A Cross-Sectional Study of English-Major Students' Receptive and Productive Vocabulary Knowledge

Peeraya Utsajit

Submitted in accordance with the requirements for the degree of
Doctor of Philosophy

The University of Leeds
School of Education

March 2022

The candidate confirms that the work submitted is her own work and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgement

First and foremost, I would like to express my deepest appreciation to my supervisors, Dr. Richard Badger and Professor Alice Deignan, for their valuable advice, encouragement, unwavering support, and tireless patience. Without their help, this thesis would not be impossible. Throughout these four years of their supervision, they have also shown me how good supervisors should be. They will always be role models for my teaching career in the future.

I would also like to extend my deepest gratitude to the committee at the transfer and the viva examination for their precious and practical suggestions and to Mahidol University, my funder, for giving me this invaluable opportunity to enhance my academic knowledge and to have an unforgettable experience living in the UK. As an English teacher, I will definitely pass these on to my students.

I am deeply indebted to teachers, students, staff, raters at Mahidol University, staff and my Thai friends at University of Leeds for their priceless contributions and time in both the pilot study and the main study. I am also grateful to my colleagues at our lively office in Hilary Place, to my Thai friends, and to Dr. Joanna Nair and Dr. Petpiboon Prasit at the PestalozziWorld organisation for your concern, encouragement, support, positive attitudes, and ingenious suggestions. I have learnt a lot from their experiences as PhD students. I am also thankful to a friend at Thammasat University for voluntarily checking the statistical data for me. I really appreciate all their help and had great pleasure of working with them.

Last but not least, the ones who are always be by my side and support me in any ways are my family. My husband sacrificed everything for me and for our daughter. He always took care of and played with our daughter when I was busy writing up the thesis. I always talked to him about my thesis. He has been a good listener even though he does not work in my academic field. He cooked, did household chores, did grocery shopping, took our daughter to the park and many others when I could not. He has always encouraged and supported me. I also thank my daughter for always being a good girl, for settling well in the UK, and for her patience when we travelled back and forth between the UK and Thailand. This little girl has always made me smile and cheered me up. I am grateful to my parents for your support and for your patience wating for me to be back home, especially to my father for taking care of my mother when I was still in the UK. I would also like to thank my mother-in-law for delicious Thai food as a welcome every time we travelled back to Thailand. Therefore, this thesis is dedicated to them all. I do really appreciate their contributions.

Abstract

This study explores the relationship between receptive and productive vocabulary knowledge. The relationship between productive and receptive vocabulary can be framed as dichotomous (with two separate stores), or developmental (with words that start as part of the receptive state moving to the productive state). This study draws on both understandings.

The relationship was investigated at frequency levels and different years of study. The study also makes a distinction between controlled productive and free productive knowledge. Receptive knowledge was analysed using the first four categories (a word-recognition task and a translation task) of the Vocabulary Knowledge Scale (VKS) (Paribakht and Wesche, 1997). Controlled productive use was investigated by the fifth category of the VKS (a sentence-writing task). Free productive use data was collected with an argumentative essay-writing task by Laufer and Nation (1995).

To ensure consistency of the analysis, the same words and the same scoring systems were applied in these tests. The words produced in the free productive test were lemmatised, grouped based on frequency levels, and graded in terms of correctness of usage in order to facilitate comparison with the other data sets.

The data was quantitatively analysed within both the dichotomous and the developmental understandings of the relationship between receptive and productive vocabulary knowledge. Within the dichotomous approach, a three-scale scoring system was used to grade the correctness of the translations and the words used in the tests. Within the developmental approach, I tracked how the participants' word knowledge changed by adopting Paribakht and Wesche's (1997) five-scale scoring.

The data showed that all forms of vocabulary knowledge were all affected by frequency levels and years of study. The same data also showed that the knowledge moved forward and backward on a continuum. The findings were triangulate with qualitative analysis. Overall, the findings suggest that words cannot be simply classified into receptive or productive vocabulary stores. The study shows that we need a more sophisticated view of vocabulary knowledge that allows for different patterns of development for different aspects of vocabulary knowledge. Word knowledge gradually moves along the cline with its aspects moving to receptive or productive states at different degrees and at different time.

List of figures

List of tables

List of equations

Table of contents

Chapter 1 Introduction

1.1   Rationale of the study

From my teaching experience, Thai university students have difficulties with reading and writing in English because of their limited vocabulary knowledge. Also, only a few of the non-English-major students I taught in a foundation English course could cope with their reading materials. Some of the students from this group (personal communication) even said that in one line of the texts there were three or four words they did not know the meaning of, and this affected their comprehension. The problem was worse when they were assigned to write, whether a sentence, paragraph level or essay. They were not able to think of some English words they really needed and ended up using Google Translation to translate the whole sentence or the whole essay for them. These might be the results of inadequate vocabulary knowledge since they do not use the English language in everyday life. They are mostly exposed to the Thai language both in and outside class. The students majoring in English also experience these difficulties.

Throughout their four years as English majors, they take many of reading and writing courses such as Analytical Reading and Writing for Academic Purposes. The vocabulary they are mostly exposed to are general vocabulary, academic vocabulary, and some technical vocabulary. In Semester 2 of the academic year 2013, I had a chance to teach the first-year English-major students general English in an English reading and writing course called English 2. I found that when they read the teaching materials or some reading passages in the textbook, they seemed to comprehend the reading texts. They were trained to use some strategies to guess the meaning of an unknown word from contexts, and sometimes they were allowed to look up some words in the dictionary or on their mobile phones. These first-year students preferred using only English-Thai dictionaries, while those third-year and fourth-year students were encouraged to use English-English dictionaries. In pair-work, they might ask their classmates for the meanings if the contexts did not help them to get the word meaning. In writing tasks, they, however, seemed to have problems in producing words and they would use only the words that they were comfortable with, and few words taught in class were used in their essay writing tasks.

My colleague discussed with me her second-year English-major students in an academic writing class (College Writing 1). In one activity, the students were asked to match the target words – adjectives describing personality – to the sentences describing a person. In the controlled matching exercises (exercises with the target words provided), the students were able to match correctly. However, in a free or an open-ended exercise that the students were required to use adjectives they had not learned before and describe the persons explained in the sentences provided, only a few of the students were able to complete all the blanks correctly. These various experiences have generated my interest in the students' vocabulary knowledge, especially the vocabulary that English-major students understand when reading and that they can produce when writing.

In addition, when I read research articles about receptive and productive vocabulary knowledge, I found the debates on the relationship between these two types of vocabulary, whether receptive and productive vocabulary is separate stores of words in a person's mental lexicon, or it is a group of words moving along the cline. Scholars have tried to convince those who have different beliefs, but there has not been a consensus until now on the relationship between the two types of vocabulary.

My teaching experience, the discussion with my colleague, and what I found in the literature then lead to the research aim of the present study (Section 1.2).

1.2   Research aim

English-major students do not seem to have a problem understanding words when reading but have difficulty producing words that they desire to express the meaning of or are asked to produce in their writing tasks. These problems then have made me interested in examining words that can be understood and words that can be recalled and used, which I initially consider them as receptive vocabulary and productive vocabulary, respectively. Only vocabulary for reading and writing is the focus because university students need to practice these two skills considerably in their study throughout the four academic years.

Reviewing the literature, I found that there is no consensus on the relationship between these two types of vocabulary. The challenges are the debates on the relationship itself. The challenges also include deciding which tests are suitable for exploring receptive and productive vocabulary knowledge and deciding which scoring methods should be used. Therefore, the research aims are to explore vocabulary understood when reading, vocabulary produced in writing tasks, the

relationship between the two, and to find suitable methods for the consistency of the data analysis. I intended to explore this issue more fully in the hope that my research results can contribute to the literature on vocabulary knowledge, research methodology and pedagogy. When we have more insights into this we can move into the same direction and study more deeply to expand our knowledge in the field. For vocabulary assessment, more understandings will help guide ways to devise vocabulary tests that can capture both receptive and productive vocabulary in a consistent and systematic way. Pedagogically, the understandings will also help designing suitable materials, developing facilitating teaching strategies as well as efficient learning techniques. Suitable teaching materials, teaching strategies and learning techniques will help students to acquire more words and develop their productive vocabulary knowledge more efficiently.

Chapter 2 Literature review 1

Receptive and productive vocabulary knowledge

2.1 Introduction

In this chapter, I firstly review how scholars have theorised vocabulary knowledge (Section 2.2), how the notion of reception and production fit in such models, whether these models show the relationship between receptive and productive vocabularies, whether these models show how the two vocabularies are acquired and roughly how research adopts them when exploring receptive and productive vocabularies. All these fundamental ideas lead to the research questions (Section 2.3) and they are behind methodology of previous studies and research findings regarding receptive and productive vocabulary knowledge, which I review in detail in Chapter 3. These ideas, therefore, are essential for my understanding before I design my study and select the suitable tests to explore English-major students' vocabulary knowledge.

I conducted the literature search by selecting only articles about receptive and productive vocabulary or passive and active vocabulary through the library websites of University of Leeds and of Mahidol University. These articles are related to the nature of vocabulary knowledge, L2 vocabulary acquisition, English vocabulary assessment, and also classroom experiments for receptive and/or productive vocabulary knowledge. The participants from these research experiments are mainly university students and adult learners.

2.2 Vocabulary knowledge models and the position of receptive and productive knowledge

Meara (1984, p. 230) says "there seems to be a firm belief that the lexicon, unlike our knowledge of syntax and phonology, is an inherently messy part of our linguistic competence". Consequently, researchers have been attempting to characterise vocabulary knowledge for more comprehension, and how they examine vocabulary knowledge has mainly been based on how they characterise it. The characterisation can be divided into two major models: 1) the framework or separate traits model and

2) the global model (Henriksen, 1999). According to Milton and Fitzpatrick (2014), the two models can be explained by using the component approach or the developmental approach. The component approach emphasises what are included in such model, while the developmental approach addresses how vocabulary knowledge changes from one state to another, assuming that such a change takes place (Milton and Fitzpatrick, 2014).

2.2.1 The framework model within the component approach

Meara (1996a) writes that the framework model appears to have developed from Richards's (1976) assumptions about what it means (for a native speaker of a language) to know a word. These assumptions draw on research findings in the disciplines of linguistics, psycholinguistics, and sociolinguistics and from the analysis of types of vocabulary teaching exercises. Richards (1976) considers these as assumptions because in the 1970s vocabulary did not receive much attention from scholars and the knowledge and theory were still changing. He proposes these assumptions in the hope that they can be used to determine the objectives of vocabulary pedagogy. That is, he aimed to encourage teachers, material developers, and course or curriculum designers to realise that there are different aspects of vocabulary usage to concern and focus on when teaching, developing teaching materials, or designing the syllabus.

The eight assumptions made by Richards (1976, pp. 78-82) are as follows:

"Assumption 1: The native speaker of a language continues to expand his vocabulary in adulthood, whereas there is comparatively little development of syntax in adult life" (p. 78). He suggests that vocabulary teaching aim at vocabulary expansion so that learners have better productive and receptive language skills.

