data were adopted and adapted from past exam papers that were considered authoritative and used standard language. Adjustments were made based on the needs expressed in the pilot study, i.e., the students wanted the material to be more directly related to their future English exams. The first major part of the data for the

DIY corpus was pulled from the CET 4, which was first held in 1987 and is sponsored by the MOE. The CET 4 is conducted by the national examinations centre and is a well-regarded large-scale standardised test in China. It serves as a nationwide test with the goals of advancing college English instruction, measuring English competency in an accurate and objective manner, and offering support for enhancing college English course instruction. The CET 4 is reputable because over 20 million English learners take the test every year. Some universities treat passing the CET 4 as a requirement for non-English majors to graduate. The second major part of the DIY corpus data were extracted from the Gaokao English test, which typically comprises two parts: a computer-based English listening and speaking test (CELST) and a written test (Table 17). The CELST is comprised of three sections: reading aloud, role playing, and retelling. The written test includes elements such as cloze, reading, error correction, and writing. The Gaokao English test is designed and administered by the MOE and is held every June. Forty articles (Table 17) were adapted from a book containing 40 essays embedded with the most essential vocabulary in Gaokao English tests. Although commercially produced materials are the property of the publishers and were inaccessible to teachers, there was a large number of open-source corpora freely available online (Charles, 2019). Most of the data used in this study were pulled from open-source corpora. Because the DIY corpus data were not used for commercial purposes and will not be published, and most of the data are open to public use, no copyrights are being violated.

4.3.1.2 Essay Writing

Based on the Involvement Load Hypothesis (ILH; Laufer & Hulstijn, 2001), the higher the involvement, the better the outcome. When composition writing is compared with reading (reading with fill-in-the-blank exercises for target words), composition writing yields significantly better outcomes than reading for studying target words (Laufer & Hulstijn, 2001). Zou (2017) reached similar conclusions in her comparative study of cloze exercises, sentence writing, and composition. Considering that the contents of all four course units were taught during the semester and that the final exam also included a similar writing exercise, essay writing was adopted to facilitate students' vocabulary learning and as the primary data collection approach in this study. To match the teaching content, there are five writing tasks (Appendix II), which are primarily practical writing exercises (letters) based on the textbooks in use and the syllabus standard (*College English Standards for Higher Vocational Education,* 2021). The first four tasks were administered in the context of daily teaching. The fifth composition was on writing a reply to a complaint letter; the exercise was administered during the final exam as the final test for the research.

Table 18. Writing tasks and the corresponding topics in the textbook used at SZIIT.

Writing task	**Topics in the textbook used during the semester**
1). Letter to a foreign friend who wants to study Chinese.	Unit 1 - Learning English: An Easier Way By singing, watching films, and through

	news
2). Short essay on how to best handle the relationship between parents and children.	Unit 2 - Building Relationships Maintaining relationships and learning to make up
3). Letter complaining about the service at a hotel.	Unit 3 - Travelling Abroad Sharing travel stories, eating out, etc.
4). Letter replying to a complaint.	Unit 4 - Selling How to apologise
5). Writing a reply to a complaint letter (depending on the actual final exam)	

The first task involved writing a letter to a foreign friend who wanted to study Chinese. The essay was an exposition adapted from the Gaokao English exam, consistent with a topic in the textbook used in the course: *Learning English: An easier way*. The units included lessons such as learning English by singing, learning English by watching films, and learning English through the news. The second writing task required students to write a brief essay about the best method to handle parent-child relationships. This assignment, which aligns with Unit 2 (Building Relationships), was modified from the CET 4's writing part. The third task was writing a letter of complaint about a hotel at which they had stayed. The fourth writing assignment was selected from past papers of the Practical English Test for Colleges (PRETCO), which used to be a prerequisite for SZIIT students to obtain a degree. The task was inspired by the topic of travelling in Unit 3. The fifth task involved writing a letter

replying to a complaint, which was a crucial part of the textbook series because it is considered commonplace and is an often-utilised skill in the field of practical writing, as indicated by the syllabus. The exercise had also been used in the end-of-term exam in previous school years.

4.3.1.3 Target Vocabulary

AntConc, a free tool developed by Anthony (2020), was used to develop the frequency list of the target vocabulary with reference to the institute's English material. *AntConc* is a free offline retrieval search engine that can search corpus data files. This software primarily provides concordance lines with the keyword in the middle, known as the *keyword in context* (KWIC). A snapshot of a search for the verb *value* in *AntConc* is presented in Figure 8.

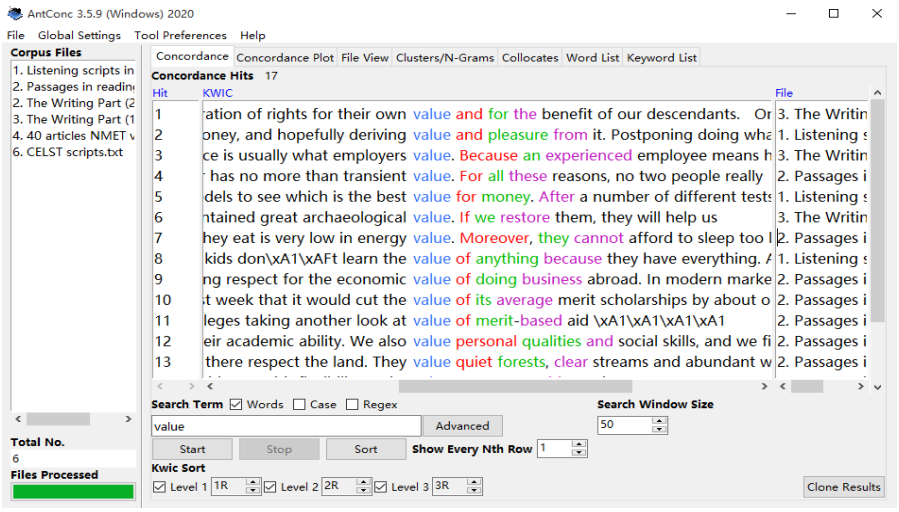


Figure 8. Screenshot of the *AntConc* search function used in this study

The *Word List* function in *AntConc* helps with generating a word list by retrieving a word from the texts (i.e., the corpora data compiled by the researcher) and deciding the target verbs in this

study is presented in Figure 9; the words are ranked in order of their frequency.

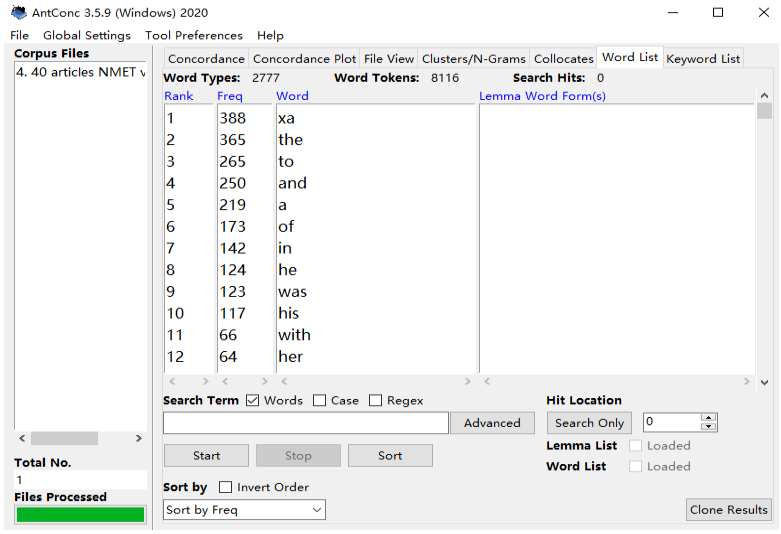


Figure 9. Screenshot of the *Word List* function of *AntConc* used in this study

The target vocabulary (Table 19) was selected based on the final exams for each school year using the frequency analysis feature in *AntConc*. The text files contain all the reference articles provided by the institute, and it was found that *take*, *cause*, *apologise*, *expect*, *deliver*, *chat*, and *serve* were the most frequently used verbs (n = 4–7). The verbs used in Writing Task 1 for familiarisation with the corpus tools are not on the frequency list, while the target vocabulary in Writing Tasks 3 and 4 is all from the frequency list. The target vocabulary is vital for kindling students' interest in learning for the purpose of passing the exam, as students can build on these verbs to write a composition that includes the essential information required in the complaint letter in this study.

Table 19. Target vocabulary for each writing task.

Writing task	**Required vocabulary embedded in the writing**

1	participate, read, watch, listen, sing, make friends
2	value, identify, compromise, order, return, send, assure
3	take, cause, apologise, expect, deliver, chat, serve
4	establish, locate, convince, give, spot, submit, confront

4.3.1.4 Collocation Error Identification

This study focuses only on verb collocation. Previous research generally indicates that students often make mistakes in verb collocations (Fang et al., 2021; Wang & Zhou, 2020; Yang et al., 2005). These mistakes include the inappropriate use of the verb tense, incorrect spelling of verbs, inappropriate use of the third-person singular, and incorrect verb structures (e.g., verb + prepositions and verb + nouns). The number of verb collocation errors was calculated and then divided by the total number of words in the essay; the percentage values were compared. Because the fuzziness in the subject of collocation is relatively substantial, and this area of language has by no means been comprehensively described (Nesselhauf, 2005), the researcher identified any incorrect use of verb collocations, and this was crosschecked by another English teacher at the same institute. However, when the degree of acceptability of the combination was in dispute, dictionaries like the Oxford Collocations Dictionary, corpora like the BNC or the COCA, and a native speaker (a foreign tutor at the same institute) were also consulted.

4.3.1.5 Measuring Vocabulary Using the Vocabulary Knowledge Scale

The vocabulary knowledge test for the target vocabulary embedded in the writing task was adopted from research by Paribakht and Wesche (1996). According to "selective attention, recognition, manipulation, interpretation, and production" (Paribakht & Wesche, 1996), the entire test is divided into distinct classification schemes. It is based on the information processing framework for L2 knowledge acquisition. The VKS assesses students' word knowledge at several levels, from complete unfamiliarity (through recognition and meaning) through the use of target terms in sentences, using a five-point scale. Because the length of time needed to complete the VKS varies depending on the target words, it is an efficient choice for this investigation. Additionally, it enables students to show that they have some knowledge of each item, which lowers anxiety among students who don't speak English well. For example, in the third level of the test, given the question: "3. I have seen this word before, but I think it means _____ (synonym or translation)", students can choose to fill in the blank with a synonym or the equivalent word in their first language to gain the full score. However, they did not need to use the target vocabulary correctly in level 5 of the scale. The VKS is adopted extensively by researchers (Lee & Lin, 2018), indicating that it is a scale with high reliability. Before the pretest, the researcher gave explicit instructions in Chinese, and assistance was also given as the students filled out the scale.

4.3.2 Quantitative Data Collection Procedure

Because this study utilises an explanatory sequential mixed method, quantitative data collection was implemented first, followed by qualitative data collection. As mentioned earlier, the adopted quasi-experimental research design demands that effort be made to minimise the differences between the two groups on the pretests. Thus, the first stage of the data collection involved preparations for the experiment, including corpus compilation and piloting the instruments. The second stage focused on corpus learning sessions and more posttests, with some basic data analyses conducted right at the end of the experiments, and then qualitative data collection. Finally, the third stage focused on data cleaning and mining. In sum, the entire data collection procedure was designed to be conducted in three stages.

Table 20. Data collection procedure.

Stage 1: Preparation (6 months before Stage 2)	Select the participants and their writing tasks √ Plan the corpus learning sessions √ Compile the DIY corpus data √ Use *AntConc* to select the target vocabulary √ Prepare the DIY corpus handouts √ Pilot and administer questionnaires and tests√ Ready the platform for data collection √
Stage 2: Experiments + data collection (complete 19-week	Before training: Use a pre-survey to collect personal information, information on learner autonomy, student consent, etc. （Week 1 & 2） √

semester)	Analyse personal information to ensure no significant differences. Corpus training sessions: Pretests, DIY corpus learning sessions, writing tasks, collect essays, and posttests (Weeks 7, 12, 15, and 18). After training: Use questionnaires to collect information on learner autonomy and students' perceptions of using DIY corpora. Plan and conduct a follow-up interview with the study participants (Week 19).
Stage 3: Data cleaning + analyses	Grade essays, and double marking. Count verb collocation errors. Paired/independent sample t-test. Code the qualitative data.

Before the experiment, the students completed questionnaires capturing their personal information. English proficiency was carefully checked as a critical variable, and no significant differences were found. For the pilot study, the questionnaire on students' perceptions regarding their English learning responsibilities had a Cronbach's alpha score of >.89 (subscales for perceptions of responsibilities for English learning were >.90, and those perceptions for the frequency of English learning outside of the classroom were >.87),

indicating a high level of reliability. Most importantly, informed consent was obtained from every student in person during lessons in Weeks 1 and 2—before the corpus learning session.

4.3.2.1 Corpus Learning Sessions

The corpus learning sessions were designed with four steps, as reviewed in the literature above. Recently, Ma et al. (2022) designed a corpus-based lesson with four steps: "(1) testing student knowledge, (2) hands-on corpus search by students, (3) inductive discovery by students, and (4) output activities". The lesson for the corpus learning session in this study was also designed in a four-step format (altered adaptation of the Ma et al. format), in which Steps 2 and 3 were different for the experimental and control groups (Table 21). Based on the implications of the findings in the pilot study, students are reluctant to bring their laptops to the classroom to perform the corpus searches, and many of them did not even have a computer at all. In addition, corpus websites and tools are not user friendly on mobile devices. Thus, Step 2 is different in the main study: the students study handouts with concordance lines and deduce the patterns needed. In Step 3, the students are supposed to write and revise their essays based on their deductive perceptions of the handouts. It is not only a deductive discovery, but also a 'pushed output' (Izumi, 2002) activity, although vocabulary knowledge is the ultimate output in Step 4.

Table 21. Different conditions applied to the experimental and control groups.

Experimental Group	**Control Group**

① Pre-test for target vocabulary	① Pre-test for target vocabulary
② Study corpora handouts	② Study e-dictionaries
③ Students write and revise using the corpora handouts	③ Students write and revise by themselves (without corpora handouts)
④Post-test of keywords in writing	④ Post-test of keywords in writing

Step 1 (Table 21) takes 10 minutes and is intended to arouse students' attention. That is, students are administered the VKS test to kindle their interest in the target vocabulary. The instructor posted quick response (QR) codes on the whiteboard for approximately 10 minutes to provide the scale. The students were then supposed to use their mobile devices (e.g., smartphones) to scan the code and fill out the questionnaire by themselves. The instructor provides detailed instructions and informs them of their right to withdraw their data and other critical information. If they wish to complete the 5th part of the scale using the target vocabulary in a sentence, they also need to provide the translation or synonyms in the 4th part. Because the test was translated into their first language (Figure 10), the researcher only needed to supervise the students to ensure that they completed the test independently and without using any reference tools.

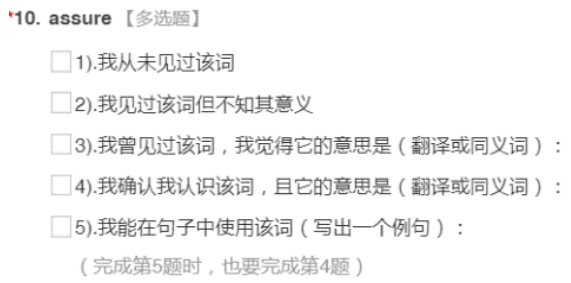


Figure 10. Excerpt on the word *assure* from the Chinese version of the VKS test

Step 2, which takes approximately 25 minutes, focuses on vocabulary studies. The teacher handed out printed concordance lines to the participants and guided them in discovering the word meaning and the target patterns for the vocabulary. In this step, four to seven concordance lines (emboldened and underlined) were provided for each target vocabulary (Figure 11). The students were given 15 minutes for discovery by themselves. The instructor spent 10 minutes demonstrating the use of *AntConc* with all the DIY corpora inserted, and then instructed the students to deduce the most frequent patterns for the target vocabulary and the meanings. The participants were instructed to take notes and learn the target language only utilising DIY concordance lines during this process. However, for 15 minutes, the students in the control group were required to consult dictionaries on their mobile devices. The instructor then displayed similar findings on the whiteboard, stressing some of the frequently asked questions for 10 minutes.

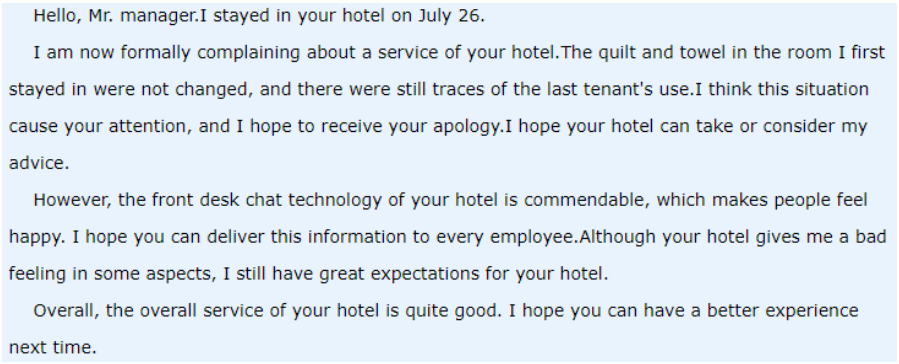
7. serve (60)↵

Welcome to English Club! It could **serve** as a platform to show your outstanding abilities.↵
schools that better **serve** our children and our nation by encouraging students↵
you will probably find a restaurant would **serve** the food of your own native country↵
a waitress will come to **serve** you as soon as you sit down.↵

Figure 11. Handout given to participants for the word *serve*

Step 3, which takes about 35 minutes, is tagged *Practice*. The students were given a writing task in which they used all the vocabulary studied in Step 2. Whenever they encountered problems using the target vocabulary, they were initially encouraged to refer to the concordance lines and their notes. If they were still unsure, they could also request help from

the instructor, who would again explain the vocabulary using instances from the concordance lines. They were allowed 35 minutes for this exercise. The instructor invigilated the students to ensure that they were consulting only the printouts. After completing the writing task, the students submitted their work to a teacher's data collection application, through which the teacher could quickly note their names and generate a list of those who had not finished the composition and broadcast a reminder to them. Students in the control group were free to use e-dictionaries at their leisure.



Hello, Mr. manager.I stayed in your hotel on July 26.

I am now formally complaining about a service of your hotel.The quilt and towel in the room I first stayed in were not changed, and there were still traces of the last tenant's use.I think this situation cause your attention, and I hope to receive your apology.I hope your hotel can take or consider my advice.

However, the front desk chat technology of your hotel is commendable, which makes people feel happy. I hope you can deliver this information to every employee.Although your hotel gives me a bad feeling in some aspects, I still have great expectations for your hotel.

Overall, the overall service of your hotel is quite good. I hope you can have a better experience next time.

Figure 12. Essay submitted by a participant for Writing Task 3 (complaint about hotel service)

Step 4 (10 minutes) is tagged *Produce*. The students retake the VKS scale with the target vocabulary reordered. On this occasion, they are all given 10 minutes, as in Step 1.

Despite the fact that the students were eager to take part in the experiments at the beginning, the pilot study findings indicate that they have a low motivation for English learning and there is a high possibility of students withdrawing from the experiment by not submitting

their compositions. Therefore, to ensure the participation of both groups, the original marking scheme assigns 50% for daily performance and 50% for the final exam papers. Usually, regular performance is awarded for punctuality and participation, and this was adjusted to 25% for after-class assignments, especially writing essays.

Table 22. Adjustments to marking scheme for participants as marked *.

Original version (total: 100%)	**Revised assessment structure (total: 100%)**
Punctuality and participation: 50%	Punctuality: 10% Participation: 15% *After-class assignments (writing essays): 25%
End-of-term exam: 50%	End-of-term exam: 50%

After each corpus learning session, the pretests and posttests on the target vocabulary and essays were collected and stored on a safe platform (supplied in the preparation stage). At the end of the last session, the students retook the learner autonomy questionnaire and reported their perceptions regarding using DIY corpora. As each lesson was recorded, actual participation during each learning session was further analysed. Grading essays and calculating verb collocation errors were part of this stage, and unclear errors were reviewed with the participants whenever possible. Appointments for the interview sessions were scheduled before the end of the course. The researcher carefully observed and facilitated the overall research to ensure participation.

4.3.3 Quantitative Data Analysis Procedure

4.3.3.1. Collecting and Measuring Student Collocation Learning in Essay Writing

Given that there is a fair amount of ambiguity around collocation and that it has not yet been fully articulated, this area of language is particularly fuzzy (Nesselhauf, 2005). Any improper verb collocation was noted by the researcher and verified by another English teacher working at the same institute. Nevertheless, dictionaries like the Oxford Collocations Dictionary, corpora like the BNC and the COCA, and a native speaker (a foreign tutor at the institute) were also consulted when needed to determine the level of acceptability of the combination in question.