"Assumption 2: Knowing a word means knowing the degree of probability of encountering that word in speech or print. For many words we also "know" the sort of words likely to be found associated with the word" (p. 79). He adds that a native speaker of a language tends to be able to identify words as common, rare, unfamiliar, or unknown ones, and tends to have knowledge of collocation.

"Assumption 3: Knowing a word implies knowing the limitations imposed on the use of the word according to variations of function and situation" (p. 79). According to Richards, this is known as register characteristics, which is word variations used in different periods of time, places, situations, fields of discourse, modes of discourse, and by people from different social classes.

"Assumption 4: Knowing a word means knowing the syntactic behaviour associated with that word" (p. 80). He points out that vocabulary and grammatical properties can hardly be separated, and even though grammar is useful, learners should be aware of exceptions as well.

"Assumption 5: Knowing a word entails knowledge of the underlying form of a word and the derivations that can be made from it" (p. 80). A native speaker of English learns different forms of words, which include underlying forms (base forms), regular derivatives (inflections), and derivatives; learners can benefit from direct teaching that encourages them to recognise base forms and inflexional and derivational affixes (Richards, 1976).

"Assumption 6: Knowing a word entails knowledge of the network of associations between that word and other words in language" (p. 81). When asked to supply words associated to the prompt, native speakers seem to provide words with these relationships: antonymy, synonymy, subordinative classification, coordinative classification, and superordinative classification (Richards, 1976). He also adds that in Henning's (1973) study L2 learners tend to store words with similar sounds together in their early stage of learning and in later stage using semantic relationships to store words instead.

"Assumption 7: Knowing a word means knowing the semantic value of a word" (p. 82). Richards states that in L1 a native speaker tends to know whether this word represents a living or non-living thing and know emotional connotation attached to it as words are used in everyday life with actions, gesture, and emotion. He adds that learning L2 words mostly happens in classroom and students tend to attach plain sense to such words. He then recommends vocabulary exercises with context and discussion of word choices would be useful for L2 learners to improve knowledge of semantic value of the target words.

"Assumption 8: Knowing a word means knowing many of the different meanings associated with the word" (p. 82), and a word meaning is dependent on its context and on the thought that the speaker or writer wants to express.

From these eight assumptions, Richards has attached the idea of reception and production by saying that vocabulary size is linked to performance of receptive language skills (reading and listening) and productive skills (writing and speaking) and by mentioning what a native speaker tends to know when hearing or seeing the word or when using it. However, most of the assumptions emphasize 'knowing' which seem to refer to only receptive knowledge.

Some scholars (e.g., Meara, 1996c; Milton and Fitzpatrick, 2014) have criticised Richards' framework of vocabulary knowledge. According to Milton and Fitzpatrick Milton and Fitzpatrick (2014, p. 4), while the other seven assumptions show the features of a word, the first one seems to involve vocabulary growth and is "out of place in a component analysis". Also, Richards' s framework is criticised by Meara (1996c) that it does not place aspects in the right order because the feature of word meaning should have been placed at the top of the list instead, so this is not a coherent theoretical framework.

"Although Richards was quite explicit about the pedagogical emphasis on his paper, this has not stopped other people from developing his ideas into a 'word knowledge framework' (Meara, 1996c, p. 4).

Nation (1990; 2001; 2013) consolidated these points into a taxonomy of vocabulary knowledge. He divided underlying form into two aspects of spoken form and written form and by grouping inflections and derivatives into one aspect of word parts. Nation shows how these aspects are interrelated by grouping all these nine aspects into three major categories: form, meaning, and use, which seems to be the order of vocabulary acquisition. Nation also explicitly maps the idea of reception and production onto this taxonomy. Each of the aspects proposed can be divided into two types of knowledge: receptive knowledge and productive knowledge. Receptive knowledge of an aspect is knowing or recognising that aspect when reading and listening, while productive knowledge of an aspect is being able to use that aspect when speaking or writing.

Morgan and Oberdeck (1930) mention that a learner's receptive or passive vocabulary refers to words with which he or she is so familiar that he or she can understand them while reading or listening and their productive vocabulary refers to words with which he or she is so familiar that he or she can use them when speaking or writing. Similarly, Crow (1986, p. 242) writes:

> "Productive knowledge of a word is traditionally defined as what one needs to know about a word in order to use it while speaking or writing (productive channels); receptive knowledge is what one needs to know in order to understand a word while reading or listening (receptive channels)".

Nation also states:

> "Receptive carries the idea that we receive language input from others through listening or reading and try to comprehend it. Productive carries the idea that we produce language forms by speaking and writing to convey messages to others" (p. 46-47).

He clarifies this in the form of questions and explains the question by giving an example of knowing the word 'underdeveloped' (pp. 48-50). Table 2.1 shows a taxonomy of vocabulary knowledge proposed by (Nation, 2013, p. 49), and I also include his explanation of each aspect.

Form	spoken	R	What does the word sound like? (Being able to recognise the word form when it is heard)
		P	How is the word pronounced? (Being able to say it with correct pronunciation including stress)
	written	R	What does the word look like? (Being familiar with its written form so that it is recognised when it is met in reading)
		P	How is the word written and spelled? (Being able to write it with correct spelling)
	word parts	R	What parts are recognisable in this word? (Recognising that it is made up of the parts 'under-', '-develop-' and '-ed' and being able to relate these parts to its meaning)
		P	What word parts are needed to express the meaning? (Being able to construct it using the right word parts in their appropriate forms)
Meaning	Form and meaning	R	What meaning does this word form signal? (Knowing that 'underdeveloped' signals a particular meaning) (Knowing what the word means in the particular context in which it has just occurred)
		P	What word form can be used to express this meaning? (Being able to produce the word to express the meaning 'underdeveloped')
	concept and referents	R	What is included in the concept? (Knowing the concept behind the word which will allow understanding in a variety of contexts)
		P	What items can the concept refer to? (Being able to produce the word in different contexts to express the range of meanings of 'underdeveloped')
	associations	R	What other words does this make us think of? (Knowing that there are related words like 'overdeveloped' backward' and 'challenged')
		P	What other words could we use instead of this one? (Being able to produce synonyms and opposites for 'underdeveloped')
		R	In what patterns does this word occur?

Use	grammatical functions		(Being able to recognise that 'underdeveloped' has been used correctly in the sentence in which it occurs.)
		P	In what patterns must we use this word? (Being able to use the word correctly in an original sentence)
	collocations	R	What words or types of words occur with this one? (Being able to recognise that words such as 'territories' and 'areas' are typical collocations)
		P	What words or types of words must we use with this one? (Being able to produce words that commonly occur with it)
	constrains on use (such as register, frequency)	R	Where, when, and how often would we expect to meet this word? (Knowing that 'underdeveloped' is not an uncommon word)
		P	Where, when, and how often can we use this word? (Being able to decide to use or not use the word to suit the degree of familiarity of the situation; at present 'developing' is more acceptable than 'underdeveloped' which carries a slightly negative meaning)
Note: 'R' stands for 'Receptive knowledge' and 'P' stands for 'Productive knowledge'			

Table 2.1: Nation's (2013) taxonomy of word knowledge aspects and explanation for knowing the word 'underdeveloped'

This taxonomy shows that when one knows a word, one can have the knowledge of these nine aspects of such word and can also have receptive and productive knowledge of each aspect. It is not necessary to know all these for a particular word, though.

Some researchers defined receptive and productive knowledge for some aspects slightly differently from Nation's taxonomy (e.g., Schmitt, 1998).

> Since at the time when Schmitt conducted his research in 1998, there was "no real guidance in the literature about how to relate productive and receptive meaning knowledge to each other …, [he] assumed unprompted explanations of meaning sense demonstrated productive knowledge…[and] prompted explanations to be related to receptive knowledge" (pp. 292-293).

Later in his book (2010), Schmitt has probably changed what he thinks about receptive and productive knowledge as he accepted that what Nation (2013) described about vocabulary knowledge (receptive/productive knowledge and aspects) in his taxonomy is the most comprehensive taxonomy.

The explanation of receptive knowledge proposed by Nation involves many key phrases such as 'being able to recognise' 'being familiar', and 'knowing', while that of productive knowledge involves the phrases like 'being able to say', 'being able to

write', 'being able to construct', 'being able to produce', 'being able to use' and even 'being able to decide to use or not use'. This shows that receptive knowledge and productive knowledge can be divided further, depending on which aspect is the focus.

Based on the framework model, Webb (2007b) writes about the development of aspects of word knowledge:

> "Although [the] assumption [that receptive knowledge precedes productive knowledge] seems logical, it might not be entirely correct. Learners might often gain productive knowledge of some aspects of vocabulary knowledge before they gain some aspects of receptive knowledge. Advanced learners are likely to gain productive knowledge of form and partial productive knowledge of grammatical functions from seeing or hearing the form of an unknown word. This might occur before they gain receptive knowledge of its meaning" (p. 90).

It can be inferred that we should look at one aspect at a time when we mention that reception precedes production because not all aspects develop at the same pace.

Also based on this framework, vocabulary researchers design different vocabulary test formats. For example, a checklist format by Meara and Jones (1988) or by Meara and Miralpeix (2015) measures the ability to recognise word forms of the target words, and Webb's (2008) translation test examine the ability to recognise the written form and to know what it means. This will be discussed in more details in chapter of previous studies on receptive and productive vocabulary knowledge – Chapter 3.

Although Laufer and Goldstein (2004) acknowledge that vocabulary knowledge consists of many 'subknowledges', or aspects, they state that the most important subknowledge in vocabulary use and vocabulary assessment is the knowledge of form and meaning (Laufer et al., 2004, pp. 204-205). They argue that a word is a unit of meaning – even when used with wrong grammatical function, it can be understood. They also add that the link between form and meaning has been the main feature of various lexical tests. This can be seen even in tests that do not seem to test the understanding of word form and meaning such as Read's (1993) Word Associate Test. That is, a test taker needs to recognise the target words and know the meaning of the target words and that of the choices before they can choose the choices related to the target words (Laufer and Goldstein, 2004, pp. 402-403). Therefore, Laufer and Goldstein (2004) and Laufer et al. (2004) propose that the knowledge of form and meaning is the central component of vocabulary knowledge.

Also, they propose that the aspect of form and meaning itself contains four more aspects (I will call them sub-aspects instead to avoid confusion with the nine aspects proposed by Nation (2013)). When they analyse the formats of the vocabulary tests available, they notice that vocabulary test types often involve the ability to recognise and recall the form or the meaning. They then suggest that the aspects of form and meaning can be further broken down into four sub-aspects, where they use the term 'passive' for 'meaning' and 'active' for 'form': passive recognition, active recognition, passive recall, and active recall. They first hypothesised that these four aspects have a hierarchy of difficulty or strength – passive recognition is probably the weakest sub-aspect and active recall is the strongest one. However, before their study, they were not sure whether active recognition or passive recall was the stronger. They produced the Computerised Adaptive Test of Size and Strength (CATSS).

> "The proposed test, known as the CATSS, aims to assess vocabulary size, i.e., knowledge of word meaning. It should be noted that we do not consider that strength of knowledge to equivalent to depth of knowledge, as the term 'depth' includes features in addition to meaning, e.g., grammatical properties, collocations, pronunciation, etc." (Laufer et al., 2004, p. 209)

Hence, the CATSS is the vocabulary size test that focuses on the knowledge of form and meaning, but not the knowledge of word use. The monolingual version of which was devised and validated by Laufer et al. (2004). The bilingual (English-Arabic and English-Hebrew) versions of which was devised by Laufer and Goldstein (2004). Later in 2019, Aviad-Levitzky, Laufer and Goldstein developed and validated the new monolingual (English-English) version of CATSS (Aviad-Levitzky et al., 2019).