Considering that each composition varies in length, it is more difficult to compare the amount of inaccuracies or accuracy of the target verb collocations than it might first appear. The researcher thus adopted Wang and Zhou's (2020) method for examining verb-noun collocations to measure the quality of the writing of the students in terms of verb collocation use. I counted the frequency of correct verb-noun collocations and totalled the frequency of all verb-noun collocations, including incorrect collocations. For instance, if there are ten total verb-noun collocations and only three of them are accurate, the accuracy rate is 0.3 or 30%.

4.3.3.2. Measuring Student Knowledge of Target Vocabulary Items

The data collected, as described in the preceding section, were triangulated. Two experienced English teachers — who did not participate in the experiment — acted as third-party reviewers.

They crosschecked the VKS using the same grading schemes used in the experiment. The final exam writing tasks were graded using the same criteria. The collected data were analysed using SPSS 27.0.

Table 23. VKS scoring scale (Paribakht & Wesche, 1996).

Self-report categories	Possible scores	Meaning of scores
I	1	The word is not familiar at all
II	2	The word is familiar but its meanings is not known
III	3	A correct synonym or translation is given
IV	4	The word is used with semantic appropriateness in a sentence
V	5	The word is used with semantic appropriateness and grammatical accuracy in a sentence

Category I of the VKS leads to a score of 1, which suggests learners have no knowledge of a word. Category II leads to a score of 2, which indicates that learners can recognise word form (i.e., demonstrate knowledge of word form). A score of 3 on Categories III and IV is achieved when a learner provides a correct synonym or the L1 equivalent of a target word (demonstrating knowledge of the word's meaning) or a score of 2 when a wrong synonym or L1 equivalent is provided. A score of 5 on Category V is achieved when learners use a word in a semantically and grammatically appropriate way in a sentence (knowledge of word use) or a score of 4 when they use it in a sentence in only a semantically appropriate way.

The first two research questions were more suited to quantitative data analyses. For Question 1 (To what extent does incorporating a DIY corpus improve writing quality and the use of

verb collocations?), participants completed four different writing assignments related to the textbook required by the syllabus, and each assignment was uploaded and reviewed. The total number of errors (especially those related to verb collocations) in both groups was collated and compared. In addition, mistakes were marked out and triangulated by the instructor. The errors and writing scores were then compared using a paired sample t-test. Question 2 (To what extent do the participants improve their knowledge of target vocabulary within writing tasks after incorporating a DIY corpus?) could be answered using pretests and posttests. With seven to eight target vocabularies embedded in the essays, the entire test was designed in the form of the VKS on an online questionnaire platform Wen Juan Xing (www.wjx.cn). It might be more interesting to work out one after all four compositions as a delayed posttest during the preparation week for the end-of-term examination. The VKS (Paribakht & Wesche, 1996, p. 180) was modified and used in a pretest and an immediate posttest to assess participants' knowledge of the target verbs. The scores for each target vocabulary were used in a between-group comparison of the experimental and control groups via an independent t-test, and within-group comparisons using a paired sample t-test.

4. 3. 3. 3. Measuring Students' Learning Autonomy and Their Reactions to Using DIY Corpus Materials in Questionnaires

Quantitative data were collected using the Likert scale sections of two questionnaires. One questionnaire captured data on students' learning autonomy from two dimensions: the student's perceived responsibilities towards English learning, and the frequency of English learning activities outside the classroom. The second questionnaire was concerned with the

student's reactions to using the target DIY corpus materials. The second questionnaire was administered to the participants in the experimental group only after the experiments because the students in the control group did not receive the DIY corpus intervention. The first questionnaire was administered during the pretest and posttest. After the data were collected using quantitative tools (e.g., questionnaires), sequential qualitative studies in the form of interviews were conducted to generate insights that would explain the initial quantitative results.

The first questionnaire on learner autonomy was designed to quantitatively answer the third research question ('To what extent does the use of a corpus facilitate students' learner autonomy?'). This questionnaire was originally developed by Spratt et al. (2002). It was adapted to examine changes in the participants' learner autonomy before and after the intervention, based on its relevance and efficiency. 'We can measure the variables that are logically related to a construct even when we cannot actually measure it. As a result, the researcher develops an operational definition that details the variables' methods of measurement' (Lodico et al., 2010, p. 25). Spratt et al. (2002) delineated the concept of autonomy into students' perceptions of their teachers' responsibilities vis-à-vis their own responsibilities for various aspects of their English learning, and students' views of their motivation and frequency of engaging in learning activities outside of class. The questionnaire survey was conducted in Hong Kong, and the 508 participants were taking similar majors (e.g., mechanical engineering, and electrical engineering); 20 minutes was assigned for filling out the questionnaire. After adaptation, the questionnaire was translated

into Mandarin Chinese and piloted among students (N = 80) taking the same major; these students did not participate in the main research. The Cronbach's alpha score was over .85. This study investigates the perception of responsibility towards English learning and the frequency of English learning activities outside the classroom. It grades learning autonomy on a scale, which Benson (2013) describes as complex because testing autonomy is an 'anti-autonomy' activity. In the experiments, the students examined any errors or mistakes made during the process of drafting. Most of the corrections were made during their spare time. In addition, the time spent drafting and rewriting, the materials used in learning, and the decision to accomplish the task in a group or individually all reflect autonomous behaviours. After reviewing the pilot study, it was determined that conceptualising learning autonomy is highly relevant to this study for identifying the differences between the control and treatment groups and between performance on the pretests and posttests for the experimental group. Before the interview, a questionnaire on learner autonomy that assessed students' perceived learning responsibility towards English learning and the frequency of their English learning activities outside the classroom (Appendix IV) was adopted from material in Sections 1 and 2 of this study (Spratt et al., 2002), translated into Chinese, and piloted with some peers of the study participants at the same HVI.

The second questionnaire was modified from study by Nam (2010), who investigated the impact of corpus-based language acquisition on 21 international undergraduate students' knowledge of useful ESL vocabulary at an American public university. The 21 participants wrote seven writing samples using either a concordance or a thesaurus, and their samples

were analysed for changes in writing quality and attitude towards a corpus-based vocabulary reference tool, much like this study. The fourth research question ('What are the participants' attitudes and perceptions towards using corpus data in English writing?') concerns the attitudes and perceptions of students towards using the DIY corpora and improving learner autonomy. A questionnaire was translated into Chinese, and some of its questions were modified to fit this study. After the questionnaire was administered, a semi-focused group interview was conducted to obtain more qualitative data from the students on incorporating corpora data into their writing assignments. Notably, using corpora in the classroom from a student's perspective informs the answers to the fourth research question ('What are the participants' attitudes and perceptions towards using corpus data in English writing?') and corpus-based teaching methods for further research. Participants in the interview had used the DIY corpus tool, and their comments and experiences could be different from those in the literature and that of the researcher. Owing to the explanatory sequential mixed research design and the qualitative data collection being based directly on the quantitative results, the interviewees' individual responses to the dependent variables (Creswell, 2014, p. 330), such as vocabulary knowledge, writing quality, and learner autonomy.

4.4.2. Qualitative Data Collection

Finally, one semi-focused group interview was conducted with participants in the experimental group to assess corpus usage, feedback on the adopted approach, and the degree of learner autonomy among participants in the experimental group. This semi-focused group

interview was conducted to obtain more qualitative data on incorporating corpora data into writing assignments. The entire process of using corpora was approached from the perspective of the student, and it informs the method of corpus-based teaching to be used in further studies. More importantly, the qualitative data, in combination with the quantitative data, can provide additional understanding of the phenomenon and research questions, achieving 'hybrid vigor' (Dörnyei, 2007; Creswell, 2012).

Before the interview, a question concerning the participants' availability to attend an interview was added to the second questionnaire, asking if they would like to attend a face-to-face interview with the instructor before leaving campus for the summer holiday. After a confirmatory follow-up on the respondents regarding the question, five eventually agreed to attend the interview. The number of interviewees was five, and their collective wisdom should provide more qualitative data because interviewees can collaborate on ideas, motivate and push one another, and respond to newly developing problems and ideas (Dörnyei, 2007, p. 144). The time and location of the interview were later agreed to by the interviewees. Informed consent for sound recording and anonymous use of the interview data was obtained before the interview started. A smartphone and sound recorder were used for the recordings. Some of the materials used during the experiments were printed as realia for the interviewees to reflect on the preceding experiments in the study. These materials included DIY corpus concordance printouts, the two questionnaires, and some of their VKS results. During the interview, the interviewees were allowed to interrupt their peers if they had anything to add.

Some guiding questions were asked in response to the research questions. The first question was about their writing quality and the accuracy of their verb collocation: 'Did you notice your writing proficiency has improved, especially that the verb errors decreased significantly? How or why did this happen?' And 'The VKS tests show significant difference from that of the control group. How or why did this happen? From your perspective, is this difference related to the use of concordance printouts?' was asked concerning vocabulary knowledge within the writing tasks. Subsequently, the following questions were asked in relation to the third question about the interviewees' learner autonomy change: 'It was found that your perception about learning responsibility towards English learning has improved significantly. Do you feel the same way? Why do you think so, or why not?' Finally, questions were also asked during the interview regarding whether the interviewees used *AntConc* after the training on its functions in or after class and whether they read the DIY corpus printouts after class. By the end of the interview, the interviewer had also asked the students if they had more to add or anything else they wanted to share with the teacher.

Considering the low English proficiency of the interviewees, the interview was conducted in Chinese, their L1. The transcribing began immediately after the interview, and transcriptions of the qualitative data were sent to the interviewees to check if their intended meanings were clearly and appropriately conveyed in the texts. After receiving confirmation, the transcription was translated into English and subsequently double-checked by an English teaching fellow at the same institute.

4.4.3 Qualitative Data Analysis Procedure

The students' reactions towards the DIY corpus-based writing were collected via questionnaire, and each term in the survey was weighted and subsequently checked during the interview.

After the interview, the data were transcribed into English and then given to the interviewees to confirm that their intended meaning had been retained. Before the data analysis, all the interview transcripts and questionnaires were read thoroughly multiple times until the researcher had a broad understanding of the major issues and ideas expressed in the transcripts. The data then underwent content analysis and a coding process comprising three types of coding (open, axial, and selective) based on grounded theory, and were subsequently reread by a colleague acting as a third-party reviewer (Richards, 2003).

Content analysis is defined as a strict and systematic set of procedures for summarising and reporting written data — primarily the contents of the data and their message (Cohen et al., 2007). Any written work can be subjected to content analysis, which focuses on language, linguistic aspects, and meaning in the context (p. 475). Because its analysis rules are explicit, accessible, and available to the public, it is systematic and verifiable. Furthermore, verification via reanalysis and replication is possible because the data are permanently in text form.

Content analyses, as Cohen et al. (2007) propose, involve coding, categorising meaningful

units like words, phrases, and sentences, and then comparing categories and making links between them. Finally, theoretical conclusions can be drawn from the text while presenting the analyses in as economical a form as possible — mentioned earlier as the fourth method.

Considering that the qualitative data in this study were captured using two questionnaires and a semi-focus group interview, the qualitative data acquisition serves as the second phase of the explanatory sequential mixed methods research design, and it is rather straightforward which is followed by 'progressive focusing'. The salient characteristics of the scenario become apparent after the researcher collected data using a wide-angle lens and then sorted, reviewed, and thought about it. In this regard, the qualitative data analysis in this research comprised the following five steps: After transcribing the data on learner autonomy and perceptions on using the DIY corpus during writing exercises, the researcher first extracted the interpretive written comments in the data using open coding, and then sorted the data into key headings using axial coding. In the third step, the researcher listed the topics within each key heading and noted the frequency of the items being mentioned. In the fourth step, the researcher went through the list generated in step four and dealt with issues into groups to avoid category overlap. In the fifth step, the researcher commented on the results in the fourth step and reviewed the interviewees' messages.

In addition to content analysis, grounded theory — as a more inductive method — was also adopted across the entire qualitative data analysis process. All three types of coding: open, axial, and selective coding, were adopted to facilitate a deeper understanding of using DIY

corpus. Exploring the data and selecting units of analysis to code for meaning and actions were part of open coding. In order to properly code the data, the researcher added additional codes, categories, and integrations where appropriate. In order to evaluate how closely related categories are to one another, the researcher used axial coding to find connections between categories and codes. The interrelationships were examined, and codes and categories were compared based on existing theories. Finally, the selective coding involved identifying a core code, and the relationship between that core code and the other codes was made clear. The 'story line' (p. 493) was identified to integrate the categories into an axial coding model.

An essential component of data reduction and selection is careful data display. The first way to organise a qualitative data analysis is by groups. This involves automatically grouping the data and enabling themes, causing patterns that are similar to be observable at a glance. However, this method is often used in a single-instrument approach; otherwise, it becomes unwieldly. The second way of organising a qualitative data analysis is by the individuals. Following the presentation of each participant's responses, attention shifts to the following person. The qualitative data in this study came primarily from a semi-focused group interview, which was not suitable for this type of organisation. The presentation of all the data pertinent to a certain subject is a third method of organizing qualitative data. When comparing respondents, this strategy is cost-effective, but because the data is somewhat decontextualised, the integrity, wholeness, and coherence of each individual respondent may be compromised. By research question is the fourth way to organise the study of qualitative data. This approach would gather all pertinent information on the precise subject that

interested the researcher and maintain the material's coherence (e.g., the questionnaires and interview transcripts in this study), and it might offer a comprehensive, systematized response to the research question. By instrument is a fifth way to arrange the data. The earlier indicated methodology is typically utilised in conjunction with this one. Ultimately, it would become particularly challenging for the researcher of this study.

Of the five ways of organising and presenting qualitative data analysis, the fourth method was adopted because of its appropriateness for this study. Three instruments were adopted to facilitate the analysis of the qualitative data, which may not be rational to adopt if using the first method of organising data analysis (by groups). Furthermore, the interview was conducted in a group, and the data generated by each individual were inadequate for the second method of presenting data (by individuals). In addition, the economy of the third method of organising data (by a specific issue) weighed less than the importance of the contextualisation of the participants' responses. The fifth method was too challenging because it would require more than one approach to be used at the same time while organising the data by instrument. The collective answers from the questionnaires and the interview could help achieve systematisation. Thus, the fourth method was adopted.

4.5. Ethical Considerations

Researchers use different guidelines for ethical concerns in research studies (Creswell, 2012). Considering these various perspectives, the first vital step is not to disclose any information

that will harm the participants. Following the guidance of the Human Research Ethics Committee (HREC) of the Education University of Hong Kong, the essential information needed for informed consent was presented in the consent form. The researcher explained the introduction, methodology, and potential risks of this study to the vice dean of the School of Foreign Language Studies of the Shenzhen Institute of Information Technology in detail and obtained her permission to access the research setting. The participants agreed to the educational experiments by signing informed consent sheets. Anonymity has been maintained, and pseudonyms have been used to protect the participants. However, the research objectives were revealed to the participants to obtain their support. The data, including the interview recordings and transcriptions, have been stored safely.

Chapter 5: Results

Both groups of participants in the present study majored in a similar science field, and their English proficiency indicated no significant difference ($p = 0.68$, two-tailed) through an independent sample t-test comparing their scores in the Gaokao English test. In addition, the number of participants varied in different corpus learning sessions for various reasons. First, some students dropped out of school for work or other personal reasons. Second, some students changed their majors during the research of the second semester, which was conducted according to the method design outlined previously. Third, some students did not complete the target tasks, such as the vocabulary knowledge scales and writing tasks.

The quantitative results were mainly computed in SPSS version 27 for data analysis, and the results are presented according to the research questions. This draws all the relevant data to each issue of concern and preserves the coherence of the material, such as the questionnaires and interview transcripts, in the present study. This systematically provides collective answers to the research questions.

The qualitative data were coded by the researcher and independently by a colleague of the researcher, and the inter-coder reliability reached almost 90%. The codes were then merged to generate multiple themes in the subheadings of the material after all of the content's contested situations had been settled through discussion. The five interviewees in the semi-structured focus group interview are referred to by the pseudonyms Jack, Tom, Jerry, Alan and David to

protect the interviewees' privacy. The interviewees came from the experimental group. Except for Alan, the others had all learned English as a foreign language in high school. The interview lasted for 60 minutes and was conducted at 3 p.m. in a meeting room after conferencing with the interviewees.

Table 24. Demographic Information for Interviewees

Name	Age	Home-town	Major	Years of English Learning	Daily English Learning in Hours	Difficulty of English Learning	Motivation to Learn English
Jack	20	Guang Dong Province	Mobile Communication Technology	12	0 ~ 1	Vocabulary	Not at all
Tom	21			9			Slightly
Jerry	20			15			Slightly
Alan	20			6*			Not at all
David	19			12			Slightly

*This participant switched to learning Japanese as a foreign language for the College Entrance Exam.

The interviewees participated in the interview voluntarily after appointments with the researcher. Table 24 displays interviewees' personal information, gleaned after they gave informed consent. The participants were all around 20 years old, and they all came from cities in Guang Dong Province. They all majored in mobile communication technology. They each

had about 10 years of English learning experience, except for Alan, who gave up English and chose to study Japanese for his College Entrance Exam. Although the interviewees received many years of compulsory English education, they still found vocabulary to be the most difficult to learn, and they spent almost no time on English learning outside class hours. Furthermore, they were all either unmotivated or only slightly motivated to learn English.

5.1 Writing Quality and the Accuracy of Verb Collocations

Research Question One: To what extent do participants improve their knowledge of target vocabulary within writing tasks after incorporating a DIY corpus?

The first research question aimed to reveal the effective incorporation of the DIY corpus on students' writing quality in terms of the grading scheme adopted at the researcher's institute and the number of target verb collocations used. The data for this question came from the grades given to participants' writing practices and verb collocation calculations.

It is relatively easy to mark out correct verb collocations by referring to dictionaries, corpora and native speakers. However, comparing the number of these errors is not as simple as it may seem because the length of each article is different. Thus, to quantify students' writing quality in terms of verb collocation use, the researcher followed Wang and Zhou's (2020) method for studying verb-noun collocations. They calculated the total frequency of all verb-noun collocations (including wrong collocations) and counted the frequency of accurate verb-

noun collocations. For example, if the total number of verb-noun collocations is ten and the number of accurate collocations is five, then the accuracy rate is 5/10=50%. After multiplying by 100, the total correct score is five.

In addition, the number of participants in the first four writing tasks varied from time to time because some did not submit their essays due to illness or other valid reasons. However, the fifth writing task was conducted as the last part of the final exam. The 31 students in the control group were supposed to take the exam; however, two of them dropped out, and four of them just copied and pasted the requirements of the writing part, where students were supposed to skim and reply to a complaint letter, instead of writing their own essays. Thus, there were only 25 valid scores. Similarly, the 46 students from the experimental group were supposed to take the final exam, but four changed their majors, two were Japanese or Spanish learners, and two did not take the exam, leaving 38 valid student writings for further data analysis.

5.1.1 Writing Quality

The writing grading scheme at the researcher's vocational institute was adopted to grade students' essays since it has been validated and used for many years in this institute, and it is highly recommended and accepted by English teachers. Moreover, the scores for every composition submitted by the participants were also checked by a third party, another English teacher, who ensured the grading scheme's equal application.

As previously mentioned, the experiments mainly included four writing tasks, and the fifth task was the written essay for the school's final examination. For ease of presentation, the data are displayed in Table 25.

Table 25. Descriptive Data of the Gradings for Writing Tasks

Task	Group	N.	Mean	Max.	Mini.	SD
1	Control	30	13.94	16.5	10.8	1.436
	Experimental	38	13.90	15.8	8.5	1.586
2	Control	13	14.42	17.4	7.5	2.894
	Experimental	30	14.81	17.3	9.8	1.635
3	Control	20	15.09	16.6	11	1.272
	Experimental	36	15.35	16.5	10.5	1.191
4	Control	20	13.68	15.5	9.9	1.527
	Experimental	31	13.42	15.9	7.1	1.536
5	Control	25	8.98	15.0	0.0	4.765
	Experimental	38	13.00	17.0	0.0	2.493

As shown in Table 25, the number of participants in each group varied in different essay writing sessions. The size of the control group ranged widely, from a maximum of 30 to a minimum of 13. However, there were originally 38 students in the experimental condition and the majority of them (30) stayed throughout all writing sessions. In contrast to the low participation rate of the control group, this suggests that the participants in the experimental

group were motivated and eager to try out the novel method, corpus-based writing. The scores of their writing tasks did not display a significant difference until the final exam, that is, the fifth writing task. As seen in Table 25, a similar performance of the previous four writing tasks could be seen from their mean scores. Moreover, the scores' deviations were also similar, though sometimes (task two and task five) the standard deviation was more significant than that in the experimental group. The first four writing tasks were conducted during the usual classroom hours; participants were free to use their mobile devices during the drafting process and handed in their final versions through a mobile application. The last writing task was part of an examination, precluding the use of mobile devices, and the teachers monitored the students. It is this last session that can truly measure the effect of the use of DIY corpus on students' writing quality.

Table 26. Independent Sample T-test for Gradings in Writing tasks

Task	t	df	Sig. (2-tailed)	Mean Difference (Experimental Minus Control)	Cohen's D
1	0.106	66	0.916	-0.04	N/A
2	-0.554	41	0.582	0.39	
3	-0.742	37	0.463	0.30	
4	0.574	49	0.568	-0.26	
5	-3.879	33	0.000*	4.02	1.13

*$p < 0.01$.

Table 26 clearly shows the differences between the control and experimental group participants regarding their writing scores, as shown by the independent sample t-test. The only significant difference came from the fifth writing task, with a p-value of .000 (two-tailed) lower than 0.01 and large effect size (Cohen's d=1.13), and the mean difference was 4.02. This means that the participants in the experimental group statistically outperformed their counterparts in the final exam when mobile use was not allowed. Thus, the lack of significant differences between the groups for the first four tasks may be attributed to an extraneous variable, such as the use of mobile phones on which students could browse for translations and use word processing software. The examination environment did not allow students to follow these habits. In other words, corpus-based learning did improve the quality of their writing, though it did not result in significant differences in daily exercises.

5.1.2 Accuracy of Verb Collocations in Writing Tasks

Despite the fuzziness of the definition of collocation as being 'predictable' or '[tending] to occur together' (Lewis, 2000, p. 42), the researcher turned to large general corpora, dictionaries, and native speakers for reference; additionally, another more experienced English teacher at the same institute checked the identification of the verb collocations. Table 27 describes the descriptive data of the frequency of correct target verb collocations, and Table 28 displays the independent sample t-test results from SPSS version 27.

Table 27. The Frequency of the Correct Target Verb Collocation in Writing Tasks

Task	Group	N.	Mean	Max.	Mini.	SD
1	Control	30	4.01	8.11	2.00	1.511
	Experimental	38	3.51	9.43	0.72	1.473
2	Control	13	5.04	10.34	0.00	2.655
	Experimental	30	3.29	8.97	0.00	2.323
3	Control	20	3.80	7.81	1.06	1.526
	Experimental	36	3.15	5.34	0.00	1.244
4	Control	20	2.58	6.09	0.00	1.725
	Experimental	31	3.63	7.69	0.00	1.706
5	Control	25	0.43	1.64	0.00	0.547
	Experimental	38	1.35	3.92	0.00	0.878

As shown in Table 27, the number of participants fluctuated from time to time, but the control group's second writing activity had the lowest number of participants overall. Of all the other participants, who did not submit the writing task, these 13 were relatively good students, and the maximum number of correct verb collocations peaked at 10.34 for every 100 words. The minimum number was 0 in some of the writing tasks, as seen in Table 27, and this phenomenon could be attributed to students not cooperating and following the teacher's instructions or being unwilling to write or incorporate the new approach. Unlike the first four tasks, the participants used the least correct target verb collocations in the fifth writing task. The reason for this could be the exam circumstances; students did not have a chance to refer

to writing tools. They had to rely on their memory of the target verbs instead, indicating that they retained different vocabulary levels. Considering that the final exam was held about one and a half months after the experiment, this decrease was deemed acceptable. As seen in Table 27, the lowest mean of correct target verb collocation frequency was 0.43 for every 100 words.

Table 28. Independent Sample T-test of the Frequency of Correct Target Verb Collocations between the Control Group and the Experimental Group

Task	t	df	Sig. (2-tailed)	Mean Difference	Mean of the C.G.	Mean of the E.G.	Cohen's D
1	1.379	66	0.173	-0.50	4.01	3.51	N/A
2	2.163	41	0.582	-1.75	5.04	3.29	
3	1.701	54	0.095	-0.65	3.80	3.15	
4	-2.141	49	0.037*	1.60	6.09	7.69	0.613
5	-4.673	61	0.000**	0.92	0.43	1.35	1.20

*p <0.05; **p <0.01.

Note: C.G. is short for Control Group and E.G. is short for Experimental Group.

The results of the independent sample t-test of the frequency of correct target verb collocations between the two groups are broadly consistent with the writing scores; the experimental group did well towards the end. As seen in Table 28, the p-value (two-tailed) of the first three writing tasks regarding the frequency of correct target verb collocations was

much higher than .05, indicating no statistically significant differences. However, in the data from the fourth session and the fifth session (the final examination), the p-value (two-tailed) was lower than .05 or .01, indicating a significant difference. In addition to the p-value, the effect size is medium for the fourth writing task, while the fifth writing task exhibited a much bigger effect size with Cohen's D value 1.20, which is bigger than 0.80. This, on one hand, suggests that the use of mobile phones may have led to insignificant results for the first three sessions (similar to the T-tests for writing quality; see Table 26). On the other hand, the result may also suggest that a more extended study may achieve better results. At the same time, the participants in the experimental group were able to retain more vocabulary knowledge and put it into writing in the exam. Compared with their peers in the control group, they were all low-proficiency English learners, and they were all unwilling or reluctant to embrace the changes in the corpus-based writing approach.

Table 29. The Frequency of Erroneous Verb Collocations in Writing Tasks

Task	Group	N.	Mean	Max.	Mini.	SD
1	Control	30	1.54	7.00	0	1.89
	Experimental	38	2.05	7.55	0	1.92
2	Control	13	0.28	1.72	0	0.56
	Experimental	30	0.79	4.51	0	1.20
3	Control	20	2.52	7.81	0	2.28
	Experimental	36	3.2	9.30	0	2.45
4	Control	20	0.92	4.03	0	1.16

	Experimental	31	1.00	4.81	0	1.38
5	Control	25	2.98	7.02	0	2.25
	Experimental	38	2.34	5.56	0	1.55
All	Control	108	1.79	7.81	0	2.04
	Experimental	173	1.95	9.30	0	1.97

Table 29 displays the frequency of erroneous verb collocations in each writing task. From the mean frequency of the erroneous verb collocations in Table 29, it may be observed that the participants in the control group made fewer verb collocation mistakes than the experimental group in the first four writing tasks. However, the situation reversed in the fifth writing task. The fifth task was one part of the final examination, and the participants were closely supervised by the teachers. The participants in the experimental group were only permitted to refer to the DIY corpus concordance printouts, while the students in the control group could browse their mobile e-dictionaries as they were used to. Since it was the first time the participants in the experimental group used the DIY corpus concordance lines, they needed time to acclimate to the experience of using the printouts and incorporating the materials into their writing processes. This contributed to the relatively higher maximum number of erroneous verb collocations, more than nine every 100 words; however, the maximum number for the control group was near eight. As is shown in the last row in Table 29, the researcher added up erroneous verb collocations all five writing tasks, and the mean number of wrong verb collocations became close between the two groups. However, the number of participants who submitted compositions varied more widely, between 108 and 173, though

N = 40 for the control group and N = 46 for the experimental group at the beginning of the research. Despite the allowed use of habitual reference tools for the control group, individuals in the control group were much more reluctant to engage in writing practices. The participants in the experimental group were more willing to step out of their comfort zones to exploit the new tool and strove to complete the writing tasks using the new tool.

The frequency of erroneous verb collocations depicted in Table 29 appears to be related to the data in Table 25 and Table 27. These data are presented as one story for the first four writing tasks but as different for the fifth task. Thus, the researcher computed the Pearson correlation coefficient via SPSS, as shown in Table 30.

Table 30. The Pearson Correlation Coefficient Test between Writing Quality and the Frequency of Erroneous Verb Collocations

Task	Group	N.	Pearson Correlation Coefficient	p. (2-tailed)
1	Control	30	-0.290	0.119
	Experimental	38	-0.641	0.000**
2	Control	13	-0.622	0.023*
	Experimental	30	-0.452	0.012*
3	Control	20	-0.818	0.000**
	Experimental	36	-0.648	0.000**
4	Control	20	-0.353	0.127
	Experimental	31	-0.151	0.417

5	Control	25	0.162	0.440
	Experimental	38	-0.451	0.005**
All	Control	108	-0.242	0.012*
	Experimental	173	-0.309	0.000**

**p < .05; **p <.01*

As shown in Table 30, the gradings and the frequencies of erroneous verb collocations are generally negatively related, indicating that the more verb collocation errors exist, the lower the gradings of a writing task will be. Only the Pearson Correlation Coefficient for the essays submitted by the participants in the control was positive fifth writing (the final exam). This was caused by the two factors. First, a student might fail the exam but still be cooperative and well behaved in the classroom, earning a relatively higher overall grade. Second, the word count for these essays was limited, and students had fewer chances to make verb collocation mistakes. Although the ecoefficiency was insignificant in the fourth essay, the p value was .012 (two-tailed) for the control group and .000 (two-tailed) for the experimental group, as shown in the last two rows in Table 30. Therefore, the frequency of erroneous verb collocations was negatively correlated with corresponding grades, or with writing quality.

5.1.3 Referring to the DIY Corpus Printouts Is an Effective Writing Strategy

The interview data show that referring to the DIY corpus printouts is an effective writing strategy. The students generally believed that the DIY corpus printouts offered them standard or desired examples to follow. When they were unsure about the collocations, especially the

verb + preposition structures, the printouts were able to directly help students. After they skimmed the KWIC (Keyword in Context) concordance lines, the interviewees were able to easily spot the collocations after the target verbs. The concordance lines gave interviewees reference sentences for their writing. The lines not only displayed the usage of the target words but also served as example sentences for students' reference, especially when the meanings of the concordance lines could meet their needs. According to the interviewees,

> Jerry: I could directly imitate this word in concordance lines.
>
> Alan: I could also just refer to the example sentence and see how to use it.
>
> Jerry: Then I would write according to that sentence.
>
> David: Sometimes, when I go through the printouts, I intentionally look for the prepositions that follow certain verbs, just like fixed sentence structures. Because these structures can become helpful when I want to use collocations in writing.

In addition to active involvement in the experimental research, the interviewees paid more attention to the structure of writing compositions. According to Jerry, 'Sometimes when I have other courses, I will think of the writing tasks and even picture the contents of the essays'. In addition, Alan said the concordance lines gave him more 'clues' for his compositions. The concordance lines with key verb collocations displayed in KWIC attracted the attention of the readers. On the one hand, participants could see the usage of the target verb; on the other hand, the meaning conveyed by the concordance lines also provided participants with ideas about where these lines or words could be incorporated into their

essays. The concordance lines could be directly incorporated once they were fit in, saving students the trouble of brainstorming ideas to express in their compositions.

However, some of the students were used to convenient translation apps. Despite being instructed to exclusively use the DIY corpus in the research design, the students in the experimental group were unable to break their practice of using translation tools in private:

> That is, when writing a composition, I don't know how to write sentences that I don't know how to write smoothly. I can use *Youdao* translation [a mobile application], input Chinese, and then several English sentences appear. This way, I can choose a more appropriate sentence translation according to my article. (Alan)

5.1.4 Summary

The grades of the writing compositions (20 total) imply that the writing quality was generally in consensus with the frequency of correct target verb collocations in the five writing tasks. The first three tasks did not show any statistically significant differences between the groups, but the fourth task, incorporating correct target verb collocation frequency, began to show substantial differences between them. Moreover, the fifth writing task was held under exam conditions in which no reference materials were available. The students in the experimental group not only obtained better grades (indicating better writing quality) but also had higher frequencies of correct target verb collocations (higher rate of target vocabulary retention).

Many factors led to the later significance that the significant difference appeared in the fourth or fifth writing task. The participants in both groups were lower-proficiency English learners. They had relatively bad English learning habits, such as relying too much on translation applications and being reluctant to write. Once the participants experienced the benefits of incorporating the DIY corpus, however, they began to voluntarily use the newly introduced approach for English writing and learning. These factors contributed to the fifth writing task's statistically significant difference between the groups.

In short, incorporating the DIY corpus into writing processes improves writing quality and target verb collocation use.

5.2 Target Vocabulary Knowledge

Research Question Two: To what extent do participants improve their knowledge of target vocabulary within writing tasks after incorporating a DIY corpus?

The variance in the number of participants for each writing task showed even more significant change than for the writing tasks outlined in section 5.1. To calculate the mean differences of the vocabulary knowledge scale (VKS), the researcher needed to confirm that the subjects took both the pre-tests and the immediate post-tests. The students took all four pre-tests and the delayed post-test. The data computed in SPSS version 27 are displayed in Tables 31 and 32 for immediate learning effect and 33 and 34 for vocabulary retention. Mean

difference 1 refers to the VKS scores between the immediate post-test and the pre-tests; mean difference 2 refers to the VKS scores between the delayed post-test and the pre-tests.

Table 31. Descriptive Data of Mean Differences between Pre- and Immediate Post-VKS

Task	Group	N.	No. of Words	Mean of Pre-Test	Mean of Immediate Post-Test	Mean Difference 1 (Immediate Post Minus Pre)	SD
1	Control	24	6	24.33	23.92	-0.42	6.52
	Experimental	10		20.70	23.40	2.70	4.90
2	Control	14	7	21.64	22.21	0.57	4.52
	Experimental	20		18.40	20.70	2.30	5.15
3	Control	9	7	25.11	27.33	2.22	6.82
	Experimental	16		20.00	23.25	3.25	3.75
4	Control	14	8	23.36	23.43	0.07	4.68
	Experimental	26		22.77	26.46	3.69	4.48

Table 31 describes the data of the VKS tests for the performance of both groups in the pre-tests issued before the writing tasks and the immediate post-tests given to participants after they completed the writing tasks. Only the first task for the control group exhibited a negative improvement close to zero (-0.42), indicating that these students did not improve their vocabulary knowledge in this case.

The scores of the immediate post-tests were frequently larger than those of the pre-tests,

though the participants in the control group improved less than their peers in the experimental group. Thus, an independent sample t-test was computed to determine whether there was a significant difference, as shown in Table 32.

Table 32. Independent Sample T-test between Difference of Mean Difference 1 in VKS between the Control Group and the Experimental Group for Immediate Post-Tests

Task	t	df	Sig. (2-tailed)	Difference of Mean Difference 1 (Experimental Minus Control)	Cohen's D
1	-1.356	32	0.185	3.117	N/A
2	-1.011	32	0.319	1.729	
3	-0.490	23	0.629	1.028	
4	-2.368	26	0.026*	3.621	0.80

*p < .05

As seen in Table 32, there was a significant difference in the fourth VKS test. The participants in the experimental group outperformed their counterparts in the control group in the fourth VKS test. The contrast of mean difference 1 was over 3.6, with a p-value of .026 (two-tailed), lower than .05. Moreover, the effect size is a large one (Cohen's d = 0.80). The students in the experimental group all grew their vocabulary knowledge more than their counterparts in the control group, despite the fact that there were no significant differences between the groups on the first three VKS assessments. It was thus found that incorporating the DIY corpus into writing practices improved target vocabulary knowledge based on

immediate post-test scores.

Regarding vocabulary knowledge retention, the researcher conducted a delayed post-test with a gap of five weeks after the last writing practice. The delayed post-test had fewer vocabulary items (20) than the combined quantity of words in the pre-tests. Considering that some of the words were already known to the participants in both groups through the two rounds of VKS tests, there was no need to re-test them. Furthermore, a VKS test with too many words could drain participants' attention as these students were low-proficiency English learners.

Table 33. Descriptive Data of Difference between Pre- and Delayed-Post VKS Tests

Group	N.	No. of Words	Mean of Pre-Tests	Mean of Delayed Post-Test	Max. Diff. Pre/Post	Mini. Diff. Pre/ Post	Mean Difference 2 (Delayed Minus Pre)	SD
Control	10	20	74.30	76.30	90/ 84	30/47	2.00	10.0
Experimental	15		63.80	74.47	91/ 93	52/53	10.67	9.4

Table 33 describes mean difference 2 (the delayed post-test minus the pre-tests). There were 10 participants in the control group and 14 in the experimental group. Excluding students who dropped out or changed majors, the researcher needed to confirm that the participants took all four pre-VKS tests and the delayed post-test. This data mining decreased the number of participants to the lowest level in the research. Since the number of words tested in the

delayed post-test was 20, the researcher also checked the scores for those commonly known words and directly deleted them for equal comparison. As is shown in Table 33, the maximum difference of the VKS score for the pre-tests and the delayed post-test was about 90, while the minimum difference score was 30 in the pre-test 47 in the posttest for the control group. As for the minimum difference for the experimental group is larger than 50. In addition, mean difference 2 was two points for the control group but more than 10 points for the experimental group.

Table 34. Independent Sample T-test for the Difference of Mean Difference 2 in VKS between the Control Group and the Experimental Group for Vocabulary Retention in the Delayed Post-Test

Group	t	df	Sig. (2-tailed)	Difference of Mean Difference 2	Max.	Mini.	SD	Cohen's D
Control	-2.202	23	0.038*	8.67	22	-11	10.01	0.90
Experimental					33	-6	9.39	

Note: *p < .05. Mean Difference 2 is the VKS score difference between the delayed post-test minus the pre-tests

Table 34 displays the independent sample t-test for the differences of mean difference 2 in the VKS test for the delayed post-test. As is shown in Table 34, the maximum mean difference 2 for the control group was 22, meaning that this student might have retained more than four

words (the maximum score for a word in VKS is 5), while the best student in the experimental group retained more than six words. However, the minimum scores were presented negatively, implying that students in both groups forgot the target words. After t-testing mean difference 2, the findings showed that those in the experimental group outperform those in the control group by more than eight points with a p-value of .038 (two-tailed). In addition, the effect size is also a large one (Cohen's d =0.90). Thus, incorporating the DIY corpus improved the participants' vocabulary retention according to the t-test results compared to mean difference 2.

5.2.1 Qualitative Data Regarding Benefits of the DIY Corpus for Vocabulary Knowledge

5.2.1.1 DIY Corpus Materials Facilitate Vocabulary Learning

The interviewees stated that using the DIY corpus tool could facilitate their English vocabulary learning. Interviewees were asked whether these concordance lines could help them learn the target vocabulary. They all agreed that it played a crucial role:

> Just like that kind of individual words, not a dozen words are listed, and then what is impressive is that they can have different meanings but the same word as the words they have known before. For example, *read* means reading, but it has different meanings, refreshing my impression. (Jack)

> The [different concordance] lines could help learn more about the meanings and word

usages. (Jack)

What influenced most to me most was that the bolded target words could attract my attention much easier. (Jerry)

5.2.1.2 Vocabulary Knowledge was Reinforced in the Writing Process

Second, the importance of vocabulary was reiterated by the possibility of testing the target words in the final exam. Using target words in the writing process left a more profound impression:

Maybe I didn't know it in the questionnaire [VKS] for the first time, and then I was very impressed after the explanation. (David)

In addition, our final exam composition was a letter of complaint, and these words were more important [in completing this composition]. It was possible to use these two words [expect, serve] deeply when writing a composition. (Jerry)

I remember you said these two words especially, and I just wrote them after you finished. (Alan)

It would deepen my impression if I could use the target word in my writing. Especially

if it is still misused, it would become more impressive after I modify it. (David)

When I wrote a composition in class, my desk mate would laugh at me if he found my clauses were uttered with mistakes. I would immediately reflect on the wrong use, refer to the printouts and revise the sentence accordingly. The process would consolidate the impression of the target words. (Jerry)

5.2.2 Summary

The quantitative data analysis was conducted using descriptive and inferential statistics in SPSS. Students from the two groups improved their target vocabulary knowledge after being required to use it in their writing. Not until the fourth VKS test did the experimental group participants significantly outperform their counterparts, as was found by comparing mean difference 1 (different scores between the immediate post-tests and the pre-tests). Moreover, in the later delayed post-tests, the experimental group participants also outperformed their peers in the control group, as was found after comparing mean difference 2 (different scores between the delayed post-test and the pre-tests). In conclusion, the participants moderately improved their target vocabulary knowledge after they put these words into writing practice, and a higher rate of vocabulary retention was shown by the students who used the DIY corpus.

5.3 Students' Perspectives of Learner Autonomy

Research Question Three: To what extent does the use of a corpus facilitate participants' learner autonomy?