> "The new CATSS overcomes [the] sampling limitation by testing words from 14 sequential frequency bands in a similar way to Nation and Beglar's (2007) VST. …Each frequency band in the new CATSS includes a sample of ten items" (Aviad-Levitzky et al., 2019, p. 349).

The CATSS includes four different tasks based on the four sub-aspects. In order to prove the hypotheses in terms of "a hierarchy of difficulty" (Laufer and Goldstein, 2004, p. 399) or "the sequence of the four strength modalities" (Laufer et al., 2004, p. 212), Laufer and Goldstein administered this test with 435 high-school and university students, whose L1 languages are 278 Hebrew, Arabic, and Russian. In the study by Laufer et al. (2004), the participants were adult learners from Melbourne and Auckland Universities. They found that these four sub-aspects are in a hierarchical order for the bilingual versions of the CATSS, passive recognition (the easiest), active recognition, passive recall and active recall (the most difficult) for all the frequency levels tested (the 2,000-, the 3,000-, the 5,000, and the academic word

level). They also found that passive recall best correlates with the learners' language performance.

For the new monolingual version, "the recall of meaning scores and the recall of form scores could be used to estimate the vocabulary size expected of learners at different proficiencies" (Aviad-Levitzky et al., 2019, p. 361).

The four sub-aspects can represent the strength of the knowledge of form and meaning. However, the scores of the passive recognition and the active recognition modalities in the bilingual version were slightly significantly different, but they were not significantly different in the monolingual version. Laufer et al. (2004) explained that this might be because the test takes needed to know the meaning of the distractors, which were L2 definitions. Hence, the two modalities were equal in terms of difficulty in the monolingual CATSS, and the version of the test also affect the participants' test scores as well.

Pedagogically, the test can be used as a screening and a placement test (Laufer et al., 2004, p. 202). Test takers are presented with total scores for each frequency level and for each of the four modalities. The test shows how well an individual word is known (Aviad-Levitzky et al., 2019, p. 350). For example, a learner with a good score in the passive modalities but a low score in the active modalities can indicate that he seems to be advanced in a reading class but not in a speaking or writing class (Laufer and Goldstein, 2004, p. 412).

Regarding the use of terms, as the sub-aspects of 'passive recognition' and 'passive recall' relate to meaning (or L1 translation) and the sub-aspects of 'active recognition' and 'active recall' relate to 'L2 word form', Schmitt (2010) then suggests that they be called 'meaning recognition', 'meaning recall' for receptive knowledge and 'form recognition' and 'form recall' for productive knowledge.

However, there is inconsistency in using terms here. While Schmitt (2010) considers 'form recognition' as productive knowledge, it is regarded as receptive knowledge to some scholars such as Meara (1990a) as he develops the Eurocentres Vocabulary Size Test (EVST), a yes/no test format, which tests learners' ability to recognise the written form of the target words when seeing it. Also, Nation (2013, p. 47) explains receptive and productive vocabulary use: "Essentially, receptive vocabulary use involves perceiving the form of a word while listening or reading and retrieving its meaning. Productive vocabulary use involves wanting to express a meaning through speaking and writing and retrieving and producing the appropriate spoken and written word form." The phrase 'perceiving the form of a word while listening or reading' seems to be related to the ability to recognise the spoken/written form while listening or reading and then 'form recognition' should be receptive knowledge. The

inconsistency in using terms also appears in the papers by Webb (2009, p. 363) and Schmitt (1998, pp. 290-292). While they consider that the ability to produce paradigmatic associates as productive knowledge of word associations, Nation (2013) views that as the receptive knowledge instead and the productive knowledge of associations should refer to the ability to "recall [an L2 word form] when presented with related ideas" (p. 538).

Additionally, Laufer (1998) divides productive knowledge into two different types: controlled productive knowledge and free productive knowledge. This is for vocabulary assessment purpose. She describes controlled productive knowledge as either the ability to provide or recall an L2 word form to the meaning given or the ability to use an L2 word when asked to do so. She describes free productive knowledge as the ability to use an L2 word freely in speech or writing. She argues that "the distinction between controlled and free active vocabulary is necessary as not all learners who use infrequent vocabulary when forced to do so will also use it when left to their own selection of words" (p. 257). In vocabulary assessment, controlled productive knowledge is reflected in test formats such as recalling an L2 word to fit in the provided context and using a given word in a sentence level, while free productive knowledge is reflected in test formats like writing an essay using a topic or pictures provided. Therefore, 'controlled productive knowledge' seems to involve only the aspect of form and meaning, but 'free productive knowledge' may require many more aspects such as grammatical functions, collocations, constraints on use, associations, word parts, and spoken/written form.

To summarise, Richards (1976) presents aspects that an individual word carries. Nation (2013) presents how the aspects are linked to one another by grouping them into categories and relates them to receptive and productive knowledge. Laufer and Goldstein (2004) divide the knowledge of form and meaning further into four sub-aspects and Laufer (1998) divides productive knowledge into two further kinds. These scholars theorise vocabulary knowledge by employing component approach, so the models they built up go deep in details but do not explain much about how vocabulary knowledge develop or how an L2 word is acquired. That is, they do not illustrate whether a word carries all these aspects with it when it is acquired, whether some aspects can be acquired before the other, whether its receptive knowledge is acquired before productive knowledge. Also, when a word is acquired, none of the models specify whether it will be stored as one group of words used receptively and productively or whether it will be stored as receptive vocabulary and once developed it will be stored as productive vocabulary. The studies adopting his taxonomy are reviewed in Chapter 3.

2.2.2 The framework model within the developmental approach

Using the developmental approach, Jiang (2000) proposed a psycholinguistics model of L2 vocabulary acquisition and presents the 'specifications' that an L2 lexical entry carries with it when it is acquired. His list of specifications is shorter than Nation's aspects, but they can also be grouped into the three major groups of form, meaning and use. Jiang explains that the internal structure of the lexical entry consists of four specifications:

1) phonological or orthographic formal specification (i.e., spelling/pronunciation or the aspect of form)

2) semantic specification (i.e., word meaning or the aspect of form and meaning and the aspect of concept and referent)

3) syntax specification (i.e., grammatical features or the aspect of grammatical functions)

4) morphological specification (i.e., morphological variants of a word or the aspects of word parts)

He proposes that there are three stages of how an L2 word is acquired into the L2 lexicon in L2 learning context: 1) the formal stage, 2) the L1 mediation stage, and 3) the L2 integration stage (p. 47), and this process shows that some aspects are acquired before others.

> "At the first stage, a lexical entry is established for an L2 word, but it only contains formal specifications and a pointer [or a lexical link]. At the second stage, an L2 entry is a combination of L2 formal information (in the lexeme) and the semantic and syntactic information of its L1 translation (in the lemma). The integration of L2 information other than formal specifications is realised at the third stage" (Jiang, 2000, p. 71).

This implies that formal specification of an L2 word is acquired before the other three specifications. He adds that the existence of L1 lexicon in an L2 learner leads to the L1 mediation stage, where language use is affected by L1 interference – using grammatical features of the L1 equivalent – and that many L2 words are often fossilised at the second stage. In vocabulary assessment or vocabulary teaching and learning, the fossilisation can be seen when learners write a sentence with some errors caused by L1 interference.

Aitchison (1994) also explains processes of vocabulary acquisition in L1 vocabulary learning; the order of aspects acquired is similar to that described by Jiang (2000).

L1 vocabulary learning starts from 1) labelling process, where an L1 word form is linked to its concept and referent, 2) packaging process, where the L1 word's derivative forms and figurative meanings are learnt, and 3) network building process, where the word is fitted in a learner's semantic network (Aitchison, 1994, p. 170). So, Aitchison mentions the aspect of associations and collocations but does not mention the aspect of grammar. Therefore, she argues, the aspects of spoken/written form and the aspect of form and meaning are acquired before the aspect of word parts, association and collocations.

Concerning receptive and productive vocabularies, Jiang's model of L2 vocabulary acquisition also include the notion of reception and production. When explaining the first stage of L2 vocabulary acquisition, Jiang (2000) mentions two directions of word use (receptive and productive use) and how words are activated in these two directions.

> "At this initial stage, the use of L2 words involves the activation of the links between L2 words and their L1 translations…, or lexical association, … In receptive use of the language, the recognition of an L2 word activates its L1 translation equivalent, whose semantic, syntactic, and morphological information then becomes available and assists comprehension. In productive L2 use, the pre-verbal message [concept] first activates the L1 words whose semantic specifications match the message fragments. The L1 words then activate the corresponding L2 words through the lexical links between L2 and L1 words" (Jiang, 2000, p. 51).

I drew the diagram (Figure 2.1) showing the directions of receptive use (understanding an L2 word) and productive use (producing an L2 word to match the concept/thought), by adapting the lexical representation shown in Jiang's paper.

Direction of productive use = Producing an L2 word

Concept (pre-verbal message) — activates — L1 word — activates — L2 word

Direction of receptive use = Understanding an L2 word
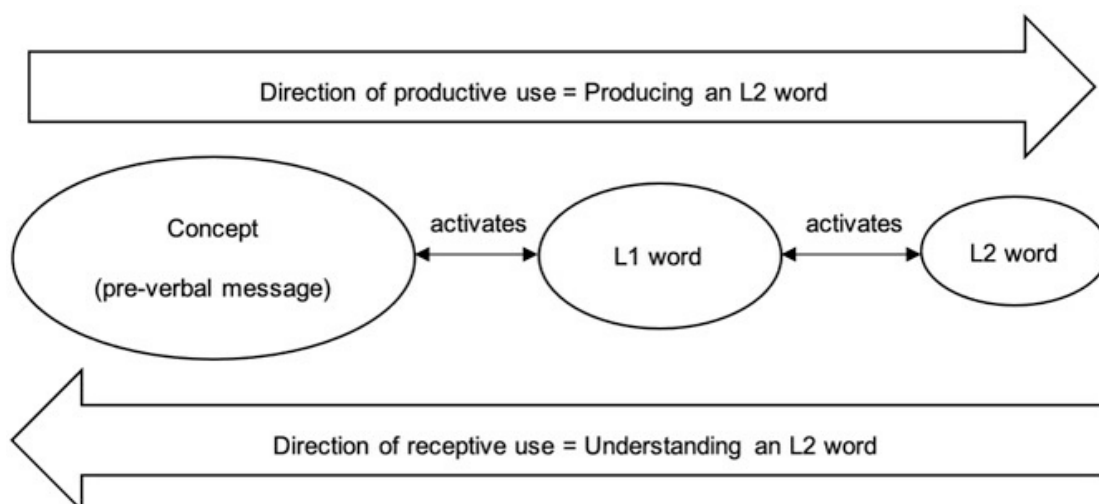
Figure 2.1: Directions of receptive use and productive use of a word in one's lexicon

If related with sub-aspects proposed by Laufer and Goldstein (2004), receptive use of an L2 word involves L2 word form recognition and word meaning (L1 translation) recall, while productive use involves L2 word form recall. Hence, Jiang's (2000) psycholinguistic model shows what specifications a word has, what specifications seem to be acquired earlier, and that a word is used receptively and productively. However, his model still does not show how words are stored in one's mental lexicon, whether they are stored as one group and used receptively and productively, or they are stored as receptive vocabulary first and used receptively and once they become productive they can be used productively. Based on his model, it is uncertain about the relationship between receptive and productive vocabularies.

2.2.3   The global model within the component approach

The models reviewed above are separate traits models built up with both the component approach and the developmental approach. Some scholars such as Meara (1996a; 1996b; 1996c) do not agree with how vocabulary knowledge is characterised into aspects and sub-aspects. Meara (1996c) argues that deconstructing vocabulary knowledge in that way does not seem to be the best way because it is too detailed and focuses too much on features of individual words; therefore, it cannot be used directly to assess learners' overall lexical competence. He also adds that for vocabulary assessment purpose it is not manageable to test all the aspects shown in Richards' (1976) framework with many test words. If all the aspects of a word need to be tested at the same time, only a few words can be included in a test battery and this again will not help researchers to explore learners' overall lexical competence (Meara, 1996c). He then suggests a new model to describe vocabulary knowledge, especially for assessment purposes. This is seen as the global model (Henriksen, 1999) built up with component approach (Milton and Fitzpatrick, 2014). Although he does not agree with how vocabulary knowledge was theorised previously, what he proposed relies heavily on the framework model.

In his papers (Meara, 1996a), Meara proposes two dimensions of vocabulary knowledge that are properties of the lexicon as a whole: 1) 'size' and 2) 'organisation'. Later, Meara (1996c) proposes another dimension 'automaticity'.

Meara (1996a, p. 2) states that the problems of vocabulary teaching and learning at that time is caused by the lack of clear understanding of vocabulary knowledge as a whole. Even some questions that might be beneficial to pedagogy and seem simple to answer such as 1) how many words are there in English? and 2) how many words do learners know? are difficult to answer. These questions are all about 'size'. The first dimension he proposes is then 'size', or the breadth of vocabulary knowledge. He considers this dimension as "the basic dimension of lexical competence" (Meara, 1996a, p. 3).

Milton and Fitzpatrick (2014) write that it is not clear which aspects the dimension of vocabulary breadth covers. What it covers varies in vocabulary assessment, but it mainly covers the aspect of spoken/written form and the aspect of form and meaning. Hence, the problem of selecting a test to measure vocabulary size is dependent on how vocabulary knowledge and the knowledge of form and meaning are defined. They can be defined differently as mentioned above when reviewing Nation's (2013) taxonomy.

Nation and Beglar (2007) write:

> "There are several reasons for wanting to measure a non-native speaker's vocabulary size. One reason is to see how close the learner is to having enough vocabulary to be able to perform certain tasks such as to read a novel, to read newspapers, to watch movies, and to listen to friendly conversations. … A second reason for measuring vocabulary size is to be able to chart the growth of learners' vocabularies. … A third reason for measuring vocabulary size is to be able to compare non-native speakers with native speakers" (p. 9).

The underlying idea for the research on receptive and productive vocabulary size is that learners' vocabulary size is strongly related to language skills. "[I]t is generally assumed that the more words learners know, the better the chance they have of understanding while reading or listening in the foreign language" (Al-Masrai and Milton, 2012, p. 15). Hence, a large vocabulary size is related to better language skills.

The link between vocabulary size and general language ability has led to the concept that learners need to pass a particular threshold of vocabulary to perform particular tasks. The most frequent 2,000 words are required for basic everyday communication, and the most frequent 3,000 words are for learners to begin reading an authentic texts or simplified text (Laufer, 1992; Schmitt et al., 2001). The knowledge of most frequent 5,000 words enables learners to guess word meaning from context, to comprehend the content of the authentic text and to cope with

unassisted reading (Hirsh and Nation, 1992, p. 689; Schmitt et al., 2001). The knowledge of most frequent 10,000 words is required for studying in a university level in an ESL/EFL context (Hazenberg and Hulstijn, 1996, p. 145; Read, 2000). Some certain amount of vocabulary is needed for some academic tasks. For example, according to Laufer (1992), only the knowledge of 3,000frequent words should be enough for comprehension when reading more advanced reading texts, and 5,000 word families was for comprehensive reading (Hirsh and Nation, 1992; Laufer, 1992).

However, there is no consensus in the number of vocabulary items needed. Nation and Beglar (2007, p. 9) and Schmitt (2008) suggest that the knowledge of around 8,000 word families is needed to cope with a range of unsimplified spoken and written texts. Laufer (1992) suggests a much small size of only around 3,000 word families. These suggestions about thresholds vary, but many researchers are still based on the idea and have measured receptive and productive vocabulary size and growth.

Vocabulary size tests vary. For instance, Meara and Jones's (1988; 1990) Eurocentres Vocabulary Size Test (EVST)) test only the ability to recognise written word form, while Nation and Beglar's (2007) Vocabulary Size Test (VST) test learners' ability to recognise the L2 written forms and the ability to recognise word meaning. If a vocabulary size test contains words with affixes, vocabulary breadth of such test then covers the aspect of word form as well as word parts. Therefore, what aspects the dimension 'size' cover depends on how researchers define vocabulary knowledge. "In practice, it seems that researchers choose the definition which suits them best and this is often dictated by the availability of the testing instruments they have to hand" (Meara, 1996c, p. 7).

The dimension of 'size' is related to the notion of reception and production. Scholars (e.g., Fan, 2000; Webb, 2008; Zhou, 2010) who explored this dimension tend to believe that there are two separate groups of vocabulary: receptive vocabulary and productive vocabulary in a learner's mental lexicon. They normally measured only receptive vocabulary size, measure only productive vocabulary size, compare the size of these two vocabularies, measure vocabulary growth cross-sectionally or longitudinally, and explore factors of the growth. The actual size of productive vocabulary is difficult to conclude, though (Palmberg, 1984; Takala, 1984). When comparing the size of receptive and productive vocabularies, they hypothesised that receptive vocabulary is larger than productive vocabulary. This has been criticised by Melka (1982, p. 5) that this is "the more or less intuitive hypothesis".

The ways that are used to distinguish receptive vocabulary from productive vocabulary might be different from researcher to researcher. For example, Meara (1990a) distinguishes these two groups of words on the basis of activation, and Corson (1983) distinguishes them on the basis of motivation.

Meara (1990a) uses the term 'active' and 'passive' interchangeably with the terms 'productive' and 'receptive'. He explains these two terms by using the metaphorical approach and Wilson and Beineke's (1979) Graph Theory, which he claims has been widely used to describe a range of phenomena, and associational structure of English vocabulary suggested by psychologists (e.g., Deese, 1965). This theory suggests that in relationships items or nodes are connected to one another via links or arcs; words are nodes, and they are linked. The direction of a link represents the direction of activation, and words that have no inward-pointing links are called passive words (Meara, 1990a). This means passive words can activate other words; they will not be activated by other related words but by external stimulation – seeing or hearing the words (Meara, 1990a, pp. 152-153). Active vocabulary is easily accessed because it is connected to inward-pointing and outward-pointing links. This means that active vocabulary can be activated by words stored in a person's lexicon. Therefore, he is certain that there are two groups of vocabulary in one's mental lexicon: active or productive vocabulary and passive or receptive vocabulary. They are distinguished by different links and different stimuli. If linked to the framework aspect, the relationship among words as he explains is related to only the aspect of word associations (with paradigmatic relations) and collocations (with syntagmatic relations).

He adds that some active words can be connected to many links, so they are keywords or central to the associational structure, while some words are linked by a few related words. In real life, it is difficult to check how many links a word has or hard to compare which word has more links than others, though.

Moreover, although Meara believes that receptive and productive vocabulary knowledge is a dichotomy, he also mentions the idea of degrees of knowledge. He remarks that the difference in number of links shows that active vocabulary exists on a continuum. This might imply that the number of links attached to active vocabulary refers to the degrees of how well we know it.

To summarise, although Meara suggests that passive vocabulary is not the same group of words as active vocabulary because the difference is qualitative, he accepts that 1) there are degrees in knowing both passive words and active words, 2) passive words can become active words if taught appropriately; he recommends

stressing links from known words to newly learned words and 3) the state of being active is more advanced and include the state of being receptive.

In terms of assessment, how he defines receptive or passive vocabulary influences how he devises a receptive vocabulary size test. As receptive vocabulary is activated by seeing/hearing a word, he then developed a word recognition test – the EVST. Defining productive vocabulary as words activated by other words, Meara and Fitzpatrick (2000) developed Lex30, which aims to elicit a learner's productive words that are related to the prompt (L2 form recall). This test is claimed to be a productive vocabulary size test, and it is in the format of recalling words to multiple activation prompts. It directly involves the knowledge of associations and collocations. There are 30 prompts in the test and a test taker is required to supply the form of any related words he can recall.

Fitzpatrick and Clenton (2017) studied performance of different productive tests and created the capture zone (p. 860) for teachers and researchers to "identify the level and range of the [productive] knowledge being tested" (p. 864). They showed that Lex30 can encourage test takers to supply various lexical items because of the multiple activation prompts and the words elicited can represent words test takers can produce, can use for appropriate referents, can use with semantic appropriateness, and use with grammar and semantic accuracy. However, words recalled in Lex30 should not be assumed to represent words that learners can use in context, and Lex30 should be categorised as a recall test, or a controlled productive test, which involves the knowledge of word association. The scores of Lex30 might be able to differentiate participants with large productive vocabulary from those with small productive vocabulary. However, it might not be able to provide the estimate of a test taker' s productive vocabulary size in one figure. Until now, none of the available productive tests has that capacity.

Similar to Meara (1990a), Corson (1983) used the same terms passive and active and explains this abstract idea by using a metaphorical approach. However, he had slightly different views about the relationship between passive and active vocabularies. Corson did not mention only activating factors but also what suppresses words; suppressed words are not used and the right words will be used instead. Corson stated that a learner's active vocabulary is an important part or subset of his/her passive vocabulary when he explained the phenomena of selecting words for use in speech. He mentioned that at the encoding level of production many words will be recalled in a learner's mind and only those with appropriate meanings will be chosen for use in speech.

Besides meanings and syntax, the necessity of explicitness of word use is also another factor in using certain words. These words are words ready to be used or easier to access, high-frequency words, and almost totally known. Corson called these words 'motivated words', and the rest will be hold passively because of both linguistic and extralinguistic constraints. Linguistic constraints involve inappropriate meaning or syntax, but extralinguistic context refers to psychological constraints such as willingness and inadequate knowledge of how to use a word, and sociological constraints such as the learner's dialects or varieties (Corson, 1983). All these influence a learner's word choice in their production. Corson called words that are not selected and will be held passively 'unmotivated vocabulary'.

According to Corson, unmotivated words can be divided further into two sub-groups: partially-known words and words that are not regularly used on everyday basis such as an unfamiliar term and swear words. These two sub-groups tend to be independent of each other but sometimes can overlap (Corson, 1983).