The researcher adopted a questionnaire for the quantitative inquiry and a semi-structured focus group interview for the qualitative inquiry. Spratt et al. (2002) developed the questionnaire. It was adapted to examine changes in learner autonomy before and after the intervention because of its relevance and efficiency. The researchers defined the concept of autonomy as comprising student perceptions of their teachers, their responsibilities for various aspects of their English learning and their views of their motivation and the frequency of their engagement in out-of-class learning activities. The questionnaire was administered in Hong Kong to 508 participants with similar majors (such as mechanical and electrical engineering); participants were given 20 minutes to respond. Chi-square tests were also conducted by Spratt et al. (2002) to determine the relationships between the different frequencies regarding the subscales, and the P value was lower than 0.01. After adaptation, the questionnaire was translated into Mandarin Chinese and piloted among students from the same major who were not part of the research. The Cronbach's alpha score was over .85, N = 84. As for the results of the questionnaire, the researcher first calculated the reliability of the questionnaire as implied by Cronbach's alpha; this was followed by descriptive data. Finally, the differences and changes between the two groups were estimated using a paired-sample test and an independent sample t-test, respectively.

5.3.1 Results of the Questionnaire

The researcher first conducted a reliability test in the form of Cronbach's alpha; the questionnaire results were then described. Finally, the self-report differences were compared through independent sample tests using SPSS. Since the researcher probed into participants' understanding of their responsibilities for English learning and their frequency of English learning activities outside the classroom, mean differences were calculated and compared.

Table 35. The Reliability Results of the Subscales for the Pre-Test and Post-Test (N = 78)

Part	Subscales	N. of items	Cronbach's alpha	
			Pre-test	Post-test
A	Perception of responsibility towards English learning	10	0.923	0.936
B	The frequency of English learning activities outside the classroom	12	0.853	0.901

Table 35 includes the subscales for both Part A and Part B of the questionnaire, N = 78, with 38 in the control group and 40 in the experimental group. The Cronbach's alpha values were all larger than .85, indicating a high level of reliability. Thus, the results may plausibly further the analysis.

Table 36. Descriptive Data of the Results of Students' Perceived Learner Autonomy Regarding Students' Perception of Responsibility towards English Learning and the

Frequency of English Learning Activities Outside the Classroom

Dimensions	Group (Control N = 38; Experimental N = 40)	Sample Survey Items, 5-Point Likert Scale Questions	Mean (SD)		Mean Diff. (Post and Pre)
			Pre-test	Post-test	
Perception of responsibilities towards English learning	Control	To check how much progress you have made. To stimulate your interest in learning English. (1 = not at all; 5 = totally)	3.67 (0.68)	3.46 (0.62)	-0.21
	Experimental		3.24 (0.37)	3.48 (0.78)	0.21
The frequency of English learning activities outside the classroom	Control	I have listened to English songs. I have practised speaking English with my friends (1 = never; 5 = always)	2.66 (0.59)	2.96 (0.63)	0.30
	Experimental		2.69 (0.52)	3.02 (0.65)	0.33

The five-point Likert scale questionnaire contained two parts: one focussed on students' perception of their responsibilities towards English learning and one focussed on the

frequency of English learning activities outside the classroom. In Part A, the participants were asked to report on the following question: 'When you are taking English classes, whose responsibility should it be?' like checking how much progress you have made, stimulating your interest in English learning, etc. The answers to this part range from 'not at all' to 'totally' their own. In Part B, they were required to answer the following question: 'Outside the classroom, how often have you conducted an English learning activity?' like listening to English songs, practicing English with friends, etc. The answers to Part B vary from 'never' to 'always'. Table 36 shows that participants in the control group had decreased perceptions of responsibility towards English learning, but those in the experimental group had increased perceptions of responsibility. The frequency of English learning activities increased by comparing the pre-test and the post-test.

Table 37. The Independent Sample T-Test comparing the Control and Experimental Groups Regarding Students' Perceived Learner Autonomy in terms of Students' Perceptions of Responsibility towards English Learning

Group	N	Mean Diff. Part A	Max.	Mini.	SD	T-test (2-tailed)	df	t	Cohen's d
Control	38	-0.216	3.0	-2.0	0.997	0.030*	76	-2.2	0.50
Experimental	40	0.245	1.9	-2.1	0.842				

*$P < 0.05$

Table 37 shows that the participants in the experimental group increased their perceptions of

their responsibilities towards English learning, while the control group decreased these perceptions. The mean difference for the control group in Part A was -0.216, yet the value for the experimental group was 0.245. Although the maximum mean difference scores were positive, the higher point was lower in the experimental group. Furthermore, the minimum points for the mean difference were around -2. The researcher came to the conclusion that the experimental group significantly outperformed the control group in terms of their perception of responsibilities for English learning after confirmation by the independent sample t-test, with a p-value of .030 (two-tailed). The effect size is medium (Cohen's d= 0.50).

Table 38. The Independent Sample T-Test comparing the Control and Experimental Groups Regarding Students' Perceived Learner Autonomy and Students' Perceptions Regarding the Frequency of English Learning Activities Outside the Classroom

Group	N	Mean Diff. Part B	Max.	Mini.	SD	T-test (2-tailed)	df	t
Control	38	0.300	3.00	-0.92	0.97	0.900	76	-0.125
Experimental	40	0.327	2.83	-1.92	0.91			

Different from Table 37, Table 38 reports no significant differences between the groups regarding the frequency of English learning activities outside the classroom, with a p-value of .90 (two-tailed), which is higher than .05. Both groups of students increased the frequency of their English learning activities outside the classroom, with the mean difference improved by .30. Although the maximum mean difference was about three points, the minimum score

was markedly different. The minimum for the control group was -0.92, while the score for the experimental group was -1.92. Thus, both groups increased the frequency of their English learning activities outside the classroom. However, some participants in the experimental group did increase their frequency, considering that the highest score for each item was five.

5.3.2 Relevant Interview Data

Thematic coding revealed the relevant interview data. The experimental condition helped participants facilitate an awareness of learner autonomy, and they began to understand the importance of it. This psychological change is attributable to their voluntary reference to the DIY corpus materials.

5.3.2.1 The Importance of Learner Autonomy

The interviewees offered to reflect on the autonomous learning experience in the higher vocational institute compared with their learning in secondary school. These interviewees hit upon the importance of learner autonomy.

> High school teachers are closely [watching over] [our English learning], and high school students will be detained if we cannot recite the required articles. (Jerry)

> Senior high school teachers will arrange [our English learning]. Senior high school students will learn all kinds of knowledge only through reciting words and doing drills.

If you want to study at a university, choose the materials according to your reality. (David)

[We should] have plans and goals, including going to the library. (Jerry)

Otherwise, you cannot learn anything. (David)

That method [incorporating DIY corpus] can only have a limited effect. If you want to learn English well, you can learn it well whether the teacher is here or not. (Alan)

Learning English can promote my sense of achievement and my peers may envy me. (Jerry)

[Even though] the high school teacher is extremely strict, if you do not want to learn [English], you cannot understand English well. (Alan)

5.3.2.2 Voluntary Reference to the DIY Corpus

Students in the experimental group could voluntarily refer to the DIY corpus printouts, and this type of activity aroused students' reflection on the experience of autonomous learning.

Sometimes, when I look at the papers [DIY corpus printouts], I will go to find the

> preposition behind the verb when I write a composition, just like a fixed sentence pattern. When I write a composition, I will see how to match it because it is often useful when writing. (David)

Interviewees in the experimental group had increased their English learning activities outside the classroom and cultivated an awareness of learning outside the classroom.

> It's mainly English songs. If you are interested, you will see the meaning of Chinese and its sentences. What kind of sentences will be smooth? The lyrics should be rhymed, fluent and free from language defects. (Alan)

> I prefer TikTok. I will pay more attention to English songs when I see them. And when I see some [advertisement] slogans, I will also pay attention. (Jerry)

5.3.3 Summary

The participants in the present research improved their perceptions of learner autonomy in terms of their responsibilities towards English learning and their frequency of English learning activities outside the classroom. The experimental group statistically outperformed the control group in terms of perceived responsibilities towards English learning. In contrast, there was no significant difference in the frequency of English learning activities outside the classroom between groups, and some students in the experimental group even decreased their

activities more than the control group.

5.4 Students' Perceptions towards Using DIY Corpus Data in English Writing

Research Question Four: What are the participants' attitudes and perceptions towards using corpus data in English writing?

A questionnaire designed to measure participants' attitudes and perceptions towards using corpus data in English writing was administered to the participants in the experimental group at the end of this study. Forty-six students filled out the questionnaire, but three of them completed it in less than 60 seconds, much lower than the average time spent on the questionnaire, 152 seconds. Therefore, there were 43 valid responses. The researcher next used Cronbach's alpha to calculate the questionnaire's reliability, presented the results, and ultimately made a connection between the quantitative and qualitative data. The information was then summarised.

5.4.1 Results from the Questionnaire about Students' Reactions to Using the Printouts from the DIY Corpus

Table 39. The Reliability Results of the Subscales (N = 43)

Part	Subscale	N of items	Cronbach's alpha
A	English Writing Proficiency Improvement	7	0.967

B	Reactions to the DIY Corpus Printouts	9	0.945
C	Outside Assignment Usage	9	0.963

Table 39 shows the three parts of the questionnaire pertaining to participants' reactions to using the printouts from the DIY corpus. The questionnaire used a five-point Likert scale, from 'strongly disagree' to 'strongly agree'. There were seven items in Part A concerning participants' self-reported focus on their English writing proficiency improvement, Part B was about using the DIY corpus printouts and Part C was about participants' use of the corpus outside their assignments. As shown in Table 39, the Cronbach's alpha values were all higher than .85, indicating a high level of internal consistency.

Table 40. Results in the Questionnaire Using a Likert Scale for the Experimental Group (N = 43)

Dimensions	No. of Items	Sample Survey Items 5-point Likert Scale Questions (1 = strongly disagree; 5 = strongly agree)	Mean (Max./ Mini.)	Std.
English Writing Proficiency Improvement	7	e.g., I would feel confident writing in English. Using the corpus printouts would be helpful for dealing with preposition usage in my writing.	3.73 (5.00/1.43)	0.89
Reactions to	9	e.g., The DIY corpus printout search	3.58	0.77

the DIY Corpus Printouts		technique would be easy to learn. I would have difficulty using the DIY corpus printouts due to the time and effort spent on concordance lines.	(5.00/1.56)	
Outside Assignment Usage	9	e.g., The DIY corpus would be more useful for writing than reading. I would use the DIY corpus for my English writing in the future.	3.57 (5.00/1.44)	0.82

Table 40 shows that all three parts of the questionnaire received generally positive feedback. The mean score for the things on the list was above 3.5, which is close to 'agree'. The participants who used the DIY corpus during their writing tasks agreed that their English writing proficiency improved, the DIY corpus was easy to use and they even used the materials after the assignments were finished. In contrast, some participants felt differently, as evidenced by minimum scores of 1.4, 1.6 and 1.4, somewhere between 'strongly disagree' and 'disagree'. Thus, the DIY corpus was still challenging to use for these students or did not affect their English learning or writing proficiency.

5.4.2 Benefits of DIY Corpus Use

The data from the questionnaire were combined with the interview data. After three rounds of open, axial and thematic coding, themes were generated and divided into the benefits and

difficulties of using the corpus. Afterwards, the themes were elaborated with excerpts and items from the questionnaire.

As shown in Table 40, all 25 items averagely fell on the scale close to agreement, with three representing 'undecided', four representing 'agree' and five representing 'strongly agree'. Additionally, more than 70% of participants agreed or strongly agreed with items 2, 12, 16, 20 and 24.

Table 41. Items from the Questionnaire with which Most Students Agreed or Strongly Agreed

Item	**Statements**
2	It would be helpful to write essays by using the DIY corpus.
12	The DIY corpus output would provide enough information to determine the usage of the vocabulary.
16	The DIY corpus feedback would give me useful references when I write and revise essays.
20	The more I use the DIY corpus, the more I like it.
24	I would recommend using the DIY corpus for writing to my friends.

In other words, for participants in the experimental group, the DIY corpus helped them determine vocabulary usage and provided them with valuable references when they wrote and revised their essays. They came to like the DIY corpus tool and would recommend it to their friends. More than 80% agreed or strongly agreed that using the DIY corpus would help with selecting the correct vocabulary between synonyms, dealing with preposition use and

improving word choice in their writing. According to David,

> sometimes when looking at those papers [DIY corpus printouts], and when writing essays, I will look carefully for a preposition after a verb, a fixed sentence structure, for instance. When writing an essay, I will look at how to match writing with it [concordance line] because of its usefulness.

In general, writing with the DIY corpus proved to increase English writing proficiency and be a useful resource for improving English proficiency.

In the interviews that followed, participants expressed that the concordance lines gave them authentic examples to imitate, as stated by Alan, who expressed that 'when they discovered that they need to use the target word, they would look it up in the concordance lines and directly follow the example to use it in their essays'. In addition, after they read through the lines, they would have more thoughts about the essays when they brainstormed for the outline of the writing task. These authentic instances from the DIY corpus attracted the attention of the participants, and they reported wanting to refer to them even after the class. The concordance lines can help students 'learn more about words and usage from multiple perspectives to understand more meanings' (Jack) of the target vocabulary. Moreover, 'the vocabulary could leave [a] deeper impression if I could use the new word in the composition', said David.

5.4.2.1 DIY Corpus Printouts Are More Comprehensible

The DIY corpus printouts were more comprehensible to the subjects because of their content and the format of the presentation. The DIY corpus comprised Gaokao materials and College English Test band 4, indicating the high school level and the tertiary level. The participants felt it was easier to understand than large general corpora (Alan called it 'slightly better', indicating that he had less trouble understanding the concordance lines). However, these difficulties did not affect their 'imitation' or 'reference' or 'writing according to that [concordance line]'.

AntConc software was easy for the researcher to adopt, but it proved difficult for the participants during hands-on use. Because the buttons and menus of the software are in English, the participants were challenged. David said, 'It was after several times of wrong operation that I know how to use it and begin to understand the meaning of each option'. In contrast with the hands-on challenge, the printouts were given out to students after the researcher intentionally selected those concordance lines that could adequately display the contexts of target words and carefully edit the format of the concordance lines to the participants' interests. Jack expressed that he would 'read it every time', and three of the interviewees agreed on this point. In addition, David said he would also read through the printouts sometimes when reviewing the course.

5.4.2.2 Concordance Lines Provide Useful Examples

The authentic materials derived from participants' interests and needs, and the concordance

lines can provide examples before, during and after writing practices.

The participants felt challenged and reluctant to write essays due to their low motivation, low proficiency and limited vocabulary. Without the desired vocabulary, Alan said, ‘even though I could come up with some ideas to write the composition, I cannot write out what I want to express because of my small vocabulary size’. Similarly, Jerry said he could write one sentence after thinking of it for a long time. However, the participants were able to skim the printouts with target vocabulary embedded in the concordance lines, and Jerry could even write according to the concordance lines.

In the process of writing, the concordance lines helped the participants with correct collocations and other sentence structures. Some of the concordance lines became handy as soon as the students decided to put them into their compositions, and in any case the collocations and sentence patterns were also of great value to them. As David said,

> Sometimes when I read through the printouts, I will pay more attention to the prepositions after the verbs. Since the target vocabulary will be incorporated into the compositions, I will take a closer look at the patterns presented by the example.

After the drafting process, some participants found erroneous sentences, and they would also refer to the concordance lines to be sure, as reported by David. Browsing the concordance printouts helped students derive correct usages, especially the collocations of target verbs,

and students were able to derive possible understandings of sentence patterns or collocations. After they put their understanding into writing practice, they felt they may still be wrong, thus the participants turned to the printouts again for confirmation.

5.4.3 Difficulties

The interviewees mainly expressed that their English proficiency affected their English learning and their DIY corpus use. The inaccessibility of some of the corpus websites or tools hindered the popularity of the corpus, and the interviewees offered some valuable suggestions for similar applications of the DIY corpus in English learning at the higher vocational institute.

Facing the harsh fact that he was a low-proficiency English learner, Alan chose to change his language course for the College Entrance Exam. According to Alan, 'I scored 40 or 50 in my English course of high school [out of a total of 150], and I was obliged to transfer to learning Japanese'. These low-proficiency English learners were interested in English learning, but they became discouraged by their academic performance, as reported by Jerry: 'For example, my high school English was very frustrating. I spent a lot of time studying science and mathematics, but I didn't like Chinese and English. I scored over 40 in English for the College Entrance Examination'.

The hardware and time needed to browse the corpus materials both hindered the popularity of

the corpus. Most participants did not have laptops or desktop computers. Even those who had computers were not willing to carry them for English writing or corpus use because ‘After we finish our English class on Friday, we have to walk to Zhi Xing Building 5 for class. Actually, bringing a computer is very heavy, which can be quite troublesome’, according to the interviewee Jerry. Full schedules of daily courses left students with limited time to cover the long distance of walking from one building to another on a big campus, which is often the case in China for students in higher vocational institutes. Furthermore, the software was not very friendly to lower-proficiency English learners. This research piloted hands-on experiments before finally turning to printouts. The software was posted online on an overseas source, which sometimes required a virtual private network to browse or exploit the software. Finally, the English interface created trouble and misunderstandings when low-proficiency English learners tried to register, not to mention use, the software.

The interviewees from the experimental group did offer some insightful suggestions for the future application of corpus-related materials in English classrooms.

5.4.4 Suggestions from Interviewees to Improve the Effectiveness of Using the DIY Corpus

The interviewees mainly provided their suggestions from the aspects of the format of the DIY corpus for better comprehension of the materials, stating that the content should include at least some L1 translations for target vocabulary. They also emphasised the need for more guidance or support from teachers and in regard to using the corpus and writing.

Table 42. Interviewees' Critical Suggestions

Interviewee	Key Suggestions	Aspect
Alan	The number of concordance lines should be between four and eight for a word.	Format of printouts
Jerry	Larger font size, bigger line spacing.	
David	Prefer more L1 translation/directly from CET 4 or 6.	Content
All	Prefer more reliance on teacher's explanation and supervision.	Method

As shown in Table 42, the participants first commented on the format of the DIY corpus printouts. They also preferred more Chinese translations and more teacher explanations and supervision during their learning. At the end of the interview, the interviewees were invited to offer suggestions for revising the printouts of the concordance lines. It was suggested that the number of concordance lines should be between four and eight, as suggested by Alan, and the interviewee Jerry stated that 'The characters should be a little larger, with larger line spacing, and then a translation should be provided below. It would be better to provide an understandable example sentence'. Jerry added, 'I preferred to translate the meaning directly. A word, then a collocation, and then a translation. The best words have an example sentence below the target words'.

As mentioned previously, some interviewees still preferred translation to concordance lines:

> The composition is more thoughtful, and then I find my vocabulary is small, and I can't write it. (Jack)

> It could write only one sentence after a long time of thinking. (Jerry)

> I can draft in Chinese, but I can't translate it into English. I always feel wrong here or there when my ideas are written in Chinese and translated into English. (David)

By the end of the interview, the participants had provided several opinions about incorporating the corpus materials. One of them would still turn to recite the word lists directly coming from the College English Tests. As David said, 'I disagreed that the [concordance] lines greatly affected vocabulary learning. I believed that it was more direct to recite words from the College English Test band 4 and 6'.

Alan and Jerry even commented on the process of incorporating DIY Corpus materials, saying that they would prefer more teacher-centred teaching, shown by their need for the teacher's explanations and scaffolding of the vocabulary knowledge: 'It will be more effective if keywords are distributed and explained again. Let's find it ourselves. It's unlikely to be effective. You can be more confident if you don't know how to translate'.

5.4.4 Summary

The participants from the experimental group self-reported their attitudes towards using the DIY corpus materials in the writing tasks through a questionnaire and an interview. The questionnaire yielded generally positive feedback on participants' feelings about improving their writing proficiency and using the DIY corpus printouts, both for and after assignments. Besides this, they also provided some practical and insightful advice for future applications of the DIY corpus. First, they preferred fewer concordance lines for particular words, with bigger, more prominent fonts and better line spacing. Second, they reported preferring more Chinese translations for the target words and key concordance lines. Third, they still relied greatly on the teacher's explanation of the target words and supervision of their English learning.