Corson gave an example of a conversation between a physician and a layman. He believes that selecting words to use in speech is affected by explicitness. When explicitness is required, a physician is likely to use medical terminology in his/her career and medical terms such as 'duodenum' instead of 'stomach' become motivated vocabulary or active vocabulary, which is ready to be used at all times in a physician's mental lexicon (Corson, 1983). Corson concluded that it is difficult for him/her to not use it in a conversation, even when he/she talks with a layman. Also, even though a layman understands the word 'duodenum' and is able to use the word if needed, he/she might prefer using 'stomach' in speech instead, but the word 'duodenum' is still in a layman's passive vocabulary and not used actively (Corson, 1983). To conclude, Corson also views that there are degrees in knowing words and understanding precedes use, or words are stored passively first and once they are forced or motivated by some factors, passive vocabulary can be used actively.

In addition to Meara (1990a) and Corson (1983), there are also many researchers who believe that receptive vocabulary and productive vocabulary are separate and compared the sizes of learners' receptive vocabulary with that of productive vocabulary (e.g., AbManan et al., 2017; Fan, 2000; Webb, 2008; Zhou, 2010). Their findings show that active or productive vocabulary is smaller than passive or receptive vocabulary. Melka (1982) summarises the explanations for this phenomenon from research findings. There are four explanations which include the following Melka (1982, pp. 17-18).

1) Some words in the lexicon are not used productively because of factors such as culture, education, social circumstances, psychological circumstances, and

these words should be considered as part of possible use knowledge or receptive knowledge.

2) Some words are not used because of learners' apparent avoidance or avoid strategies such as being unwilling to use taboo words, so these words are part of receptive vocabulary.

3) Some words are not produced because learners lack knowledge of how to form a word or use the word (knowledge about morphology or rules).

4) With the support of context, only partial knowledge is required to understand the word, so there tend to be more receptive words than productive words.

The first three explanations support what Corson explained about receptive and productive vocabularies. The fourth explanation is also mentioned by Crow (1986, p. 243), "A vast body of information is required for productive control of vocabulary items, while a much smaller body of knowledge is required for general receptive control".

Therefore, dividing words in one's lexicon based on motivation should be fine in natural settings, i.e., people can choose what words to use depending on their knowledge, their willingness and on many extralinguistic factors as Corson mentions. However, I would like to argue that in teaching and learning settings using motivation to distinguish receptive vocabulary from productive vocabulary would be problematic. If those unmotivated words appear in a controlled productive task or test, learners are forced to use them, and they use them correctly, it is not clear whether these words should be grouped in unmotivated, motivated, receptive, or productive vocabulary.

The second dimension proposed by Meara (1996a; 1996b; 1996c) that should be explored together with the dimension of 'size' is 'organisation' – the depth of vocabulary knowledge. 'Organisation' refers to how rich lexical structure is linked (Meara, 1996c). It is connected to lexical breadth because when vocabulary size is larger, more links will be built among words in the lexicon (Meara, 1996a; Milton and Fitzpatrick, 2014).

For vocabulary assessment, Meara (1996a, p. 11) suggests, "Ideally, what we need is a characteristic that is able to distinguish between someone who has a large vocabulary because they have just learned a long list of words, and someone whose vocabulary is more structured than that".

Meara (1996a) points out that the aspect of word associations is related to a large number of words and to how they are structured; therefore, word association is the property of vocabulary as a whole. He mentions that vocabulary depth does not require a "detailed understanding of the way individual lexical items function" but

rather how words are related to one another (p.15). In addition to the aspects of word associations, Meara (1997) argues that it is also related to collocational links as well as grammatical links.

Meara (1996a) adds that lexical organisation might be an important dimension of vocabulary knowledge because it is directly linked to learners' language performance. There is evidence that word structure in L2 learners' mental lexicon is weaker than that in native speakers' mental lexicon. That is, native speakers can make connections to a variety of words with ease, while L2 learners can link a word to fewer words (Meara, 1996a).

In vocabulary assessment, one example of vocabulary depth tests is Read's (1993) Word Associate Test (WAT). This test asks a test taker to choose the words from the choices that have the same or related meanings with that of the target words (testing the knowledge of associations) and to choose the words that are always used with the target words (testing the knowledge of collocations). Since this test involves the ability to recognise the answer, it is also viewed as a receptive vocabulary test.

However, scholars define the term 'depth' differently. While Meara (1996a; 1996b; 1996c) and Read (1993) define 'vocabulary depth' as the lexical organisation or the knowledge of word associations and collocations, Schmitt (2010) and Zhong (2016) define 'vocabulary depth' as the knowledge of many aspects such as the aspects of written form, form and meaning, word parts, grammatical functions, word associations, and collocations. Based on a developmental approach, Wesche and Paribakht (1996) devised a Vocabulary Knowledge Scale, a vocabulary depth test. The depth they define include the progression of how well a word is known, ranging from the ability to recognise a word to having some idea of what it means, and to being able to use the word.

The third dimension of lexical competence is 'automaticity'. It is proposed later in Meara's (1996b; 1996c) paper because it can also provide bigger picture of learners' vocabulary knowledge, not merely focusing on features of individual words. Meara (1996c, p. 5) states, "we would need to be able to specify how automatically the items in a lexicon could be accessed". This dimension is called 'fluency' by Milton and Fitzpatrick (2014, p. 7), Kroll et al. (2002) and Færch et al. (1984).

Færch et al. (1984) explain the term 'fluency' in their paper as an aspect of communicative competence. They state that 'fluency' in all skills is almost the same, but the modes and time involved are different. They used a spoken communication as an example for simplicity in explaining. "Essentially fluency involves the capacity to be able to put what one wants to say into words with ease. This is the broad sense in which we shall refer to fluency… Fluency has to do with how well the speaker is

able to link the units of speech together smoothly and without strain. These units can belong to any of the following levels of the linguistic system: individual sounds, words and their combination, semantic units" (Færch et al., 1984, p. 143)

In vocabulary assessment, "[t]his three-dimensional framework makes it possible to distinguish between learners who know lots of words and lots about them but struggle to use them (declarative knowledge), and learners who can quickly and naturally activate this knowledge for communication (procedural knowledge)" (Milton and Fitzpatrick, 2014, p. 7). 'Fluency' can be tested together with both receptive and productive vocabulary knowledge. For example, Kroll et al.'s (2002) psycholinguistic study combine the dimension 'fluency' with learners' receptive knowledge when examining the effect of L1 interference on time the participants spent on translating words because it is widely believed in the field of psycholinguistics that L1 interference is an obstacle for fluency in language use. However, as the aim of this study is not about fluency, the previous studies tapping this dimension are not reviewed in Chapter 3.

To conclude, Meara suggested that researchers examine a bigger picture of learners' lexical competence by exploring vocabulary 'size', 'organisation' and 'fluency'. However, his three-component model does not show the idea of vocabulary acquisition or how words are stored in our mental lexicon.

2.2.4 The global model within the developmental approach

The model of vocabulary knowledge in this section draws on the dimensions proposed by Meara (1996a) and the notions of receptive and productive vocabulary knowledge. Henriksen (1999, p. 304) states that she "see[s] a need for being more specific and suggest[s] three dimensions as a balanced position between the global and the separate trait view". Her model emphasises vocabulary acquisition or stages showing how much words are familiar or acquired. It is built up from the idea that vocabulary knowledge is not an all-or-nothing phenomenon but rather a movement along a continuum. Henriksen (1999) creates the model for lexical development based on how previous research studies have explored vocabulary knowledge and on the test formats utilised.

The first dimension proposed by Henriksen (1999) is the partial-to-precise dimension, which corresponds with the dimension of size proposed by Meara (1996a). She notices that studies of vocabulary size often use different test tasks; these tasks deal with different levels of knowledge. Some studies of vocabulary size use test tasks such as matching L1 to L2 words, choosing the correct meaning of a

target word from choices, and translating L2 to L1. These tasks require knowledge of form and meaning, which she considers precise comprehension.

Some studies of vocabulary size examine only the ability to recognise words by using a checklist test format (e.g., Meara, 1996a). Meara (1996a) argues that when a learner selects the word they know, besides the ability to recognise the word form, the student also needs to have knowledge of the word meaning as well. However, as the test format itself apparently focuses on word form recognition, Henriksen thinks that this task should be clearly distinguished from the ones mentioned above. Knowledge of form and meaning should require precise comprehension while the ability to recognise word form needs only partial knowledge. Researchers can tap different levels of understanding by including many test tasks as a test battery in their studies, such as a word recognition task and a multiple choice format. Mostly, the test formats focusing on the aspects of spoken/written form and the aspect of form and meaning can be considered as the research on partial-precise dimension or on the vocabulary breadth dimension.

Henriksen proposes the dimension of partial to precise dimension so that it can cover the aims of these studies or range of test formats. She adds that as vocabulary knowledge is not an all-or-nothing phenomenon learners can have vague knowledge of word meaning (partial knowledge) and this knowledge does not need to be fully developed to full comprehension (precise knowledge); understanding will gradually change depending on experience of the world and of the language. She suggests that it is possible to tap levels of understanding by combining test formats and track the development longitudinally.

However, scholars define the investigation of many aspects of vocabulary knowledge differently. Henriksen (1999) and Read (1998) also considers it as the investigation of learners' partial-precise knowledge. Read (1998) examined the knowledge of spoken form, meaning, part of speech, associated words, and derivatives by using an interview procedure. Schmitt (2010) and Zhong (2016) examined these aspects but consider their work as exploring the vocabulary depth, not the partial-to-precise dimension.

The second dimension of vocabulary development proposed by Henriksen (1999) is depth. She proposes this dimension by summarising and combining the ideas of other scholars (e.g., Cronbach, 1942; Dolch and Leeds, 1953; Nation, 1990; Richards, 1976) and by depending on the features of some tests such as Meara's (1982) word association test. Mostly, this dimension involves the aspects of words associations and collocations.

Therefore, Henriksen's second dimension related to knowing different meanings of a word and its related words is similar to Meara's (1996a) second dimension 'depth'. She writes that the development of this dimension involves the development of the vocabulary size. That is, "while developing a general understanding of a word, the learner will primarily have to develop a link between sign and referent" (Henriksen, 1999, p. 312).

Henriksen proposes the dimension of receptive-productive vocabulary knowledge as the third dimension of vocabulary knowledge, based on the test types available such as a multiple-choice test of vocabulary (receptive vocabulary test) and picture-description task (productive vocabulary test). In vocabulary assessment, Henriksen suggests that many test tasks be combined in a test battery to track vocabulary development along the receptive-productive dimension, but these tasks need to focus on the same target words.

Regarding the relationship between receptive and productive vocabularies, Henriksen points out that most researchers believe that receptive vocabulary and productive vocabulary are two separate groups that compete for threshold levels of language proficiency (e.g., Fan, 2000; Meara, 1996a; 1996c); receptive vocabulary includes the words that one is able to comprehend, and productive vocabulary is the words one can use. However, what she believes is different. She writes:

> "I would hesitate to draw a sharp and well-defined line between receptive and productive vocabularies in order to emphasize that I am not dealing with a dichotomy (i.e., completely distinct sets of vocabularies) but am operating on a continuum … Furthermore, it is quite unclear where one would have to draw such a dividing line or threshold between reception and production" (Henriksen, 1999, p. 313).