5.5 Summary of the Results

The quantitative results show that the DIY corpus helped students' vocabulary learning and composition writing to some extent. As shown by the writing quality in the quantitative data, the participants received significantly higher grades and displayed higher frequencies of correct target verb collocations in the fifth writing task. The fifth task was conducted under exam conditions, which was closer to the fact of students' real writing quality than the first four writing tasks done in the classroom. In terms of vocabulary knowledge, mean difference 1 (comparing the immediate post-test with the corresponding pre-test) and mean difference 2 (comparing the delayed post-test with the pre-tests) were calculated. The significant

difference appeared in the fourth VKS test for mean difference 1, and the participants in the experimental group also outperformed their peers in the retention of target vocabulary knowledge according to mean difference 2. Jack expressed that certain target words in the concordance lines could display new meanings, and these new impressions could help build links to what was already known to the students, thus leaving a more profound impression. In addition to the importance of the target words that could appear in the exam, the teacher's explanation of the target words could help enhance students' awareness of the words. Then, after students used these words in the writing tasks, they could develop a better understanding of the target words. As for the benefits for composition writing, some concordance lines provided a reference for students to imitate in their writing, according to Jerry. The correct usage of the words in the KWIC style helped students to revise and proofread their drafts at different stages of completing the writing tasks. If the improper use of a target word was detected by the students or their classmates, they directly corrected their use accordingly. This process of trial and error stressed the meanings of the target words, as stated by Jerry and David.

The process of exploring the DIY corpus printouts helped students become more autonomous. As shown by the quantitative data, the participants in the experimental group outperformed their counterparts in the control group in terms of responsibilities towards English learning, yet they equally increased the frequency of their English learning activities outside the classroom. After several months of the DIY corpus experiment, students like David thought back to their years of high school learning, drafting their writing tasks in class.

The teachers closely monitored these high school years, while the higher education system depends more on the students. The interviewees began to form the idea of learning motivation and learner autonomy, as indicated by Alan: ‘Even if the high school teacher was very strict, you could not learn if you do not want to learn’. In addition, students like Alan and Jerry began to pay more attention to the English slogans in advertisements and the lyrics of English songs instead of the rhythm, especially the sentence structures, rhymes and slang.

The experimental methods received generally positive feedback from the participants in the experimental group. In the questionnaire items regarding students’ reactions to incorporating DIY corpus materials into their writing, some of the items were agreed or strongly agreed upon by more than 80% of the students regarding vocabulary learning and use, such as selecting the correct vocabulary between synonyms, dealing with preposition usage and improving word choice in their writing. Most importantly, the students felt more confident in improving their English proficiency.

Apart from the positive feedback, the interviewees also provided several insightful and practical suggestions from many aspects of the process of incorporating the DIY corpus materials. First, some of the participants in the experimental group still preferred their habit of translation and even mentioned providing more translation in the concordance lines. Before writing, some interviewees would first write a Chinese version of the composition in their heads and then attempt to translate it into English. Additionally, some mobile translation applications were mentioned in the interview. Second, students offered several format

suggestions for the printouts. They would prefer fewer concordance lines for each target word, ranging from four to eight. The line spacing and font size should be made more prominent. They also preferred Chinese translations. Third, the content of the concordance lines was advised to be directly derived from the English exams, such as CET 4 and CET 6; students would also prefer that the teacher spend more time explaining the concordance lines since some of them were still difficult to understand.

Chapter 6: Discussion

The current study's findings reveal the effects of incorporating the DIY corpus into the English classrooms of higher vocational institutes, where corpus-based research is seldom targeted. A mixed-method research design was adopted in this study. Pre- and post-tests investigated the writing quality, verb collocation use and target vocabulary knowledge displayed in assigned writing tasks. In addition, the quantitative research design, aided by questionnaires and a qualitative, semi-structured focus group interview, reveal the participants' perspectives of learner autonomy regarding their responsibilities towards English learning, the frequency of their English learning activities outside the classroom and their reactions to using the DIY corpus in their English writings. The following chapter focusses on a discussion and interpretation of the results. First, the chapter summarises the essential findings before discussing them in relation to the literature in terms of writing quality and the accuracy of verb collocation, vocabulary knowledge, learner autonomy in terms of perceived responsibilities towards English learning and the frequency of outside English learning activities, the feasibility of the DIY corpus for low-proficiency English learners and corpus-based language pedagogy for lower achievers in English. Finally, a summary of the discussion and the significance of the study are presented.

6.1 Key Findings

The first research question was about writing quality and target verb collocations in writing with the incorporation of the DIY corpus. Students in the experimental condition improved

their writing quality by increasing their correct target verb collocations after incorporating the DIY corpus in their writing practices. In the fifth writing activity, participants in the experimental group significantly outperformed those in the control group, which was completed under exam conditions with teacher supervision. In addition to the writing quality, the frequency of accurate target verb collocation was significantly different in the fourth and fifth writing tasks. Moreover, writing quality was negatively correlated with the number of erroneous verb collocations.

The second research question asked whether participants improved the target vocabulary knowledge embedded in their writing tasks. The research design made use of a pre-test, an immediate post-test and a delayed post-test in the form of a VKS, and the mean differences between the immediate post-test and the pre-test and between the delayed post-test and the pre-tests were later calculated. Based on the mean difference of the former comparison, participants in both groups improved their vocabulary knowledge. A t-test was conducted to compare the immediate post-test and the pre-test. It was found that the experimental group students significantly outperformed their counterparts in the fourth VKS test. Analyses of the mean difference between the delayed post-test and the pre-test revealed that both groups of participants retained target vocabulary knowledge. However, the experimental group significantly outperformed the control group in terms of vocabulary retention.

The third research question was designed to explore the effect of DIY corpus use on participants' learner autonomy regarding their responsibilities towards English learning and

the frequency of English learning activities outside the classroom. A highly reliable questionnaire adapted from Spratt et al. (2002) and a semi-structured focus group interview were employed to investigate the research question. The questionnaire was administered to both groups before the experiments as a pre-test and after the research as a post-test. In further analysis, the mean difference between the pre-and the post-test was calculated. The results show that the participants in the control group perceived lower responsibilities towards English learning, while the experimental group showed improvement in their perceptions of their responsibilities towards English learning. In addition, both groups increased the frequency of their English learning activities outside the classroom. In a later independent sample t-test of the mean differences, it was revealed that a significant difference existed in regards to the perception of responsibilities, with the experimental group showing a significantly higher rating than the control group. However, no significant difference was found in the frequency of students' English learning outside the classroom.

The last research question focussed on the participants' attitudes and perceptions towards using corpus data in English writing. A questionnaire adapted from Nam (2010) was used to collect these data, and the experimental methods received generally good feedback from the participants in the experimental group. The coding for the qualitative data from the interview and regarding the benefits of incorporating DIY corpus materials revealed that the DIY corpus was perceived to be more understandable than other corpora, and the concordance lines provided valuable examples for the writing process. The interviewees also expressed their concerns about their low English proficiency and the limited hardware availability that

obstructed popular access to the corpus. Finally, they also provided suggestions on the format of the DIY corpus printouts and approaches to incorporating the corpus.

6.2 Writing Quality and Verb Collocation

The present research is aligned with the research of Luo and Liao (2015), who found that students' writing quality can be improved across different dimensions after a corpus-based intervention. Gilmore (2009) found that Japanese university students made 61.14% of their total word count in lexical and grammatical problems between their first and second drafts, making their writing sound more natural. Additionally, corpora utilised as reference materials are more beneficial than online dictionaries for assisting undergraduate students without an English major in making precise edits and minimizing errors during free output (Luo & Liao, 2015; Luo, 2016). The present research was designed to incorporate DIY corpus printouts into students' writing practices. The writing quality in this study's writing tasks was evaluated in the form of a grading scheme that considered elements such as content, grammar, vocabulary and structures. The scheme was comprehensive and not only focussed on specific aspects. The participants displayed no significant differences in their first four writing tasks due to their use of mobile applications. Still, the participants in the experimental group significantly outperformed the control group in the fifth writing task. This task was completed under exam conditions, and the participants had to rely on themselves to complete their writing. In previous empirical research, students have usually been asked to write about only one topic (Gilmore, 2009; Luo, 2016). By contrast, this research adopted five different

topics, following Huang (2014) and Luo and Liao (2015). Huang (2014) focussed on using abstract nouns in the writing of 40 students in their third year of college studying in English for business purposes. Luo and Liao (2015) compared the BFSU CQP (Beijing Foreign Studies University Corpus Query Processor) and an online dictionary among 30 undergraduate students. The participants in these studies were relatively high-proficiency university students. The current research was conducted at a higher vocational institute, and the participants were lower-proficiency English students. These participants were habitual mobile phone users and heavily depended on translation applications. Considering their low motivation for English learning and their histories of failed English learning experiences, the findings of this study show that the DIY corpus could be a valuable resource to help low-achieving vocational institute English learners improve their English writing.

As demonstrated by the use of noun-verb collocations by native Chinese speakers (Chan & Liou, 2005), verb-adverb collocations by native Macedonian speakers (Daskalovska, 2015), and a variety of collocations by native Arabic speakers (Cobb, 1997), DDL is a successful method for teaching and learning L2 collocations. Collocation as part of lexical-grammatical knowledge is a crucial part of writing. Yılmaz's (2017) experimental group, as compared with the control group, used a wider range of collocational and colligational patterns and made less grammatical mistakes while utilising abstract nouns. Incorporating a corpus could enable learners at all three proficiency levels produce considerably fewer collocations than native speakers, and the amount of collocations only grew at the advanced level, despite the participants' varied competence levels (Laufer & Waldman, 2011). In the present research,

the participants were lower achievers in English. Correct target verb collocation frequency significantly differed in the compositions of the experimental and control groups in later writing tasks. The experimental group participants made even more target verb errors in the first three writing tasks. However, the experimental group students made significantly more correct target verb collocations in the fourth and fifth tasks.

More importantly, verb-noun collocation errors peaked among writing errors for Chinese EFL learners at around 30% (Wang & Zhou, 2020) or 57.14% of the total collocation errors (Zou, 2019). The present research discovered that verb collocations had more to do with writing quality. The frequency of erroneous verb collocations was graded, and the Pearson correlation coefficient test was performed, resulting in a negative correlation with students' corresponding grades. In other words, the more erroneous verb collocations there are, the lower the writing grade will be.

6.3 Vocabulary Knowledge

Vocabulary knowledge is fundamental to language learning, involving different aspects of knowledge (Richards, 1976; Wesche & Paribakht, 1996; Schmitt & Meara, 1997; Nation, 1990, 2022; Webb, 2005; Qian, 2008; Van & Schmitt, 2013). Language learners typically think their struggles with receptive and productive language usage are primarily due to a lack of vocabulary (Nakata, 2008), and vocabulary knowledge is highly important when there is no enough linguistic input (Nation, 1990). However, the truth is that a low vocabulary can be

improved through input modes (Vidal, 2011). Unlike previous input-based vocabulary learning, this study explored vocabulary learning by writing. The vocabulary knowledge improvement was affected by various factors.

It is essential to realise that the output hypothesis (Swain, 1995) does not minimise the importance of input in any manner. Instead of substituting input-based language learning methods, the objective is for students to go beyond what is minimally required to comprehend the overall message (Swain & Lapkin, 1995; Izumi & Bigelow, 2000). The concordance lines provided input, and the writing tasks were output, triggering the participants to pay attention to the collocations of the target words in KWIC, followed by the concordance lines. In addition, when the allotted time for tasks depends on the amount of time needed for completion, with writing tasks requiring more time, a writing task is more practical; moreover, productive vocabulary learning tasks are more effective than receptive tasks (Webb, 2005). Writing in the present research was conducted during a two-week cycle. The participants were given a VKS test for the target verbs, then they were guided to study the DIY corpus printouts and refer to vocabulary consultation tools. Finally, they were required to complete the writing tasks using target verbs. According to the research schedule, the drafting and revising took at least 35 minutes. In addition, students were instructed to use the target vocabulary in their writing. In this sense, essay writing can be beneficial for vocabulary learning. The two mean differences were calculated to probe into the effect of writing on vocabulary learning. The first mean difference came from the pre-test and the immediate post-test; the second was between the pre-tests and the delayed post-test.

Regarding the first mean difference (the pre-tests and the immediate post-tests), the scores on the immediate post-tests were higher than those on the pre-tests. However, the participants in the control group made less progress than their counterparts in the experimental group. Moreover, there was a significant difference between the groups in the fourth VKS test, and the participants in the experimental group significantly outperformed their peers in the control group. Concerning the second mean difference (the pre-tests and the delayed post-tests), it was found that the participants in the experimental group retained significantly more vocabulary knowledge than their counterparts in the control group. In other words, the students in the experimental group could utilise more target verbs in their final exam writing, which was in accordance with the significance of the writing quality results of the final exam.

However, the data analyses did not show a significant difference between the pre-tests and the immediate post-tests until the fourth test. On the one hand, the low achievers needed more time to become experienced in referring to the new corpus tool; on the other hand, the vocabulary acquisition was complicated.

Successful L2 vocabulary acquisition through reading depends on three things, say Peters et al. (2009). L2 learners should start by learning the definitions of new words. Second, they should elaborately process lexical information. For example, they should focus on the semantic function of the target words in context. Third, the meaning of repetition should reinforce the form-meaning connections of these words; however, students fail to notice the vocabulary, which is where learning starts. It was suggested by the interviewees that the DIY

corpus materials were printed too densely, and their low English proficiency discouraged them from reading more than eight concordance lines. The participants had more creative latitude with the essay-writing assignments, which provided more significant contexts for language creation but made it more challenging for them to explicitly compare their inter-language output with the model input. Participants may have paid attention to other aspects of that input as a result (Izumi & Bigelow, 2000).

If participants fail to discover the meanings of unfamiliar words, then elaborative processing, which is crucial for learning L2 vocabulary (Hu & Nassaji, 2016), will not occur. The third aspect is form-meaning connections (Vidal, 2011); prior to the study, the majority of students were unfamiliar with some of the target words. Instead of adding a new record to an already existing notion, they had to develop new meanings for these terms and tie them to the forms they came across in the text. However, for those words that the participants had some prior knowledge about before the study, the participants established the connections to the pre-existing knowledge, which helps with vocabulary retention in the long term. This can be seen from both the quantitative data and the qualitative data. In the quantitative data, the mean difference between the pre-test and the delayed post-test exhibited a significant difference between groups, indicating that the participants in the experimental group retained statistically significantly more vocabulary knowledge than those in the control group. According to David,

> Maybe I didn't know it in the questionnaire [VKS] for the first time, and then I was

> very impressed after the explanation [concordance lines]. It would deepen my impression if I could use the target word in my writing. Especially if it is still misused, it will become more impressive after I modify it.

Apart from the factors affecting vocabulary acquisition, the L2 learners' learning factors were also influential. L2 incidental vocabulary acquisition is significantly influenced by a learner's L2 competency, anxiety, and strategy mastery. L2 learners tend to pick up more words accidentally by reading if they have stronger L2 competence, better strategy mastery, and higher degrees of incidental vocabulary acquisition anxiety (Zhao et al., 2016). The personal information questionnaire conveyed the participants' low English proficiency, anxiety about passing the compulsory English course and the need for a certain English level for future career development. In addition, writing was perceived as one of the most challenging skills for the participants, and the corpus, as a new strategy, was also a considerable challenge.

6.4 Learner Autonomy

In the contemporary educational environment, which encourages lifelong learning and is facing an expanding need for distance learning, autonomy has emerged as an essential educational aim (Spratt et al., 2002). Learning autonomy is the capacity to take responsibilities of their own study that students have or exhibit in various situations (Benson, 2001). The participants in this study enhanced their understanding of learner autonomy in terms of their responsibilities for English learning and the frequency with which they engage

in extracurricular English learning activities. Regarding perceived obligations to English study, students in the experimental group significantly displayed more perceived obligations than those in the control group. However, the frequency of English learning activities outside of the classroom did not change significantly, and compared to the control group, certain students in the experimental group shown more significant progress.

First, the inductive nature of corpus-based language learning promotes learner autonomy. A corpus is a constructivist and inductive language learning method, favouring cognitive and metacognitive growth, critical thinking and observational abilities, linguistic awareness and sensitivity to texts, autonomy and lifelong learning (Cobb & Boulton, 2015). The experimental group participants had their own writing procedure pace and were free to consult the corpus materials. They were also independent explorers, able to read through the concordance lines and develop their understanding of meaning, collocation and other uses of the target words.

Second, the authenticity of the DIY corpus is its first advantage (e.g., Cobb & Boulton, 2015; Gilquin & Granger, 2010; McCarthy & O'Keeffe, 2010; Timmis, 2015), exposing students to thousands of actual instances of particular language elements. Learners can better understand interlanguage aspects and produce better language by comparing models produced by native speakers in corpora and language. The DIY corpus consists of CET 4 materials and Gaokao English materials. Though these were all exam oriented, they all came from the feedback from the pilot study. The presence of more than 240,000 words in the corpus data could

match participants' proficiency levels and their needs for future career development.

Third, 'DIY' is short for 'do it yourself', and the DIY corpus has many advantages that large general corpora do not. Teachers first choose which linguistic information goes into a corpus. There are several drawbacks to large general corpora; they are criticised for overwhelming students with data (e.g., Charles, 2012). Moreover, DIY corpora are resources that do not require an internet connection for access. Once constructed, they are always available and may be accessed wherever and whenever needed. In the present research, not all participants had access to computers; intentionally selected sections of the corpus were printed out in order to overcome the problem of concordance lines sometimes being decontextualised for a limited length of contexts (McCarthy & O'Keeffe, 2010). The participants in this research reported their voluntary reference to the DIY corpus concordance lines.

6.5 Feasibility of the DIY Corpus for Low Proficiency Students

Following the advent of learner corpora, which are digitised collections of authentic texts written by language learners, Seidlhofer (2002) coined the term 'learning-driven data', and learner corpora are becoming more popular in language teaching and learning (Cotos, 2014). Cotos (2014) explored the potential of local learner corpora and examined the use of a native-speaker corpus alone and when paired with a learner corpus. She found that adverb use was favourable for 31 overseas graduate L2 students with TOEFL scores ranging from 83 to 107. In another study, secondary-level Korean EFL learners observed and unlearned their 'over

generated be' (additional *be* occurring before thematic verbs) by comparing native English speaker and learner corpora with guided induction (Moon & Oh, 2017) on the cognitive and affective benefits of data-driven learning. Different from the concept of learner corpora, with data coming from learners' production, a DIY corpus is a 'small-scale database of electronic texts built by users for specific, limited and local purposes' (Charles, 2018). The terms 'local corpora', 'disposable corpora' and 'personal corpora' have also been used to describe them (Charles, 2018, 2019). In the present research, the corpus included Gaokao English materials and CET 4 materials. The data were chosen according to the participants' needs and interests and then converted into text format. Then, the plain texts were uploaded into software programmes like *AntConc* (Anthony, 2020) for further analysis. Charles (2012) pioneered the introduction of DIY corpora into EAP courses, and 90% of that study's 50 participants found it easy to build their corpora from 10 to 15 research articles. However, in the present research, participants were lower achievers, and not all had laptops. Thus, the researcher created the corpus and selected the concordance lines, printed them out and administered them to the participants for their use.

The questionnaire adapted from Nam (2010) explored the participants' attitudes and perceptions towards using corpus data in English writing in three dimensions, their English writing proficiency improvement, reactions to the DIY corpus printouts and use of the corpus outside of assignments. The results of this high-reliability questionnaire implied generally positive feedback. The average score for the items related to the English writing proficiency improvement, students' reactions to the DIY corpus printouts and their outside assignment

usage mentioned above was more significant than 3.5, indicating that participants agreed in most cases. It was clear that the participants who used the DIY corpus throughout their writing assignments concurred that their English writing ability had improved, that the DIY corpus was simple to use and that they had continued to utilise the resource even after their assignments had been completed. In the semi-structured focus group interview, the interviewees expressed that the concordance printouts in the DIY corpus were more understandable than the large general corpora in the pilot study, and they agreed that the corpus provided participants with practical examples to follow in every step of their writing.

The difficulties mentioned by the interviewees can be overcome. First, the corpus can be carefully selected and made more suitable for the participants' proficiency levels. Second, the need for laptops and internet access to make the corpus website available overseas can be dealt with by printing out the concordance lines. Ellis (1995) considers reading to be the 'ideal medium' for vocabulary acquisition; printing materials allows learners more time to process a new language input, 'whereas in speech it passes ephemerally' (p. 106). In addition, as suggested by the interviewees, larger font sizes, wider line spacing and some critical L1 translations may also facilitate comprehension of the concordance lines.

6.6 Corpus-Based Language Pedagogy for Low-Proficiency Students

The evidence from a meta-analysis (Cobb & Boulton, 2015) indicates that corpus work is now prepared to go beyond university ESP (English for Specific Purposes) classes, where it

has primarily been utilised up to this point, and into general second and foreign language learning. Of course, its benefits can still be further explored, and the factors contributing to its effectiveness can be further explained. Moreover, most of the professional teaching community has remained largely unaware of the corpus-based linguistic approach for various reasons: the absence of corpus learning in teacher training, teachers' associations of corpus linguistics with research activities and difficulty using corpus technology (Boulton, 2008).