Similarly, Melka (1982; 1997) also states that receptive-productive vocabulary knowledge is a sort of continuum, or the degrees of familiarity with some aspects becoming productive while some still at receptive state. Also based on test types, she says that measuring vocabulary knowledge by using different test formats like a checklist, a multiple choice, and a defining or illustrating task should be seen as testing "various criteria of knowledge".

> It goes "from the most obvious and intuitive (checking words as known) to the more "reliable" (defining or illustrating and item). … there is a sort of continuum, a line of knowledge going from less familiar to more familiar; paralleling this there are procedures or techniques to test the degrees of familiarity: multiple choice, synonym, illustration, translation etc. This shows

once again that the distinction made between reception and production is simply a pro forma distinction and thus unsatisfactory" (Melka, 1982, p. 20).

Melka furthers:

"[T]he notion of familiarity (or knowledge) of a word could be represented as being a line, a continuum starting, roughly, with the first stage of recognition, passing through various intermediary points and finishing near productive knowledge, with productive knowledge itself composed of several stages or phases (e.g., knowing a range of meanings for a word or the knowing of collocations or idioms could represent a more advanced stage of production compared to simply knowing a single meaning of a word which has many). The boundaries of receptive knowledge and productive knowledge are variable: their limits for a particular subject are depending on the form of a test or circumstances that are a part of the moment of an experiment" (Melka, 1982, p. 21).

Færch et al. (1984) also suggest that one end of the continuum of vocabulary knowledge represents the "ability to make sense of a word" and the other end represents the "ability to activate the word automatically for productive purposes" and words that are acquired into the lexicon will scatter at different points along this cline (p. 100).

How Melka (1982) and Færch et al. (1984) define the distinction between receptive and productive vocabulary knowledge seems plausible. However, it is still not clear what should be included as intermediary points along the continuum. Besides, what Melka exemplifies as stages in productive knowledge are viewed as receptive knowledge in Nation's taxonomy as they emphasize the term "knowing", not "the ability to use".

Despite the unclearness of the intermediary points, some researchers agree that vocabulary knowledge includes many stages or degrees of knowledge. In vocabulary assessment, researchers who adopt this idea – vocabulary knowledge is a sort of continuum – think that receptive vocabulary and productive vocabulary are not competitive or grow with different rates. They then do not compare the size of the two but explore vocabulary development in terms of how a word is acquired or moves from one end to the other end (e.g., Paribakht and Wesche, 1993). According to Palmberg (1987, p. 203), vocabulary development can be explored qualitatively, which is to answer the questions like "how far individual words move along the continuum".

One vocabulary test in this model that has been widely used is Paribakht and Wesche's Vocabulary Knowledge Scale (VKS) (e.g., Min, 2008; Tahmasbi and Farvardin, 2017). The VKS tracks the knowledge of a word by starting from the stage of no knowledge, word recognition, knowing its meaning, being able to define it, and being able to use it in a sentence level.

As the theory of the relationship between receptive and productive vocabulary knowledge is still debateable, it is usual that the idea about the continuum will be argued against, in the same way to the idea of the receptive-productive distinction. Even a researcher who tends to prefer the idea of the separation between receptive vocabulary from productive vocabulary like Meara (1996c) still states that it is not clear what varies along the cline and that the degrees of passiveness and activeness are not clear, either.

Explaining vocabulary acquisition, Meara (1996c) suggests that vocabulary knowledge should be more unstable than a continuum or it should not be in a linear progression. Otherwise, a word should not be forgotten once it is learnt (Melka, 1982). Consequently, he proposes a multi-stage model, which include five discrete stages of vocabulary development and allow transitions to move from any one state to any higher state and to move back to lower state if such word is fully learnt and then forgotten.
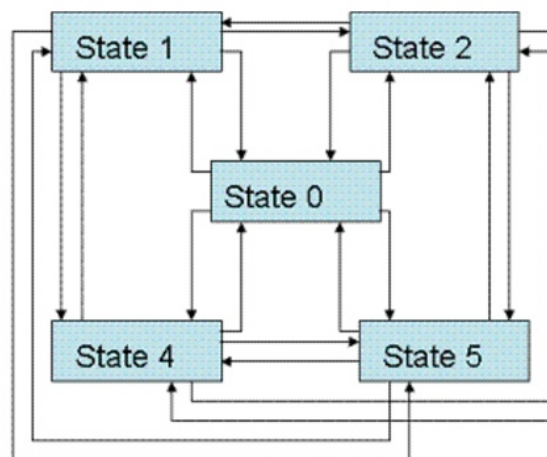
Figure 2.2: Meara's (1996c) multi-state model of vocabulary acquisition

Meara (1996c) suggests this model by adapting the states in Wesche and Paribakht's (1996) VKS. In his paper, he excluded State 3 because it is repetitive to State 4 and mentions that, in this model, we allow words to move from State 0 (no

knowledge) to State 5 and reversely move back to lower states. His model has not been developed further into a vocabulary test, though.

Meara criticises the VKS, stating that it is not a good tool to track acquisition states because the VKS format is a fixed linear progression and does not allow for shifts among acquisition states of vocabulary as a whole at a particular set of circumstances as his model presents. Nevertheless, I still see that the VKS can be used to track the increase and the decrease of vocabulary acquisition between two periods of time, and it can include many words (such as groups of frequency-based words) for the overall picture of vocabulary acquisition if that is the main purpose of a study. However, the vocabulary tests are reviewed in Chapter 4 before I can select the one suitable for my study.

In conclusion, vocabulary knowledge has been characterised in a number of ways, with both the separate traits and global models. Receptive and productive vocabulary knowledge has been considered in terms of different abilities, mostly dependent on the aspect focused. The nature and relationship between receptive and productive vocabularies and how they grow are still debatable, whether they are the same group of words moving on a cline and going through different stages, or separate groups of words activated by some factors and selected for use (whether they are two separate groups or one group as a subset of the other). This uncertainty still exists, and several details need to be explored.

2.3   Research questions

The research questions of the present study are based on the research aims and on what I have reviewed in Section 2.2.

Firstly, I aimed to explore the nature of receptive and productive vocabularies, and according to Laufer (1998), productive vocabulary should be divided into controlled productive and free productive vocabularies. My research questions cover these three groups of vocabulary. I aim to explore vocabulary knowledge in general, not the knowledge of some specific words. I then need to review in Chapter 3 on which group of words to be the focus. Also, even though vocabulary knowledge includes many aspects, the focus is on vocabulary knowledge in general, not on any certain aspects. The research questions then cover the three major categories in Nation's (2013) taxonomy: form, meaning, and use, instead. The knowledge of form and meaning is related to receptive vocabulary knowledge, while the knowledge of form, meaning, and use is related to two groups of productive vocabulary knowledge.

The relationship among receptive, controlled productive, and free productive vocabularies is still uncertain, whether they are separate groups of words in mental lexicon, or they are stored as one group of words and change their status along a continuum. I then did not take side at this stage, and I aimed to explore receptive and productive vocabulary knowledge in both ways (within the dichotomous approach and the developmental approach). I needed to review what previous studies have done and what measures they have used so that I could sort out the research design and the measures that were the most suitable for my study.

I also aimed to explore the receptive, controlled productive, and free productive vocabularies at different years of study, so the research questions are initially as follows.

Research question 1: What is the participants' receptive vocabulary knowledge?

Sub-question 1: What is the participants' receptive vocabulary knowledge at different years of study?

Research question 2: What is the participants' controlled vocabulary knowledge?

Sub-question 1: What is the participants' controlled productive vocabulary knowledge at different years of study?

Research question 3: What is the participants' free productive vocabulary knowledge?

Sub-question 1: What is the participants' free productive vocabulary knowledge at different years of study?

Research question 4: What is the relationship among the receptive, controlled productive, and free productive vocabulary knowledge?

Sub-question 1: What is the relationship among them at different years of study

To answer these four questions, I reviewed what previous studies have done when they investigated these three groups of vocabulary in Chapter 3 so that I had some guidelines on which group of words to be focused, which research design to adopt, which tests to be used, and how to score participants' answers. After reviewing the literature in Chapter 3, I revised the research questions as shown in Section 4.2.

Chapter 3 Literature review 2

Research on receptive and productive vocabulary knowledge

3.1 Introduction

Chapter 2 examines how researchers view the relationship between receptive and productive vocabulary, a continuum or a dichotomy. This chapter reviews how researchers have explored receptive and productive vocabulary knowledge in relation to these two views. Since my study focuses on reading and writing skills, receptive and productive vocabulary knowledge will be limited to the ability to comprehend words when reading and writing.

In addition, I will be reviewing studies related to English as a second/ foreign language rather than the studies on vocabulary size and growth of English as L1 or that of other languages. Although my study focuses on higher education, I will cover studies conducted with the participants at primary or secondary levels as their findings and the vocabulary development trends and the relationship between receptive and productive vocabularies might inform my study.

This chapter serves as the rationale for the methodology in my study. The review covers what types of vocabulary knowledge researchers have explored and the most widely-used research methods. The methods include the number of participants involved, which tests researchers have used, which frequency levels and which groups of words they have tested, and the research designs (a longitudinal study or a cross-sectional study). I hope that this will lead to more methodological insights and the findings in my study might shed some light on the relationship of EFL learners' receptive and productive vocabulary knowledge.

3.2 Receptive and productive vocabulary assessment

3.2.1 Research on receptive vocabulary

Checklist or yes/no test format

The researchers who assume that being able to recognise written forms means having receptive knowledge of such words often use a checklist test format to measure receptive vocabulary size (e.g., Al-Masrai, 2009; Al-Masrai and Milton, 2012; Barrow et al., 1999; Meara and Jones, 1990; Sahiruddin, 2008; Ward, 2009b; Yunus et al., 2016) and receptive vocabulary growth (e.g., Al-Masrai, 2009; Al-Masrai and Milton, 2012; Sungprakul, 2016). Two well-known vocabulary tests with this format are the Eurocentres Vocabulary Size Test (EVST) by Meara and Jones (1990) and the XK_Lex by Meara and Milton (2003) and by Al-Masrai (2009).

The first version of the EVST was devised by Meara and Buxton (1987). IT can be used as a placement test and a vocabulary size test (Meara, 1990b). The EVST is a computerised yes/no test drawing 100 words from ten 1,000-word-family levels from a general word list, developed by Thorndike and Lorge's (1944) list.

> "The checklist format of the EVST reduces task demands to an absolute minimum by requiring the test-takers just to respond, 'yes' or 'no', to say whether they know the word, to a series of words [presented in isolation] appearing on the computer screen" (Read, 2000, p. 147).

The EVST uses pseudo words and was deigned to adjust words.

> "However, when the test taker has responded 'Yes' to non-words, a more complex calculation is required. The calculation is based on statistics from signal detection theory...in a military context" (Read, 2000, p. 130).

That is, if a test taker knows fewer words at a certain level, it stops showing words in the next (infrequent) level. Raw scores should be multiplied by 100 to estimate receptive vocabulary size; it can represent the most 10,000 frequent words.