Several researchers have developed their own versions of corpus-based language pedagogy, such as the 4 Is (Moon & Oh, 2017) and a four-step approach (Ma et al., 2022). The teaching strategy used by the DDL group was based on Moon and Oh's (2017) use of Flowerdew's (2009) 4 Is framework for corpus-based activities: illustration, interaction, intervention and induction. In step one, illustration, students first look at lecturers' hand-selected, paper-based data, focussing on the concordance line patterns. Students engage in pair and group discussions and present their observations in response to the worksheet's prompts during step two, interaction. In place of clearly stating rules after step three, intervention, the teacher provides broader induction signals where necessary. The lower-level students' attention is directed towards the target patterns while the teacher helps them with any words they do not understand. This gives step four, induction, a more precise direction to look for the mismatch between learner and native speaker data. Unlike Moon and Oh (2017), based on Shulman's idea of pedagogical content knowledge, Ma et al. (2022) examined the process by which a group of prospective TESOL teachers learned corpus literacy and corpus-based language instruction. Based on Gass' (2001) L2 acquisition model, the study team developed a four-

step, corpus-based lesson plan: (1) evaluating students' knowledge, which entails spotting grammatical mistakes; (2) facilitating student-led corpus searches, such as searching for language patterns; (3) fostering student-led inductive discovery, such as eliciting language patterns; and (4) assigning output tasks, which involve actively using newly learned terms. This design combines corpus-based learning and teaching expertise with real-world classroom corpus use.

The present research adopted a four-step approach, close to Ma et al. (2022) but with several alterations. The first step was a vocabulary knowledge test to evaluate students' prior knowledge. In the second step, the printouts of the concordance lines were given out, and students were guided to discover the words' meanings and target collocations or patterns for the vocabulary. In this process, participants were encouraged to engage in inductive learning. Then, students were given much of the class time to complete a writing task and were required to incorporate the target vocabulary into their compositions. Here the writing process could trigger the student-led inductive discovery and summarising of language patterns according to their use. Finally, a VKS test was again given to the participants to measure their vocabulary knowledge development. In the present research, the writing tasks and the VKS tests were two output activities, but the writing was the purpose of students' inductive learning and part of their output.

The qualitative and quantitative interview data revealed that further alterations could be made to extend the potential for corpus-based language pedagogy. First, the lower proficiency

students were much more dependent on the teachers' instructions and scaffolding; as Alan and Jerry reported, it would be more effective if keywords were distributed and explained again. Second, the participants were more confident about turning to L1 translation. As Jerry said, 'I preferred to translate the meaning directly. A word, then a collocation, and then a translation. The best is that the target word has an example sentence below the target words'. Finally, until the last few tests, there were no statistically significant differences in the writing quality and vocabulary knowledge scales between the pre-, immediate post-, and delayed post-tests. In this sense, the lower-proficiency English learners with low learning motivation, and less confidence from their past years of unsatisfactory learning experiences, needed more time for training and to become experienced users of the DIY corpus.

From this discussion about corpus-based language pedagogy for lower-proficiency students, a revised method based on Ma et al. (2022) was proposed for applying the DIY-corpus-based language pedagogy to lower-proficiency participants. First, the evaluation of students' prior knowledge towards targets in the form of tests or exercises helped to arouse participants' interest and self-awareness of their knowledge gaps. Second, the teacher took the lead in conducting corpus searches because the less proficient participants were more dependent on the teacher's instruction, and they were less confident in their own discoveries, especially in an environment without sufficient hardware or internet access. Despite adequate training in the corpus tools, low-proficiency students cannot quickly stop engaging with their habitual learning behaviours, such as the reluctance to bringing laptops, reliance on the mobile translation applications, and so on. Third, students are responsible for inductive discovery

learning, in which the participants were able to self-exploit the corpus printouts or hands-on corpus materials using mobile devices or laptops. Fourth, the participants used what they had concluded from the corpus in the similar or identical exercises and tests administered in the first step. Still, some reordering of the test's items are incorporated. In this way, the participants can reflect on their improvement by comparing their prior knowledge with what they have acquired. By doing so, the participants can self-detect their progress, facilitating a sense of achievement and motivation for English learning. Their learner autonomy towards English learning might also be developed.

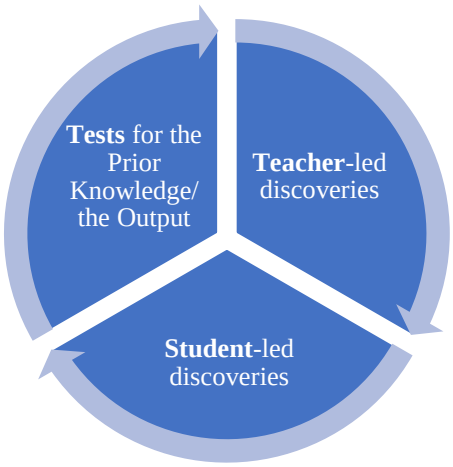


Figure 13. A revised version of CBLP for low-proficiency English learners

6.7 Summary of the Discussion and the Significance of the Study

Each of the input-based methods for learning vocabulary has a distinct drawback. Listening comprehension appears to be as complicated as vocabulary learning (Vandergrift, 2013), direct learning is too decontextualised (Oxford and Crookall, 1990, pp. 9-10), reading is too

slow and error-prone (Peters et al., 2009) and watching is more stressful for educators (Peters & Webb, 2018). In contrast to input-based methods, output-based approaches are theoretically sound in the context of the output hypotheses (Swain, 1995), and they have a relatively higher index in terms of ILH (Hulstijn & Laufer, 2001) and TFA (Nation & Webb, 2011) (a higher index indicates better learning outcomes). Corpus-based writing (Bao, 2018) facilitated the participants' writing quality improvement in this study.

Writing has not been empirically examined for vocabulary acquisition by prior research, despite being one of the main output activities. According to several studies (e.g., Chang et al., 2008; Daskalovska, 2015; Siyanova-Chanturia, 2015; Zou, 2019), the quality of collocation used when writing is a predictor of writing quality. Language is a system of interconnected words (Nation, 2022). Collocation is seen as a sign of near-native language proficiency (Chang et al., 2008), and multi-word speech comprises more than half of written discourse (Siyanova-Chanturia, 2015). Verb collocation, particularly verb-noun collocation, stands out among the various collocation errors made by Chinese EFL students (Zou, 2019; Wang & Zhou, 2020). In line with previous empirical research, writing practices helped improve the acquisition and accuracy of target verb collocations in this study, and writing quality was negatively correlated with erroneous verb collocations.

According to Webb and Nation (2017), vocabulary is the cornerstone of language and is vital to language mastery. Vocabulary acquisition is one of the numerous advantages of empirical, corpus-based studies (Bowker, 2018; Karras, 2015). However, Chinese researchers have

traditionally focussed more on describing language use, as seen in the CLEC (Chinese Learner English Corpus) (Yang et al., 2005), productive vocabulary use (Sun, 2017) and verb-noun collocation use (Wang & Zhou, 2020). Furthermore, corpus-based research abroad has been interested in advanced EFL learners (e.g., Gaskell & Cobb, 2004; Yoon, 2008; Cotos, 2014). Many of these empirical investigations, except for corpus-based studies by Luo and Liao (2015) and Luo (2016), have concentrated on writing quality rather than vocabulary development. In the present research, the mean differences between the pre-tests and the immediate and delayed post-tests revealed that the participants receiving corpus-based writing treatments were able to acquire more vocabulary knowledge and retain significantly more vocabulary knowledge.

The use of corpus-based language learning in this study was consistent with the methods of Benson (2001) and Dörnyei (2001), who advocate for fostering learner autonomy, but there is little empirical support of its efficacy with the exception of some qualitative research (e.g., Yoon, 2008; Charles, 2012). Even though Ma et al. (2022) have studied corpus-based language pedagogy, its use in the context of higher vocational institutes is rare in the field. The present research design was based on corpus-based language pedagogy (Ma et al., 2022). It was feasible to apply this pedagogy to higher vocational institutes with lower-proficiency English learners to evaluate writing quality improvement, the accuracy of verb collocation and target vocabulary knowledge. In addition, a revised version of corpus-based language pedagogy is proposed as a response to the experiences and feedback of the participants. The DIY corpus, one of the most recent advancements in corpus linguistics, has already

addressed the drawbacks of large general corpora that overwhelm learners with meaningless data (Charles, 2012). The experimental group participants welcomed the DIY corpus incorporation according to their generally positive feedback through the questionnaire, and they also provided some insightful advice on the format of the printouts and approaches to incorporating the corpus-based method.

Chapter 7: Conclusion

This section concludes the research with an overview of the research process and a summary of the major findings from the pilot study and the main study; subsequently, pedagogical implications for teachers and students, the limitations of the study and recommendations for further research are presented.

7.1 Overview of the Research Process

This study involved three stages of data collection: preparation (including a pilot study), data collection (experimental study and a semi-structured focus group interview) and overall interpretation. The pilot study explored the feasibility of incorporating corpus data into an English course for HVI students. The main study is an intervention study involving experimental and control groups to test the effectiveness of using a DIY corpus compiled by the teacher with low-achieving HVI EFL students. Tests and survey data were collected to answer the research questions from a quantitative perspective. In addition, qualitative data were collected through a semi-structured focus group interview to evaluate students' perspectives and perceptions towards the use of the DIY corpus.

The pilot study was conducted a semester earlier than the main study, in the middle of the first semester. There were three main purposes for the pilot study. Firstly, the researcher aimed to conduct a pilot study to test the effectiveness of hands-on corpus usage. Second, the researcher introduced large general corpora like COCA and BNC for students' use and

collected their feedback and comments for further research. Third, several instruments for data collection were piloted, including questionnaires regarding learner autonomy, vocabulary knowledge tests and topics for writing tasks. After two writing tasks were completed across five weeks, the researcher discovered that the participants still faced great difficulty in browsing the corpus websites due to their limited access to laptops and online barriers to accessing overseas websites. Moreover, the students' low English proficiency levels prevented them from understanding the concordance lines in the general corpora. Despite this discouraging feedback, students did evince interest in utilising the new method for English learning: the first writing task exhibited a significant difference between the experimental group receiving corpus aids and the control group. In addition, the reliability of the data collection instruments was higher than expected, though a few alterations were made afterwards.

Based on the students' reactions and feedback from the main study, the researcher turned to the DIY corpus for various reasons. First, the DIY corpus overcame the difficulties inherent in large general corpora, such as the inclusion of irrelevant data (e.g., Charles, 2012). Second, the DIY corpus has been widely applied with advanced students and to EAP courses and has received positive feedback (e.g., Zhang et. al., 2017; Charles, 2012, 2014, 2018). It seemed to be interesting to investigate whether a DIY corpus can benefit low-proficiency students. Third, the DIY corpus is theoretically related to the concept of learner autonomy (e.g., Cobb & Boulton, 2015), but there is a lack of empirical evidence. After reviewing literature, the researcher decided to compile a DIY corpus following the guidance provided by Charles

(2012, 2015), which could be tailored to suit the relatively weak students in the study.

The data collection in the main study was divided into three stages. To smooth the process of data collection, the first stage (preparation) was enacted about six months prior to the experiments. In this stage, participants from similar majors were intentionally chosen. Under the Human Research Ethics Committee's guidance, informed consent was obtained from the participants and the institute leadership to proceed with data collection. Next, the target writing tasks were selected according to the course syllabus and the participants' needs. In the meantime, the researcher compiled a DIY corpus in response to the limitations of students' busy schedules and their limited access to computers. This process entailed learning how to use AntConc software (Anthony, 2020) and compiling the DIY corpus in five steps (Charles, 2019): selecting the target texts, converting the files into text format, checking the conversions, renaming the text files as part of sub-corpora and finally cleaning up distracting symbols and mistakes. After this compilation, a more than 240,000-word corpus with two main sub-corpora, one from CET 4 and the other from Gaokao English tests, was ready for use. In addition to this preparational work, four writing tasks were selected in accordance with the syllabus in service to the institute and the participants' needs. After confirming the students' preferred platform for submitting their compositions (mobile phone), the target vocabulary for each writing task was chosen. The preparation stage ended with a planned corpus incorporation session and the administration of complementary handouts. The sessions were planned and conducted according to a four-step CBLP approach (Ma et al., 2022). Each session lasted for 80 minutes, with 10 minutes to evaluate the students' prior

knowledge, 25 minutes for studying the DIY corpus printout, 35 minutes to perform the writing tasks and 10 minutes for the post-tests, if possible.

The second stage was busier and continued for nearly an entire semester. First, the personal information questionnaire was administered during this phase to check for insignificant differences regarding demographic factors. Apart from the personal information questionnaire, the pre-tests for learner autonomy were also given out. Next, the experiments incorporating the corpus-based writing were carefully adopted in a two-week cycle, in addition to one week for the VKS pre- and post- tests. At the end of the intervention, a delayed post-test was administered for detecting vocabulary retention, a questionnaire was administered to evaluate learner autonomy and a questionnaire to evaluate participants' perceived reactions towards corpus-based English writing for the experimental group was also administered. In the last questionnaire, a short question about whether students would like to attend an experiment interview was included. A semi-structured focus group interview was successfully conducted with five participants at their preferred time and place. The participants in the experimental group offered great insights and suggestions to support the quantitative data.

The third stage occurred immediately after the data collection stage. All the compositions, VKS tests and questionnaires were graded or evaluated, as well as double-checked by another colleague with sufficient English learning and teaching experience. The quantitative data were further analysed via SPSS, and the qualitative interview data went through several

rounds of coding. Finally, the thesis is concluded in this chapter.

7.2 Summary of the Major Findings

After adding the DIY corpus into their writing practices, the students in the experimental condition enhanced their writing quality by increasing the number of proper target verb collocations in their writing. The fifth writing assignment was completed under exam conditions with teacher supervision, and individuals in the experimental group significantly outperformed those in the control group. Along with writing quality, the fourth and fifth writing tasks showed significant changes in the frequency of precise target verb collocation. Additionally, it was discovered that the quantity of incorrect verb collocations was inversely correlated with the quality of the writing. As for the target vocabulary knowledge within the writing tasks, participants in both groups increased their vocabulary knowledge. Students in the experimental group significantly outperformed their classmates in the fourth VKS, according to an independent sample t-test comparing the results of the immediate post- and pre-tests. Furthermore, compared to those in the control group, those in the experimental group retained more vocabulary knowledge.

The findings from the questionnaire and interview regarding learner autonomy indicate that individuals in the control group perceived a reduced responsibility for learning English, whereas the experimental group demonstrated growth in perceptions of this responsibility. Both groups also increased the number of times they engaged in extracurricular English

learning activities. A subsequent independent sample t-test of the mean differences revealed a significant difference in responsibility perception, with the experimental group scoring significantly higher than the control group. However, there was no discernible difference in how frequently the two groups of students learned English outside the classroom.

The final research question centred on the attitudes and opinions of the participants concerning the use of corpus data in English writing, and the experimental procedures received generally positive feedback. According to the questionnaire adapted from Nam (2010), the DIY corpus was perceived to be more straightforward to interpret than other corpora, and the concordance lines offered helpful examples for the writing process when coding the qualitative data from the interview. The interviewees also highlighted their worries about their poor English skills and the outdated equipment that prevented the corpus from being widely used. Finally, they offered advice on the structure of the DIY corpus printouts and the methods used to incorporate the corpus.

7.3 Implications

7.3.1 Pedagogical Implications

The pedagogical implications in this section fall into three categories: participants, DIY corpus and the corpus-based language pedagogy for incorporating corpus use for students. Lower-proficiency students need more time to become accustomed to and use new approaches to English learning. A teacher-compiled DIY corpus can not only offer examples

and lexical-grammatical support for creating relevant and appropriate materials (Charles, 2019), but it can also help students' development of vocabulary knowledge and learner autonomy. Corpus-based language pedagogy can be revised for the context of higher vocational institutes, and it is feasible for widespread implementation.

First, it is essential to exercise greater patience when addressing habitual misbehaviors exhibited by lower-proficiency students, such as skipping classes, dependence on mobile applications, and displaying low motivation. The participants in the present research came from a higher vocational institute, and their English proficiency was among the lowest in the tertiary education system. Moreover, their low motivation, low learner autonomy and limited access to hardware severely hindered their English development. These students still faced the pressure to pass the English exams in the syllabus and needed to improve their English proficiency for future career development in the Greater Bay Area, such as passing the CET 4 or CET 6. Consequently, the demand for English proficiency is crucial, and the disparity between students' current and desired levels of English proficiency presents a significant challenge. In the field of corpus-based language learning, corpora are widespread in EAP courses (e.g., Charles, 2012; Zhang et. al., 2017; Smith, 2020), especially for advanced students (e.g., Gaskell & Cobb, 2004; Cotos, 2014; Rahmanian & Soleimani, 2018; Charles, 2019), English majors (e.g., Daskalovska, 2015) and post-graduates (e.g., Yoon, 2008; Zhang et. al., 2017). According to the quantitative data of the current study, the writing quality and VKS tests did not show any significant differences in the first few tests; the low-proficiency participants needed more time for training and tolerance of their failure. In short, longer

training is better.

Second, teacher-compiled DIY corpora can be as effective as student-compiled DIY corpora. A DIY corpus is usually created by advanced learners themselves, as defined by Charles (2018): 'small-scale databases of electronic texts by users for specific, limited and local purposes'. Charles mainly instructed students to create their own corpora (e.g., Charles, 2012, 2014), such as a corpus of 10-15 research articles (Charles, 2012) or 20 papers from several important journals representing academic performance (Zhang et. al., 2017). The current research extended the advantage that 'a corpus can offer examples and lexico-grammatical support for creating relevant and appropriate materials' (Charles, 2019) by compiling the DIY corpus according to students' proficiency levels and needs. The present research not only demonstrated improved writing quality and target vocabulary knowledge within the writing tasks, but it also developed the low-proficiency participants' learner autonomy. A DIY corpus can help teachers become familiar with the specialised discourse they will need to teach a course in a new and uncharted topic and serve as a resource for them when they need to answer queries from students (Charles, 2019). In this vein, novice teachers can construct their own DIY corpora in order to equip themselves with specialised knowledge and useful reference materials for a specific area. Experienced teachers can respond to the needs of learners by controlling and adjusting the content of DIY corpora (Charles, 2018). Teachers can also create DIY tagged corpora with participants' erroneous utterances or electronic texts concerning students' target levels of proficiency to facilitate their learning.

Third, the present research adopted corpus-based language pedagogy (Ma et al., 2022), and the approaches for incorporating this pedagogy can be contextualised accordingly. The qualitative interview data did, however, indicate that additional adjustments may be made to increase the potential of corpus-based language pedagogy. First, during the interview, some students stated that it would be more efficient if keywords were provided and reiterated, and the lower-proficiency students were much more dependent on their teachers' instructions and scaffolding. Second, the group members felt more comfortable using L1 translations.

Thus, the four-step approach to corpus-based language pedagogy can be revised as follows for less proficient L2 learners. First, assessments of students' past knowledge of objectives in the form of tests or activities were made to get participants interested and make them aware of their knowledge gaps. Second, teacher-led corpus searches were facilitated because students with lower levels of proficiency are more reliant on teachers for guidance and are less confident in their knowledge, even in settings with adequate hardware or internet access. Low-proficiency students struggle to swiftly break their old learning habits while receiving enough corpus tool instruction. The third step is student-led inductive discovery, in which participants can independently use laptops or mobile devices to explore the printed or physical corpus. The participants used what they had learned from the corpus in the comparable or identical exercises and tests given in the first stage. By contrasting their existing knowledge with what they have learned, the participants can assess how much they have improved. The participants will be able to gauge their development in this way, which will help them feel successful and motivated to study English. Their learner autonomy

regarding English learning may also grow as a result.

7.3.2 Limitations and Future Research

The first limitation of this study is the large attrition rate of the participants for each learning session. The number of participants in the present research, especially for the control group, varied unexpectedly. Though the Cronbach's alpha values of the questionnaire and other scales indicated high levels of reliability, the varying number of participants threatened the triangulation.

The participants quit the study for various reasons, such as sickness, absence to attend to public or personal affairs or deciding they did not want to learn English and dropping the course. All these reasons imply a low motivation towards English as a compulsory selective course in the curriculum (MoE, 2021). Given their low proficiency, the students had already been discouraged due to their perceived unsuccessful attempts at learning English. The control group did not change their way of English learning, yet they faced the same stress of passing the course and acquiring a certain level of English ability to foster their future careers; they were passively placed in a dilemma in which they chose to ignore it or run away. The experimental group exhibited a more promising condition, evident from the significantly higher number of compositions submitted and questionnaires completed. The introduction of the corpus as a novel strategy for English writing and learning increased their learning interest. Despite facing years of unproductive English learning experiences, they

remained eager to explore this innovative method.

The second limitation is that students were assisted in their writing tasks by their mobile devices, giving them the chance to consult translation. This limitation partly contributed to the lack of a significant difference in writing quality in the first three writing tasks and the first three VKS tests. The writing practice was conducted in the usual classroom, and both groups of participants were habitual mobile phone users. Moreover, it was more convenient for the instructors to collect around 50 essays on the mobile platform in a shorter time. The experimental group was told to refer to the DIY corpus printouts only. However, the control group was allowed to refer to the mobile applications they liked. Some participants in both groups, especially those in the control group, used translation applications to copy and paste their essays. This limited the differences in apparent writing quality and affected the results of the VKS tests. Thus, in the final exam, the fifth task did not allow students to refer to the materials they liked, leaving them to rely on their own competencies.

For future studies, it is advisable to apply corpus-based research over an extended period and exercise increased patience when engaging with lower-proficiency students. Adult L2 learners with low proficiency tend to be less motivated and autonomous in their learning, often having experienced years of unsuccessful learning. Moreover, English courses at higher vocational institutes may have not received as much attention as other majors or career-related courses. Therefore, adopting a longitudinal research approach could potentially yield more comprehensive data on the effects of corpus-based learning for low-proficiency English

learners.

Despite its potential benefits, corpus-based learning has not become a mainstream methodology in English language pedagogy (Sun & Hu, 2020). Moreover, it has yet to gain widespread acceptance as a language teaching approach (Smith, 2020). Notably, this method remains largely unfamiliar to many educators in primary and secondary school settings (Ma et al., 2022). This is due to the absence of corpus-based learning in teacher training programs and the widespread perception that corpus linguistics is largely used for research purposes. Furthermore, the students' challenges in using corpus technology suggest that incorporating corpus-based language pedagogy should be considered as an opportunity to enhance the adoption of new technology, rather than a hindrance. The process of integrating corpus-based language pedagogy can be adapted to suit various contexts.

Comparative research offers another promising avenue, such as examining the effectiveness of corpora in comparison to traditional reference tools like e-dictionaries (Luo & Liao, 2015) and search engines (Luo, 2016). Future studies could also compare DIY corpora, such as learner corpora, with native speaker corpora (Cotos, 2014) or other types of corpora to evaluate the efficacy of researcher-compiled DIY corpora. The present research expanded upon the popular student-compiled DIY corpus approach by employing a teacher-compiled DIY corpus, which yielded promising results. Notably, the participants, who were lower achievers and less-resourced L2 learners, still relied on translation applications on their mobile phones during writing tasks. These participants were not advanced enough to compile

their own corpora and fully appreciate the autonomy that access to such resources could provide, which would allow them to reduce their dependence on native-speaking teachers and editors for text improvement (Charles, 2019).

7.4 Final Conclusion

The current mixed-methods empirical study investigated the effects of incorporating corpus-based writing on participants' writing quality, target verb collocations, target vocabulary knowledge, and learner autonomy development. Conducted within a higher vocational institute, the participants were low-proficiency English learners. The overall impact was moderate but promising. With regard to writing quality, as measured by graded essay scores, the experimental group outperformed the control group in the final writing task conducted during the final exam. Additionally, the frequency of correct target verb collocations displayed a significant difference in the last two compositions. Furthermore, writing quality was inversely correlated with the frequency of incorrect verb collocations. The three rounds of vocabulary tests clearly showed that the experimental group learned and retained a lot more language than the control group in terms of vocabulary knowledge. Concerning learner autonomy, participants exposed to the corpus-based writing experiment significantly improved their perceived responsibilities towards English learning; However, there was no discernible difference in the frequency of their outside-of-class English learning activities. In conclusion, the experimental group provided generally positive feedback, and several suggestions were offered for future research.

This study has several pedagogical implications. First, lower-proficiency L2 learners should be given more time to explore this new corpus-based approach. Second, a teacher-compiled DIY corpus can be effectively employed with low-proficiency students, contributing to the existing body of knowledge where DIY corpora are typically created by high-proficiency students as a self-learning tool. Third, corpus-based language pedagogy can also be adapted to cater to participants in varying contexts. Although the present research has its limitations, the findings regarding the implementation of the DIY corpus are encouraging and warrant further exploration.

References

Ackerley, K., & Coccetta, F. (2007). Enriching language learning through a multimedia corpus. ReCALL (Cambridge, England), 19(3), 351–370. https://doi.org/10.1017/S0958344007000730

Adamson, B. (2004). *China's English: a history of English in Chinese education*. Hong Kong University Press.

Alcaraz Marmol, G., & Almela Sanchez-Lafuente, A. (2013). The involvement load hypothesis: Its effect on vocabulary learning in primary education. *Revista Espanola de Linguistica Aplicada, 26*(26), 11–24.

Anthony, L. (2020). AntConc (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bal-Gezegin, B. (2014). An Investigation of Using Video vs. Audio for Teaching Vocabulary. *Procedia, Social and Behavioral Sciences, 143*, 450–457. https://doi.org/10.1016/j.sbspro.2014.07.516

Bao, G. (2019). Comparing Input and Output Tasks in EFL Learners' Vocabulary Acquisition. *TESOL International Journal. Vol. 14, Issue 1*, 1-12.

Benson, M. (1985). Collocations and idioms. In R Ilson (Ed.), *Dictionaries, lexicography and language learning,* (pp. 61 - 68). Oxford: Pergamon Press

Benson, M., Benson, E., Ilson, R., & Benson, M. (1997). *The BBI dictionary of English word combinations* (Rev. ed., enl. ed.). Amsterdam: John Benjamins.

Benson, M., Benson, E., & Ilson, R. (2010). *The BBI combinatory dictionary of English: Your*

guide to collocations and grammar (3rd ed.). Amsterdam: Benjamins.

Benson, P. (2001). *Teaching and researching autonomy*. Harlow: Pearson Education.

Benson, P. (2007). Autonomy in language teaching and learning. *Language Teaching, 40*(1), 21–40. https://doi.org/10.1017/S0261444806003958

Benson, P. (2013). *Teaching and researching autonomy* (Second edition.). Routledge.

Boulton, A., & Cobb, T. (2017). Corpus Use in Language Learning: A Meta-Analysis. *Language Learning, 67(2)*, 348–393. https://doi.org/10.1111/lang.12224

Bowker, L. (2018). Corpus linguistics is not just for linguists. *Library Hi Tech, 36(2)*, 358-371.

Chan, T., & Liou, H.-C. (2005). Effects of Web-based Concordancing Instruction on EFL Students' Learning of Verb-Noun Collocations. *Computer Assisted Language Learning, 18(3)*, 231–251. https://doi.org/10.1080/09588220500185769

Chang, J.-Y. (2014). The use of general and specialized corpora as reference sources for academic English writing: A case study. *ReCALL (Cambridge, England), 26(2)*, 243–259. https://doi.org/10.1017/S0958344014000056

Chang, Y.-C., Chang, J. S., Chen, H.-J., & Liou, H.-C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning, 21(3)*, 283–299. https://doi.org/10.1080/09588220802090337

Charles, M. (2012). 'Proper vocabulary and juicy collocations': EAP students evaluate do-it-yourself corpus-building. *English for Specific Purposes, 31(2)*: 93-102. https://doi.org/10.1016/j.esp.2011.12.003

Charles, M. (2014). Getting the corpus habit: EAP students' long-term use of personal

corpora. *English for Specific Purposes (New York, N.Y.), 35(1)*, 30–40. https://doi.org/10.1016/j.esp.2013.11.004

Charles, M. (2018). "Using Do-It-Yourself Corpora in EAP: A Tailor-Made Resource for Teachers and Students." *The Journal of Teaching English for Specific and Academic Purposes 6 (2)*: 217-224.

Charles, M. (2019). Do-it-yourself corpora for LSP: Demystifying the process and illustrating the practice. *Scripta Manent, 13(2)*, 156–166.

Coady, J., & Huckin, T. N. (1997). *Second language vocabulary acquisition a rationale for pedagogy*. Cambridge University Press.

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In *The Cambridge Handbook of English Corpus Linguistics* (pp. 478-497). Cambridge University Press.

Cohen, L., Manion, L., Morrison, K., & Morrison, K. (Keith R. B.). (2007). *Research methods in education (6th ed.)*. Routledge.

Cotos, E. (2014). Enhancing Writing Pedagogy with Learner Corpus Data. *ReCALL 26 (2014)*: 202, doi:10.1017/S0958344014000019

Craik, F. M., & Lockhart, R. S. (1972). Levels of processing: a framework for memory research. *Journal of Verbal Learning and Verbal Behavior,11*, 671-684

Creswell, J. W. (2012). *Educational research: planning, conducting, and evaluating quantitative and qualitative research* (4th ed.). Pearson.

Creswell, J. W. (2014). *Research design: qualitative, quantitative, and mixed methods approaches* (4th ed.). SAGE Publications.

Crookall, D., & Oxford, R. L. (1990). *Simulation, gaming, and language learning*. Newbury House Publishers.

Dafei, D. (2007). An exploration of the relationship between learner autonomy and English proficiency. *Asian EFL Journal, 24(4)*, 24-34.

Danan, M. (2004). Captioning and Subtitling: Undervalued Language Learning Strategies. *Translators' Journal, Vol. 49, No. 1*, 2004, p. 67-77.

Daskalovska, N. (2015). Corpus-based versus traditional learning of collocations. *Computer Assisted Language Learning, 28(2)*, 130–144. https://doi.org/10.1080/09588221.2013.803982

Dörnyei, Z. (2001). *Motivational Strategies in the Language Classroom*. Cambridge University Press.

Dörnyei, Z. (2007). *Research methods in applied linguistics: Quantitative, qualitative and mixed methodologies*. Oxford: Oxford University Press.

Fan, M. (2009). An exploratory study of collocational use by ESL students – A task based approach. *System (Linköping), 37(1),* 110–123. https://doi.org/10.1016/j.system.2008.06.004

Fløan, Tor Emil, F. (2015). You Click and Hold the Move Button. A Study on Incidental L2-Vocabulary Learning Whilst Playing Video Games. (Unpublished master's thesis). Norwegian University of Science and Technology, Norway.

Flowerdew, L. (2009) Applying corpus linguistics to pedagogy: A critical evaluation. *International Journal of Corpus Linguistics, 14(3)*: 393–417. https://doi.org/10.1075/ijcl.14.3.05flo

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System (Linköping), 32(3)*, 301–319. https://doi.org/10.1016/j.system.2004.04.001

Gass, S. M., & Selinker, L. (2001). *Second language acquisition: an introductory course* (2nd ed.). L. Erlbaum Associates.

Gilmore, A. (2009). Using online corpora to develop students' writing skills. *ELT Journal, 63(4)*, 363–372. https://doi.org/10.1093/elt/ccn056

Gilquin, G., & Granger, S. (2010). How can data-driven learning be used in language teaching? In A. O'Keefe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 359-370). London: Routledge.

Gohar, M. J., Rahmanian, M., & Soleimani, H. (2018). Technique Feature Analysis or Involvement Load Hypothesis: Estimating Their Predictive Power in Vocabulary Learning. *Journal of Psycholinguistic Research, 47(4)*, 859–869. https://doi.org/10.1007/s10936-018-9568-5

Gorjian, B. (2014). The effect of movie subtitling on incidental vocabulary learning among EFL learners. *International Journal of Asian Social Science, 2014, 4(9)*: 1013-1026.

Guan, X. (2013). A Study on the Application of Data-driven Learning in Vocabulary Teaching and Leaning in China's EFL Class. *Journal of Language Teaching and Research, 4(1),* 105–. https://doi.org/10.4304/jltr.4.1.105-112

Howarth, P. (1998). Phraseology and Second Language Proficiency. *Applied Linguistics, 19(1)*, 24–44. https://doi.org/10.1093/applin/19.1.24

Hu, & Nassaji. (2016). Effective vocabulary learning tasks: Involvement Load Hypothesis

versus Technique Feature Analysis. *System, 56*, 28-39.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. *Reading in a Foreign Language, 23*, 403–430.

Huang, Z. (2014). The effects of paper-based DDL on the acquisition of lexico-grammatical patterns in L2 writing. *ReCALL (Cambridge, England), 26(2)*, 163–183. https://doi.org/10.1017/S0958344014000020

Hulstijn, J. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cognition and second language instruction*. Cambridge, UK: Cambridge University Press

Hulstijn, J. H., & Laufer, B. (2001). Some Empirical Evidence for the Involvement Load Hypothesis in Vocabulary Acquisition. *Language Learning, 51(3)*, 539–558. https://doi.org/10.1111/0023-8333.00164

Huo, Y. J. (2014). Analyzing collocation errors in EFL Chinese learners' writings based on corpus. *Higher Education of Social Science, 7(1)*, 87–91.

Izumi, S., & Bigelow, M. (2000). Does Output Promote Noticing and Second Language Acquisition? *TESOL Quarterly, 34(2)*, 239–278. https://doi.org/10.2307/3587952

Izumi, S. (2002). OUTPUT, INPUT ENHANCEMENT, AND THE NOTICING HYPOTHESIS: An Experimental Study on ESL Relativization. *Studies in Second Language Acquisition, 24(4)*, 541–577. https://doi.org/10.1017/S0272263102004023

Izumi, S. (2003). Comprehension and Production Processes in Second Language Learning: In Search of the Psycholinguistic Rationale of the Output Hypothesis. *Applied Linguistics, 24(2)*, 168–196. https://doi.org/10.1093/applin/24.2.168

Johns, Tim. (1991). Should You Be Persuaded-Two Samples of Data-Driven Learning Materials. *ELR Journal*, *4*, 1–16.

Karabiyik, A. (2008). The Relationship Between Culture of Learning and Turkish University Preparatory Students' Readiness for Learner Autonomy. (Unpublished Master's Thesis). The Graduate School of Education of Bilkent University.

Karras, J. N. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL (Cambridge, England), 28(2)*, 166–186. https://doi.org/10.1017/S0958344015000154

Kim, Y. (2008). The Role of Task-Induced Involvement and Learner Proficiency in L2 Vocabulary Acquisition. *Language Learning, 58(2)*, 285–325. https://doi.org/10.1111/j.1467-9922.2008.00442.x

Kim, Y. (2011). The Role of Task-Induced Involvement and Learner Proficiency in L2 Vocabulary Acquisition. *Language Learning, 61(s1)*, 100–140. https://doi.org/10.1111/j.1467-9922.2011.00644.x

Krashen, S. D. (1995). *Principles and practice in second language acquisition*. Phoenix ELT.

Laufer, B. (1988). The concept of "synforms" (similar lexical forms) in vocabulary acquisition. *Language and Education, 2(2)*, 113–132. https://doi.org/10.1080/09500788809541228

Laufer, B. (1998). Passive and active vocabulary development in a second language: Same or different? *Applied Linguistics, 19 (2),* 255-271.

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

Laufer, B., & Waldman, T. (2011). Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning, 61(2)*, 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Lee, P., & Lin, H. (2019). The effect of the inductive and deductive data-driven learning (DDL) on vocabulary acquisition and retention. *System (Linköping), 81*, 14–25. https://doi.org/10.1016/j.system.2018.12.011

Lewis, M. (2000). *Teaching Collocation – Further Developments in the Lexical Approach*. Hove, England Language Teaching Publications.

Lewis, M. (2000). Introduction. In M. Lewis (Ed.), *Teaching collocation – Further developments in the lexical approach*. Hove: Language Teaching Publications.

Liang, M. C., Li, W. Z., & Xu, J. J. (2010). *Using Corpora: A Practical Coursebook.* Beijing: Foreign Language Teaching and Research Press.

Liu, Li-E. (2002). A Corpus-based Lexical Semantic Investigation of VN Miscollocations in Taiwan Learners' English. (Unpublished master's thesis). Tamkang University, Taiwan.

Li, K. Q. (2020, May). Report on the Work of the Government. State Council of People's Republic of China. Retrieved from http://www.gov.cn/guowuyuan/2020zfgzbg.htm

Liu Na, & Nation, I. S. P. (1985). Factors Affecting Guessing Vocabulary in Context. *RELC Journal, 16(1)*, 33–42. https://doi.org/10.1177/003368828501600103

Liu, S. (2017). An Experimental Research on the Effects of Types of Glossing on Incidental Vocabulary Acquisition through Reading *. *Journal of Language Teaching and Research, 8(4)*, 782-793.

Luo, Q., & Liao, Y. (2015). Using Corpora for Error Correction in EFL Learners" Writing. *Journal of Language Teaching and Research, 6(6)*, 1333. https://doi.org/10.17507/jltr.0606.22

Luo, Q. Q. (2016). The effects of data-driven learning activities on EFL learners" writing development. *Springer Plus (2016)5*: 1255. DOI 10.1186/s40064-016-2935-5

Ma, Q., & Sin, C. H. (2015). Teaching young learners English vocabulary with reading-based exercises in a real classroom situation. *Porta Linguarum, 23*, 125-138.

Ma, Q., Tang, J., & Lin, S. (2022). The development of corpus-based language pedagogy for TESOL teachers: a two-step training approach facilitated by online collaboration. *Computer Assisted Language Learning, 35(9),* 2731–2760. https://doi.org/10.1080/09588221.2021.1895225

Macaskill, A., & Taylor, E. (2010). The development of a brief measure of learner autonomy in university students. *Studies in Higher Education (Dorchester-on-Thames), 35(3)*, 351–359. https://doi.org/10.1080/03075070903502703

Mao, L., Liu, Y., & Zhang, M. (2018). Research on the Effectiveness of College Student English Writing Teaching Based on Data-Driven Learning. *Educational Sciences: Theory & Practice, 18(5)*, 1160–1169. https://doi.org/10.12738/estp.2018.5.017

Mayer, R. E. (2014). *The Cambridge Handbook of Multimedia Learning*. Cambridge University Press.

Ministry of Education. (2021). Curriculum Standards for College English in Higher Vocational Education (2021 edition). [高等职业教育专科英语课程标准（2021 年版）]. Retrieved from

http://www.moe.gov.cn/srcsite/A07/moe_737/s3876_qt/202104/t20210409_525482.html

Mizumoto, A., & Chujo, K. (2015). A meta-analysis of data-driven learning approach in the Japanese EFL classroom. *English Corpus Studies*, *22*, 1-18.

McCarthy, M., & O'Keeffe, A. (2010). *The Routledge Handbook of Corpus Linguistics (1st ed., Routledge handbooks in applied linguistics).* London: Routledge.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: method, theory and practice.* Cambridge: Cambridge University Press.

Moon, S., & Oh, S.-Y. (2018). Unlearning overgenerated be through data-driven learning in the secondary EFL classroom. *ReCALL (Cambridge, England), 30(1)*, 48–67. https://doi.org/10.1017/S0958344017000246

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: implications from cognitive psychology research for optimal spaced learning. *ReCALL (Cambridge, England), 20(1)*, 3–20. https://doi.org/10.1017/S0958344008000219

Nam, D. (2010). *Productive vocabulary knowledge and evaluation of ESL writing in corpus-based language learning*. Thesis (Ph.D.)--Indiana University, 2010.

Nation, I. S. P. (1982). Beginning to Learn Foreign Vocabulary: A Review of the Research. *RELC Journal, 13(1)*, 14–36. https://doi.org/10.1177/003368828201300102

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. Heinle & Heinle.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2022). *Learning vocabulary in another language.* Cambridge University Press. https://doi.org/10.1017/9781009093873

Nation, I. S. P., & Webb, S. A. (2011). *Researching and analyzing vocabulary* (1st ed.). Heinle, Cengage Learning.

Nesselhauf, N. (2005). *Collocations in a learner corpus*. Amsterdam/ Philadelphia: John Benjamins.

Paribakht, T. S., & Wesche, M. (1996). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In *SECOND LANGUAGE VOCABULARY ACQUISITION: A RATIONALE FOR PEDAGOGY*, Coady, James, & Huckin, Thomas [Eds], England: Cambridge U Press, 1997, pp 174-200 (pp. 174–200). Cambridge University Press. https://doi.org/10.1017/CBO9781139524643.013

Peters, E., Hulstijn, J. H., Sercu, L., & Lutjeharms, M. (2009). Learning L2 German Vocabulary Through Reading: The Effect of Three Enhancement Techniques Compared. *Language Learning, 59(1)*, 113–151. https://doi.org/10.1111/j.1467-9922.2009.00502.x

Peters, E., & Webb, S. (2018). INCIDENTAL VOCABULARY ACQUISITION THROUGH VIEWING L2 TELEVISION AND FACTORS THAT AFFECT LEARNING. *Studies in Second Language Acquisition, 40(3)*, 551–577. https://doi.org/10.1017/S0272263117000407

Qian, D. D. (2002). Investigating the Relationship Between Vocabulary Knowledge and Academic Reading Performance: An Assessment Perspective. *Language Learning, 52(3)*, 513–536. https://doi.org/10.1111/1467-9922.00193

Richards, J. C. (1976). The Role of Vocabulary Teaching. *TESOL Quarterly, 10(1)*, 77–89. https://doi.org/10.2307/3585941

Richards, K. (2003). *Qualitative inquiry in TESOL*. Palgrave Macmillan.