Meara and Jones (1988, p. 4) write:

> "if the testees score highly on this band, then they are tested on the next band, and this process continues until performance drops below a preset threshold. At this point, the program works out a rough estimate of how many words we think each testee knows, and tests a further 50 words form the appropriate frequency bands. … So, suppose our testee scores 100% on bands 1-4 but only 20% on band five, the program reckons that they know somewhere between four and five thousand words, and does its detailed testing on band four. The detailed testing phase actually tests one word in twenty at the appropriate level."

The XK_Lex sampled 100 words from two general word lists by Nation (1984) and Kilgarriff (2006). There are two pseudo words for each level. Not adjusted like the EVST, the XK_Lex gives a larger receptive vocabulary size than the EVST because

48

some learners know some infrequent words too. Raw scores should be multiplied by 100. If the test-taker chooses a pseudo word as a known word, 500 points are deducted from the total scores. The test is available in both computer-based and paper-based versions (Al-Masrai, 2009, p. 33).

A sample of a checklist test item is shown in Figure 3.1.

```
1.    adviser


     [ Yes ]        [ No ]

```

Figure 3.1: An EVST test item

Ward (2009b) used a yes/no checklist test to measure the size of general vocabulary and engineering vocabulary of engineering students in Thailand to check whether these students have sufficient receptive vocabulary knowledge to cope with the reading tasks in an engineering class. The students were supposed to read engineering textbooks in English. He first developed an Engineering Word List (EWL) (Ward, 2009a), sampled the words from the list and devised a new EWL vocabulary size test in a yes/no test format (Ward, 2009b). This test format was shown to be valid and reliable by Meara and Buxton (1987) and Meara and Jones (1988), and Meara (1990a). It takes little effort from the test takers. Ward writes the test was suitable for his participants, engineering students who were low-proficiency in English. He asked the participants to tick words for which they knew at least one meaning. To check if test takers really know the target words, he produced a new test with some non-words.

In the main study, there were 250 participants, and the tests were administered in a regular class. 75% of the participants sat the test of engineering words, the EWL checklist test, the focus of the study and 25% of them sat the test of general words sampled from West's (1953) General Service List (GSL). There were 120 real words and 60 non-words for each test, and raw scores were converted into percentages. Choosing a non-word, raw scores would then be deducted; Ward did not report how. The 120 real words in the EWL test were sampled from 2,000 words from the EWL, providing "95% percent coverage of a 1,000,000 word foundation engineering corpus (Ward, 1999)" (Ward, 2009b, p. 296). The 120 real words in the GSL checklist test

represent the most 2,000 frequent words in the GSL and cover around 79-80% of a running words in most academic textbooks (Nation and Waring, 1997).

The findings showed that on average the participants had relatively small receptive vocabulary size, of both engineering words and the general words. The subjects knew less than half of the words from the Engineering Word List (953 out of 2,000 word families) and an estimate of 1,200 out of the 2,000 word families of the GSL. When he converted this vocabulary size into the percentage of text coverage to see how much the learners were able to comprehend the textbooks, the findings showed that "most students appear clearly to be operating way below the 95% lexical coverage level" (p. 298). Ward concluded that the lack of sufficient receptive vocabulary knowledge of engineering words and general words can impact students' performance when reading English texts in engineering class. His teaching experience at this university and the students' low mean scores (35.6 out of 100) on the English section in the entrance exam from the year 2009 (Ward, 2009a) are possible evidence for this. The results also showed that the ideal curriculum goals set by the university were difficult to achieve.

Ward (2009a) directly benefited the teaching and learning in his context, but he sampled words from specially designed word lists, the EWL. This makes it difficult to generalise the findings of his study. However, he also measured the participants' general vocabulary size by sampling words from the GSL, and the result in this part can then be compared to that of others' studies. However, measuring only the knowledge of the most 2,000 frequent words might not provide the participants' overall vocabulary knowledge as they might know words out of this frequency level. Hence, sources of general words might be more useful for my research as the focus is on vocabulary knowledge in general, but words from other frequency levels would be considered.

Other researchers have used more general words lists. Al-Masrai and Milton (2012) used two yes/no tests to measure learners' receptive vocabulary size: EVST designed by Meara and Jones (1990) and XK_Lex by Al-Masrai (2009). The word lists used are explained earlier. They asked the participants to sit two versions of XK_Lex and then the EVST, to estimate their knowledge of words out of 10,000 word families. This study was also distinctive in that they investigated the learners' receptive vocabulary growth cross-sectionally by comparing the receptive vocabulary sizes of two groups of university students in Saudi Arabia (55 first-year students and 37 fifth-year students, majoring in English).

The findings showed that the fifth-year group had larger receptive vocabulary size than the first-year group. The first-year students had a receptive vocabulary of about

2,000-3,000 word families, and the fifth-year students had a receptive vocabulary of around 5,000 word families. According to Laufer (1998), 5,000 word families can cover 95% of running words in authentic texts and the fifth-year students had enough vocabulary to comprehend reading texts but were not fluent readers (Al-Masrai and Milton, 2012).

In 2016, Sungprakul also measured the size and growth of the receptive vocabulary of English-major students cross-sectionally by using the XK_Lex (Al-Masrai, 2009). The participants in his study were 40 students purposively sampled from all the four year cohorts at Silpakorn University. He aimed to examine the receptive vocabulary growth cross-sectionally.

Compared with Al-Masrai and Milton's (2012) study, the findings showed that the receptive vocabulary size of Thai university students who majored in English was larger. While the first-year English majors in Al-Masrai and Milton's (2012) study had a receptive vocabulary size of around 2,000-3,000 word families, the first-year students in Sungprakul's (2016) study had a receptive knowledge of around 5,800-6,000 word families even though they were EFL learners at the beginning of university level in the same major and got tested with the same vocabulary size test. Participants in higher years of study had a larger vocabulary. The average receptive vocabulary size of the second-year students was around 6,300-6,600 word families, the third-year students knew around 6,500-7,100 word families, and the fourth-year students knew around 7,300-7,400 word families.

Al-Masrai and Milton (2012) and Sungprakul (2016) used a test that samples words from more general word lists and represents a large group of words – 10,000 words. They did not assume that the participants' knowledge was limited to only 2,000 words as Ward (2009b) did. However, the result, presented as a single figure of vocabulary size, might not be beneficial for pedagogical purposes since it does not identify which groups of words are needed to be taught or learned more.

Al-Masrai and Milton (2012) and Sungprakul (2016) also investigated receptive vocabulary growth by adopting a cross-sectional research design. "A cross-sectional study is one that produces a 'snapshot' of a population at a particular point in time" (Cohen et al., 2007, p. 213) . It seems more practical for studies with limited time. The researchers did not need to spend three or four years to collect all the data and did not experience the attrition of the participants during the project, either. However, this research design cannot be used to track vocabulary growth of a particular subject, so the data on vocabulary growth can only be inferred from the difference in vocabulary scores between groups of participants from different years of study. For practicality, this research design appeared beneficial for the present study.

In addition, while the number of participants of each year cohort in the study by Al-Masrai and Milton (2012) exceeds thirty, that in Sungprakul's study does not; there were only around ten students per year group. According to Cohen et al. (2007), "a sample size of thirty is held by many to be the minimum number of cases if researchers plan to use some form of statistical analysis on their data, though this is a very small number and we would advise very considerably more" (p. 101). Therefore, ten participants per year cohort as in Sungprakul's study might be too small for the statistical analysis.

In terms of the test format, Meara (1996a) and Al-Masrai (2009) claim that checklist test format is the simplest format of vocabular size tests that allows a research study to include many words and allows a test taker to complete the test in a short time. It is also easy for the teachers or researchers to administer and to score learners' performances; rating learners' answers in this test is straightforward and it is easy to interpret the scores (Meara, 1996a).

However, there are some disadvantages of the test format as well. First, the yes/no test format cannot guarantee whether the learner really knows and understands the word. If the test takers tick a real word without knowing it, there is no task for them to verify their knowledge. Therefore, the test allows guessing and 50% chance of correct answers. some test takers would have response bias, that is, "over-willing to say 'yes' to the "imaginary words" (Meara, 1996a, p. 9). Including non-words or pseudo words might not be able to solve the problem of guessing. Cameron (2002) trialled this test format (Meara, 1992), together with another vocabulary test before measuring her secondary-level learners' receptive vocabulary size and found that "the inclusion of non-words in the Yes/No test produced unreliable results" (Cameron, 2002, p. 145).

Cameron (2002, p. 159) writes:

> "The non-words in the Yes/No test created problems for all students, but particularly the EAL students, many of whom had scores heavily reduced by the number of non-words they checked as 'known'. This was particularly evident at the 1K and 2K levels."

Second, the computerised format (the EVST) does not seem to accurately estimate learners' receptive vocabulary size. According to Al-Masrai and Milton (2012, p. 17), the "EVST assumes vocabulary is learned in strict frequency order and therefore discounts knowledge of much infrequent vocabulary. It assumes no knowledge at all in these areas, once vocabulary scores in a frequency level drop below the highest of levels". Al-Masrai and Milton (2012) used these two tests to measure Saudi university students' receptive vocabulary and found that the estimates of receptive

vocabulary size given by the EVST was smaller than that given by the XK_Lex. It showed that learners might know some infrequent words and automatically adjusting words to test in the next level does not accurately estimate a test taker's receptive vocabulary size.

Therefore, this test format might not be the most appropriate test format to be used in this study in order to investigate learners' receptive vocabulary since it allows high percentage of guessing and it cannot check whether the test taker knows the meaning of the test words.

Matching format

Many research studies used a vocabulary test with a matching format. Among many others, studies by Laufer and Paribakht (2016), Waring (1997a), Zhang and Lu (2013) are some examples.

A popular receptive vocabulary test with a matching format is the VLT (Beglar and Hunt, 1999; Nation, 1983; 1990; Schmitt et al., 2001). The VLT measures the knowledge of word form and meaning, like the EVST. It does not estimate one figure of a learner's receptive vocabulary size but indicates how much vocabulary at different frequency levels is known. It is used to "determined whether learners had gained mastery of high-, mid-, and low-frequency words as well as words that are common in academic discourse" (Stoeckel et al., 2020, p. 4).

Based on the GSL (West, 1953) and Thorndike and Lorge's (1944) list, it includes general vocabulary from four frequency levels (the 2,000-word level, the 3,000-word level, the 5,000-word level, and the 10,000-word level) and one section of academic words from the UWL (Xue and Nation, 1984). There are 18 items for each section in the original version by Nation (1983; 1990), so there are 90 items altogether. These eighteen items represent 1,000 word families of each level, except at the 2,000-level and at the academic word level. The eighteen items at the 2,000-level represent the most frequent 2,000 word families, while the eighteen items in the UWL section represent 836 word families. Hence, the scores should be calculated based on these numbers.

Later, Schmitt et al. (2001) adapted the VLT by increasing the number of items in each level to 30 items and words in the academic word level were sampled from a more recent list at the time – the Academic Word List (AWL) by Coxhead (2000). The cut-off point for mastery of each word level is 80%. That is to say, a test-taker who can score 24 word families or more out of 30 word families for each word level in the VLT is taken to master that word level (Schmitt et al., 2001).