Russell, V. (2014). A Closer Look at the Output Hypothesis: The Effect of Pushed Output on Noticing and Inductive Learning of the Spanish Future Tense. *Foreign Language Annals, 47(1)*, 25–47. https://doi.org/10.1111/flan.12077

Sakai, S., & Takagi, A. (2009). Relationship Between Learner Autonomy and English Language Proficiency of Japanese Learners. *Journal of Asia TEFL, 6(3)*, 297–325.

Schmitt, N., & Meara, P. (1997). RESEARCHING VOCABULARY THROUGH A WORD KNOWLEDGE FRAMEWORK. *Studies in Second Language Acquisition, 19(1)*, 17–36. https://doi.org/10.1017/S0272263197001022

Seidlhofer, B. (2002) Pedagogy and local learner corpora: Working with learning-driven data. In: Granger, S., Hung, J. and Petch-Tyson, S. (eds.), *Computer learner corpora, second language acquisition and foreign language teaching*. Amsterdam/Philadelphia: Benjamins, 213–234.

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

Siyanova-Chanturia, A. (2015). Collocation in beginner learner writing: A longitudinal study. *System (Linköping), 53*, 148–160. https://doi.org/10.1016/j.system.2015.07.003

Smith, S. (2020). DIY corpora for Accounting & Finance vocabulary learning. *English for Specific Purposes (New York, N.Y.), 57*, 1–12. https://doi.org/10.1016/j.esp.2019.08.002

Spratt, M., Humphreys, G., & Chan, V. (2002). Autonomy and motivation: Which comes first? *Language Teaching Research, 6(3)*, 245-266.

Stubbs, M. (1995). Corpus evidence for norms of lexical collocation. In G. Cook & B. Seidlhofer (Eds.), *Principle & practice in applied linguistics: Studies in honour of H. G.*

Widdowson (pp. 245–256). Oxford: Oxford University Press.

Sun, Dongyun. (2017). The CEFR Stratification of English Productive Vocabulary of Chinese University Undergraduates Based on DIY Learner English Corpus. *Journal of Language Teaching and Research, 8(5)*, 909. https://doi.org/10.17507/jltr.0805.09

Sun, Xiaoya, & Hu, Guangwei. (2020). Direct and indirect data-driven learning: An experimental study of hedging in an EFL writing class. *Language Teaching Research: LTR, 136216882095445*. https://doi.org/10.1177/1362168820954459

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input and second language acquisition*. Rowley, MA: Newbury House.

Swain, M. (1993). The output hypothesis: Just speaking and writing aren't enough. *Canadian Modern Language Review, 50*, 158–164.

Swain, M. (1995). Three functions of output in second language learning. In. G. Cook and G. Seidhofer (Eds.) *Principles and practices in applied linguistics: Studies in honor of HG Widdowson* (p. 125-144). Oxford University Press.

SWAIN, M., & LAPKIN, S. (1995). Problems in Output and the Cognitive Processes They Generate: A Step Towards Second Language Learning. *Applied Linguistics, 16(3)*, 371–391. https://doi.org/10.1093/applin/16.3.371

Timmis, I. (2015). *Corpus Linguistics for ELT (Routledge corpus linguistics guides)*. London: Routledge.

Todd, R. W. (2001). Induction from self-selected concordances and self-correction. *System (Linköping), 29(1)*, 91–102. https://doi.org/10.1016/S0346-251X(00)00047-6

Vandergrift, L. (2007). Recent developments in second and foreign language listening comprehension research. *Language Teaching, 40(3)*, 191–210. https://doi.org/10.1017/S0261444807004338

van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System (Linköping), 41(3)*, 609–624. https://doi.org/10.1016/j.system.2013.07.012

Vidal, K. (2003). Academic Listening: A Source of Vocabulary Acquisition? *Applied Linguistics, 24(1)*, 56–89. https://doi.org/10.1093/applin/24.1.56

Vidal, K. (2011). A Comparison of the Effects of Reading and Listening on Incidental Vocabulary Acquisition. *Language Learning, 61(1)*, 219–258. https://doi.org/10.1111/j.1467-9922.2010.00593.x

Vyatkina, N. (2016). Data-driven learning of collocations: Learner performance, proficiency, and perceptions. *Language Learning & Technology, 20(3)*, 159–179.

Wang, W. Y., & Li, X. S. (2018). Investigating the Use of Verb–Noun Collocations in L2 Writing by Advanced Learners. *Journal of PLA University of Foreign Languages, 41 (1)*, 90-98.

Wang, W. Y., & Zhou, D. D. (2020). The Use of V–N Collocations in L2 Oral Production. *Foreign Languages and Their Teaching. Vol. 3*, 54-63

Webb, S. (2005). RECEPTIVE AND PRODUCTIVE VOCABULARY LEARNING: The Effects of Reading and Writing on Word Knowledge. *Studies in Second Language Acquisition, 27(1)*, 33–52. https://doi.org/10.1017/S0272263105050023

Webb, S. (2007). The Effects of Repetition on Vocabulary Knowledge. *Applied*

Linguistics, 28(1), 46–65. https://doi.org/10.1093/applin/aml048

Webb, S. A., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford University Press.

Wesche, M., & Paribakht, T. S. (1996). Assessing Second Language Vocabulary Knowledge: Depth Versus Breadth. *Canadian Modern Language Review, 53(1)*, 13–40. https://doi.org/10.3138/cmlr.53.1.13

Wesche, M. B., & Paribakht, T. S. (2000). Reading-Based Exercises in Second Language Vocabulary Learning: An Introspective Study. *The Modern Language Journal (Boulder, Colo.), 84(2)*, 196–213. https://doi.org/10.1111/0026-7902.00062

YAMASHITA, J., & JIANG, N. (2010). L1 Influence on the Acquisition of L2 Collocations: Japanese ESL Users and EFL Learners Acquiring English Collocations. *TESOL Quarterly, 44(4)*, 647–668. https://doi.org/10.5054/tq.2010.235998

Yang, H. Z., Gui, S. C., & Yang, D. F. (2005). *Corpus-based Analysis of Chinese Learner English (1st ed.)*. Shanghai, Shanghai Foreign Language Education Press.

Yilmaz, M. (2017). The Effect of Data-driven Learning on EFL Students' Acquisition of Lexico-grammatical Patterns in EFL Writing. *Eurasian Journal of Applied Linguistics (Online), 3(2)*, 75–88. https://doi.org/10.32601/ejal.460966

Yoon, H. (2008). More than a linguistic reference: The influence of corpus technology on L2 academic writing. *Language Learning & Technology, 12(2)*, 31–48.

Yuksel, D., & Tanriverdi, B. (2009). Effects of watching captioned movie clip on vocabulary development of EFL learners. *TOJET the Turkish Online Journal of Educational Technology, 8(2),* 48–54.

Zarei, A. A. (2009). The Effect of Bimodal, Standard, and Reversed Subtitling on L2

Vocabulary Recognition and Recall. *Pazhuhesh-e Zabanha-ye Khareji, No. 49, Special Issue, English*, Winter 2009, pp. 65-85.

Zhao, A., & Guo, Y. (2012). The Effect of Four Enhancement Techniques on Second Language (L2) Vocabulary Acquisition through Reading. *Hong Kong Journal of Applied Linguistics, 14(1)*, 48-68.

Zhao, A., Guo, Y., Biales, C., & Olszewski, A. (2016). Exploring Learner Factors in Second Language (L2) Incidental Vocabulary Acquisition through Reading. *Reading in a Foreign Language, 28(2)*, 224–245.

Zhang, L. X., & Li, X. X. (2004). A comparative investigation of learner autonomy of between Chinese students and students West Europe. [中国——西欧学生自主学习能力对比调查研究]. *Foreign Language World. No. 4* (General Serial No. 102).

Zhang, F., Zheng, Yunahua, & Li, L. (2017): Using Medical Academic English Corpus for Postgraduates Students Academic Writing Training. *Theory and Practice in Language Studies, Vol. 7, No. 10*, pp. 868-873, October 2017

Zou, D. (2017). Vocabulary acquisition through cloze exercises, sentence-writing and composition-writing: Extending the evaluation component of the involvement load hypothesis. *Language Teaching Research: LTR, 21(1)*, 54–75. https://doi.org/10.1177/1362168816652418

Zou, D., & Xie, H. (2018). Personalized Word-Learning based on Technique Feature Analysis and Learning Analytics. *Educational Technology & Society, 21(2)*, 233–244.

Zou, Q. (2019). A corpus-based study of verb-noun collocation errors in Chinese non-English majors' writings. In 4th international conference on contemporary education, social sciences

and humanities (ICCESSH 2019). Paris: Atlantis Press.

Appendix I Personal Information Questionnaire

Human rights declaration 人权宣言

You are welcome to participate in the research about incorporating the corpus data into English writing. In the experiment, you are supposed to write the essays using target vocabulary designated by the teacher. Later your performance on the writing will be carefully analyzed for research purposes. If you have any concerns about this research, you can withdraw without any impact on your final score of the subject. Yet any questions will be kindly answer by the teacher.

欢迎你参与将语料库数据整合到英语写作中的研究。在实验中，你应该使用老师指定的目标词汇来写文章。稍后，我们将仔细分析您在写作方面的表现，以便进行研究。如果你对这项研究有任何顾虑，你可以退出，而不影响你的最终成绩。然而，任何问题都将由老师友好地回答。

Personal Information（个人信息）

Name（姓名）:_____ Gender（性别）:_____

Year of Birth （出生年）:_____ Hometown （家乡）:_____

Major（专业）:_____

Years of English Learning （学习英语年份）:_____

Your last English score （上一次英语得分）:_____ Total Score（满分）:_____

The time spent in English speaking country （你在讲英语国家待过的时间）:_____

Multiple Choice （选择题）

1. （多选） Which aspect of English knowledge you find most difficult to learn? 你认为英语最难知识点？

A. Vocabulary 词汇 B. Grammar 语法

C. Pronunciation 发音 D. Logical Thinking 逻辑思维

2. （单选）Which one of the four English proficiency is hardest to you? 四项英语能力哪个方面对你最难？

A. Reading 阅读 B. Writing 写作

C. Listening 听力 D. Speaking 口语

3. （单选）How would you describe yourself about motivation to learn English? 对于英语学习的动力，你要如何描述你自己？

A. Highly-motivated to learn English 高度积极地学习英语

B. Well-motivated to learn English 相当积极地学习英语

C. Motivated to learn English 积极地学习英语

D. Slightly-motivated to learn English 比较积极地学习英语

E. Not-at-all motivated to learn English 一点也不积极地学习英语

Appendix II Essay writing tasks (adapted from CET 4 tests and the past final exams of the Institute)

1. For this part, you are allowed 30 minutes to write a letter to a foreign friend who wants to study Chinese. Please recommend some methods to him. You should write at least 120 words but no more than 180 words. *You must include the following vocabulary in your composition: participate/ read/ watch/ listen/ sing/ make friends

2. For this part, you are allowed 30 minutes to write a short essay on how to best handle the relationship between parents and children. You should write at least 120 words but no more than 180 words. *You must include the following 7 vocabulary in your composition: value/ identify/ compromise/ order/ return/ send/ assure/

3. For this part, you are allowed 30 minutes to write a letter to complain about the service of a hotel you lived in during your stay. You should write at least 120 words but no more than 180 words. *You must include the following 7 vocabulary in your composition: take/ cause/ apologize/ expect/ deliver/ chat/ serve

4. You, a customer relations advisor in a company, have got a complaint letter from your customer. Write a reply. * establish, locate, convince, give, spot, submit, confront

The complaint letter:

To:	**Customer Relations Advisor**
From:	**Johnemail@ghi.com**
Subject:	**Wrong Size of the Shoes**

Dear Sir/Madam,

We are writing to you about the shoes bought from your company last Monday.

To our regret, we found that the goods do not conform to the samples. The size of the shoes is different from your samples. We ordered 40 pairs of shoes in size 45, but you sent us 40 pairs of shoes in size 46. You should have checked carefully before sending them.

We would like you to replace the whole in the correct size immediately, as we are in urgent need of them.

We expect you to pay prompt attention to this matter. If you are still unable to deliver them in 3 days, we shall reluctantly cancel our order and ask for a full refund and compensation.

Kind regards,
John Smith

Appendix III Vocabulary and Vocabulary Knowledge Scale

The bilingual Vocabulary Knowledge Scale composed of 5 levels as follows (Paribakht & Wesche, 1996):

The bilingual Vocabulary Knowledge Scale composed of 5 levels as follows (Paribakht & Wesche, 1996):

1. I do not remember having seen this word before
   我从未见过该词
2. I have seen this word before, but I do not know what it means
   我见过该词但不知道其意义
3. I have seen this word before, but I think it means _____ (synonym or translation)
   我曾见过该词，我觉得它的意思是（翻译或同义词）
4. I know this word. It means _____ (synonym or translation)
   我确认我认识该词，且它的意思是（翻译或同义词）
5. I can use this word in a sentence, e.g.: _____ (if you do this section, do section 4)
   我能在句子中使用该词（如果你做这个也要完成第4题）

Appendix IV An excerpt of Learner Autonomy on Perception of Responsibility toward English Learning

Your Perception of Responsibility toward English Learning 你对英语学习责任的认识 (adapted from Spratt, Humphreys & Chan, 2002) When you are taking classes, how much responsibility should you take concerning the following items? 当你在上课时，你认为对下列事项的应承担多少责任？

1. Not at all 一点也没有 2. Hardly 几乎没有责任 3. To some extent 有一些责任 4. Mostly 几乎都是我的 5. Totally 完全是我的

1) To decide your goal of study in one semester 去决定你这个学期的目标
2) To check how much progress you make 去检查你取得多少进步
3) To decide the learning materials you use in class 去决定你在课堂上使用的学习资料
4) To decide topics and activities you learn in class 去决定你上课的话题和活动
5) To stimulate your interest in learning English 去激励你英语学习的兴趣
6) To decide the type of classroom activities, such as individual, pair and group work 去决定课堂活动类别，比如个人、双人或小组活动
7) To decide how long to spend on each activity 去决定每项活动花费的时间
8) To decide what you learn outside class 去决定你课外学习的内容
9) To assess your study 去评价你的学习
10) To evaluate the course 去评价你的课程

Appendix V Interview guidelines

After reviewing the questionnaire about using corpus tools, can you please answer:

1. At the beginning of introduction of corpus tools I recommended the website of COCA, how do you feel about using it?

2. Did you use *AntConc* after the training of its functions in or after class?

3. Compared with corpus websites and its software, how do you feel about reading the concordance printouts?

4. It is found that your perception about learning responsibility towards English learning has been improved significantly, do you feel the same way? Why or why not you think so?

5. Did you notice your writing proficiency has been improved, especially the verb errors decreased significantly? How or why could this happen?

6. Of the 22 verbs investigated in the VKS, … have exhibited significant difference from that of the control group, how or why could this happen? Is this difference related to the use of concordance printouts from your perspective?

Appendix VI Grading Scheme

Category One (18-20 points): write all the main points of the content, and the sentences are fluent. There are basically no errors in grammar and vocabulary. A few errors are mainly due to higher-level vocabulary or complex structures. The number of words meets the requirements. Basically, no word spelling and punctuation errors.
Category Two (13-17 points): write most of the main points. The sentences are smoother. There are a few errors in grammar and vocabulary, but they do not affect comprehension. There are a few misspellings and punctuation errors.
Category Three (8-12 points): write some key points. The sentence is basically smooth. There are some grammatical and vocabulary errors, a few words and punctuation errors.

Category Four (3-7 points): write a small number of main points, and the sentences are not smooth. There are more errors in grammar, vocabulary, spelling, and punctuation. Only a few sentences are correct.
Category Five (0-2 points): the content is basically not written, and there are many errors in grammar, vocabulary, and spelling. There are basically no smooth sentences. Just write a few words and don't know what to do.

Appendix VII Questionnaire about reactions to using the printouts from the DIY corpus.

Part A. The following questions are regarding your opinions on using the printouts from the DIY corpus English writing. Please use the following scale below and check the most closely resembles your perspectives.

1: Strongly disagree 2: Disagree 3: Undecided 4: Agree 5: Strongly agree

English writing proficiency improvement

1. I would feel confident in writing in English. 我会对用英语写作充满信心。
2. It would be helpful to write essays by using corpus printouts. 使用语料库打印素材会对写文章有帮助。
3. Using the corpus printouts would be useful for selecting the right vocabulary between synonyms in writing. 使用语料库打印素材会对在写作时选择合适的同义词有帮助。
4. Using the corpus printouts would be helpful for dealing with the preposition usage in my writing. 使用语料库打印素材会帮助写作中的介词使用。
5. Using the corpus printouts would be helpful for dealing with the word phrase usage in my writing.使用语料库打印素材会对写作时的短语使用有帮助。
6. Using the corpus printouts would be helpful for my writing in consistent and organized way. 使用语料库打印素材会帮助我使得写作连贯和有条理。
7. In general, writing with the corpus printouts would increase my writing proficiency in English. 总体而言，使用语料库打印素材写作会提升我英语的写作水平。

Reaction to the corpus printouts

8. The corpus printouts searching technique would be easy to learn. 语料库打印素材搜索技能是很容易学习的。
9. I would have difficulty in using the corpus printouts due to time and effort spent on the concordance lines.
我会觉得使用语料库打印素材有难度因为在阅读索引的词条花费的时间和精力。
10. The corpus printouts would be useful for finding appropriate words than other references, such as dictionaries.
使用语料库打印素材跟其他参考如词典相比会更效。
11. I would feel uncomfortable with using the corpus printouts due to unfamiliar vocabulary on COCA output.
我会对语料库打印素材显示的不熟悉的词汇感到不舒服。
12. The corpus printouts output would provide enough information needed to find out the usage of the vocabulary.
我认为语料库打印素材为我提供了充足的内容去找到词汇合适的用法。
13. I would not have problems in analyzing corpus printouts lines.
在分析语料库打印素材显示的词条时我没有问题。
14. Corpus printouts would be useful to my first draft. 语料库打印素材会对我的初稿有帮助。

15. Corpus printouts would be useful to my second draft. 语料库打印素材会对我的二稿有帮助。

16. The corpus printouts feedback would give me useful reference when I write and revise essays. 当我再写作和修改文章时，语料库打印素材的反馈会给我有用的参考。

Part B. Please answer the following questions.

17. Using the corpus printouts, I would feel that I would be most likely to make noticeable improvements for my writing in: 使用料库打印素材我会感受到我能够在写作的以下方面有显著的进步

1). learning the meaning of the vocabulary. 学习到词汇的含义

2). correcting grammar. 改正语法

3). using vocabulary appropriately. 恰当地使用词汇

4). organization. 篇章结构的组织

18. When you would be asked to revise the essays, how was the corpus printouts feedback useful for the following components? 当你被要求去修改文章时，料库打印素材的反馈对以下哪些部分有用？

1). learning the meaning of the vocabulary for writing. 写作中的词汇含义

2). correcting grammar in writing. 写作时订正语法

3). using vocabulary appropriately in writing. 在写作中使用恰当的词汇

4). organization of writing. 写作时的组织构架

Part C. If you have any experiences (either good or bad) while using the corpus printouts and you would like to share, please provide them below. 如果在使用料库打印素材上有或好或坏的经历愿分享，请写下。

_____

_____

Appendix VIII A sample of handout for vocabulary input for writing task three.

1. take (801)

President Clinton is trying to *take* a half-hour snooze every afternoon.

it should *take* him four years of full-time study to reach the…

It will *take* me much time, so I must finish the homework.

You will believe that this will *take* more time than you actually have.

2. cause (143)

I hope the change will not *cause* you too much trouble.

Radiation from the screen can *cause* various diseases.

It seems the things that *cause* us to lose the most sleep.

3. apologize (6)

Miss Mark visits my home to *apologize* to my parents after work.

I want to *apologize* to you because I can't attend the meeting.

4. expect (57)

Supposing the man had children, what would he *expect* them to do with their pocket money?

And then state your problem and what you *expect* to have done about it. Be polite!

To invest in their daughters because they do not *expect* them to make an economic contribution to the family.

5. deliver (13)