This test can be used as a diagnostic test rather than a proficiency test as it addresses the receptive vocabulary size of different groups of words (Nation, 1983; 1990; 2001; 2013). Nation and Beglar (2007) say that the VLT is used to "determine whether learners need to focus on high-frequency words, academic words, or low frequency words" (p. 10). However, the test results can be calculated to roughly find the vocabulary size estimates. Learners with only small amount of knowledge of the target words will be able to complete the task; the target words are isolated (no contextual sentence) (Nation and Beglar, 2007). The choices are not semantically related, so it is easier for beginners to distinguish the correct choice from the incorrect ones (Nation and Beglar, 2007).

Answers in a matching-format test need to be graded with a dichotomous scoring method (correct/incorrect). The scores can be converted into the estimate of vocabulary size as suggested by Nation (1990), "Vocabulary size = N correct answers multiplied by total N words in dictionary (the relevant word list) divided by N items in test" (p.78).

The VLT would appear to be a valid and consistent vocabulary test because it measures what it is designed to measure (Beglar and Hunt, 1999; Laufer, 1998; Read, 2000). "Many studies have used the VLT in any one of its versions to test the receptive vocabulary size of subjects either for descriptive, comparative or correlational purposes" (Gallego and Llach, 2009, p. 119). Figure 3.2 is the sample of the VLT.

Instruction: You must choose the right word to go with each meaning. Write the number of that word next to its meaning.

1. concrete
2. era   _____ circular shape
3. fiber   _____ top of a mountain
4. hip   _____ a long period of time
5. loop
6. summit

Figure 3.2: Sample of the VLT

Li and MacGregor (2010) investigated the vocabulary knowledge levels of 128 Chinese university students in Hong Kong by using the VLT (Schmitt et al., 2001). The results from the test do not give a single figure of vocabulary size like the VST does, though. The VLT rather provides estimates of receptive vocabulary knowledge

at four different frequency levels and at academic word level. The participants majored in English and were selected because they tended to use a large or wide vocabulary for their study and work. There were two groups of them: 81 first-year students (BA group) and 47 part-time MA students (MA group); the majority of those in the MA group were working as English teachers at the time.

The results showed that both groups of the participants had high scores at the high-frequency levels and lower scores when the test words became less frequent. They received the highest scores at the 2,000-word level (97-99%) and the scores at the 3,000-word level and at the academic word level were roughly equal (93-96%); the average score of the academic words was between the 3,000- and the 5,000-word levels. They knew 77-79% at the 5,000-level and had the lowest scores at the 10,000-word level (30-34%). This showed that frequency levels had impacts on their vocabulary knowledge. The researchers also investigated the growth cross-sectionally, and the results showed that there were no significant differences of the scores between the two groups, "demonstrating that more years of university education and work experience do not lead to the acquisition of the low-frequency VLT words by Hong Kong learners." (p. 247).

Testing the knowledge of many groups of words based on different frequency levels seems to fit the aim of my study because it can provide a big picture of learners' overall vocabulary knowledge. Moreover, academic words are worth testing because university students need to encounter a lot of academic words throughout their four-year learning. The estimates of learners' general words and academic words can also serve as a guide for both learners and teachers so that they knew which groups of words need more emphasis. Therefore, both general words and academic words are included into my research study. However, I review some more studies using a matching format before I decide whether I would use a matching format.

Unlike Li and MacGregor (2010), Alonso and Fontecha (2014) used the VLT (Schmitt et al., 2001) to track learners' vocabulary development longitudinally. The participants were adolescent and adult learners (intermediate level or CEFR B1 level) in a two-year English course at the Official School of Languages in Spain. The data was collected twice (in 2011 and in 2012). Only the 2,000-word level of the VLT was tested. The results showed that there was an increase in vocabulary knowledge after two years of instruction. In their first year, the participants' mean score was 77.77% (23.33 out of 30 points) in their first year and 82.93% out of 30 points in their second year. They mastered the vocabulary at the 2,000-word level when they were in the second year of the course. The researchers used the formula suggested by Nation (1990) to estimate the participants' vocabulary size; the participants knew

around 1560 word families in their first year and 1660 word families in the following year. This was a statistically significant increase.

However, Alonso and Fontecha tested only the 2,000-word level. This might be because the researchers only investigated whether there was an increase in knowing words or only to check the effectiveness of the language course, so the 2,000-word level was sufficient. They did not use the VLT as a diagnostic test, so there was no need to test many groups of words. For my study, testing only words at the 2,000-word level is not sufficient.

Testing the participants' knowledge of more groups of words longitudinally, Zhang and Lu (2013) examined the vocabulary growth of the 300 Chinese university students both in English major and non-English majors from their first year to their third year and to examine the relationship between their vocabulary knowledge and their vocabulary fluency. As vocabulary fluency is not the focus of the present study, I will not review this issue.

Zhang and Lu explored the participants' vocabulary development by using Schmitt's (2000) VLT. They also examined whether frequency levels had an impact on vocabulary knowledge, so they included all the five frequency levels of the test. They administered the test three times: Two weeks after their enrolment (T1), eleven months later (T2), and another eleven months later (T3).

The results showed significant differences among their vocabulary scores over the 22 months, and frequency levels had an influence on the participants' vocabulary knowledge and vocabulary growth. The participants' average vocabulary scores at the 5,000-level had the highest increases throughout the three times, from 38.43% (11.53 out of 30 points), to 58.90% and to 77.73%. All the increases were statistically significant with the p-value of less than 0.01. The vocabulary scores of the 2,000-level had the smallest increases, from 90.63% to 97.40% and 98.93%. This might have been because of the high scores at this level and there was not much room for improvement. The vocabulary scores of the AWL level (69.27%, 86.90%, and 95.13%) were between those at the 3,000-level (69.57%, 88.37%, and 93.13%) and the 5,000-level (38.43%, 58.90%, and 77.73%).

Testing a variety of groups of vocabulary, like the study by Zhang and Lu (2013), appears to be suitable for the aim of the present study as it covers many words and the results would not be too specific to only the 2,000-word level. It can give more insights about vocabulary knowledge in general.

Regarding the research design, exploring learners' vocabulary knowledge longitudinally could provide accurate results of how vocabulary knowledge of the

same groups of participants grows. Longitudinal studies "enable researchers to…highlight similarities, differences and changes over time in respect of one or more variables or participants (within and between participants)" (Cohen et al., 2007, p. 212). This research design does not seem to be practical for the current research because of the limitation of time; the duration of data collection cannot be lengthened to three or four years.

Conducting a three-year longitudinal study, Gallego and Llach (2009) measured and tracked the vocabulary development of one group of primary students in Spain from when they were in 4th grade to when they were in 7th grade (1st year at their secondary level). As mentioned in the introduction of the chapter, I review this study even though their participants were not university students because I can learn how they conducted the research, and I might be able to relate the patterns of how their participants' vocabulary developed to those of the participants in my study.

Gallego and Llach's study initially recruited 283 students, but due to participant attrition, only the data of those sitting the VLT (Schmitt et al., 2001) at all the four data gathering moments were analysed in the study. Since they were all primary school students, only the 2,000-word level of the VLT was selected. The results showed that their vocabulary size increased constantly each year. Their average scores out of 30 points at the 2,000-word level from 4th grade to 7th grade were 18.03%, 25.43%, 31.57%, and 40.87%, respectively. When they converted these figures into the number of word families they had acquired, the vocabulary sizes were 361, 509, 631, and 817 word families.

However, the participants gained roughly the same but rather small number of words every year (approximately 120-180 word families per year), the increases were then not statistically significant. Moreover, even when they were at their secondary level, their vocabulary size did not reach half of the 2,000-word level.

When considering only their maximum scores, the researchers found that they gained lower maximum scores when they were in 7th grade than they got in the last year (two points lower). The researchers then qualitatively examined the data and found that one student was responsible for this maximum score and assumed that this decrease in score might be relevant to extra-linguistic factors such as "fatigue, disinterest, or student's entering adolescence" (p. 120). This finding can be evidence of the increase and decrease of someone's vocabulary knowledge. Nevertheless, the overall trends of receptive vocabulary size so far seem to increase when learners move to higher education levels.

Webb and Chang (2012) also longitudinally examined the growth of receptive vocabulary of the Taiwanese learners over a five year period from high-school to a

university level at one vocational senior high school. They used the bilingual version of the VLT. They adopted the test items at the 1,000- and the 2,000-word levels from the bilingual version by Nation and Wang Ming-tzu (1999) and translated the test items at the 3,000-, the 5,000-word levels and the academic word level from the monolingual VLT (Schmitt et al., 2001). Using the bilingual version made it easier for the participants to understand all the choices. They acknowledged that the test used in their study involves only the knowledge of form and meaning or the ability to recognise form and meaning, not word production. Also, the test examines the knowledge of words at some frequency levels. If the study aim is to examine the knowledge of words at the 4,000-word level or other levels, the VST (Nation and Beglar, 2007) was more appropriate.

At the beginning of the study, there were 222 participants but at the last year only 166 participants did all the five tests (the same test but administered five times annually). Therefore, only the data from the 166 participants was analysed to estimate the number of words learnt over four years of study, and to what extent the high-frequency words and the low-frequency words were learnt by the participants. The participants were divided into three groups (Groups A, B, and C); The class of Group A focused on English language learning and English literature, while the classes of the other two groups focused on general English language learning.

The findings showed that the participants from all the three groups knew more words when they moved to higher years of study, with Group A got the highest scores or acquired the most words over the four years. The highest gain was for Group A between year 1 to year 2 (12.91 points from the score in year 1, 21.24 out of 30 points). This represented 430 words approximately when one test item represented 33.3 words (Webb and Chang, 2012, p. 118). The least gain was for Group B between year 1 and year 4 (0.55 points from the score in year 1, 19.80 out of 30 points). The smallest gain was merely around 18 words. The gain of the academic words was similar to that at the 2,000-word level. In terms of mastery (a score of 26 or above), few students mastered the high-frequency words; overall, 47% of them mastered the 1,000-word level and 16% of them mastered the 2,000-word level in their final year of study. Also, a steady increase at the 3,000- and the 5,000-word levels was similar within each group. The findings showed inefficient English vocabulary learning. Webb and Chang then suggested that the curriculum have greater focus on explicit learning and learning high-frequency words so that it can help learners with comprehension of a reading text.

Some modifications for a replication of Webb and Chang's (2012) study suggested by Pellicer-Sánchez (2019) are 1) including instruments to elicit information about

out-of-class exposure because it also plays a role in vocabulary learning and 2) including a productive test so that receptive and productive vocabulary knowledge can be explored and compared. As many researchers consider that learning outside classroom is important for vocabulary learning, I included this in the questionnaire. As this is not the focus of my study, I initially hoped it might explain some unexpected findings if there were any.

Hence, vocabulary growth can be investigated cross-sectionally and longitudinally. Both ways might have their own advantages and disadvantages. Some disadvantages of the longitudinal research design I learned from Gallego and Llach's (2009) study is that participants dropping out in the middle of the study can happen and it needs a long time to complete a project. With time constraint, I decided to adopt a cross-sectional design.

Concerning the test format, even though the matching format is a simple test task that is easy to complete by test takers and easy to grade by researchers (using a dichotomous scoring method), one problem with this format is that it allows guessing (Read, 2000; Schmitt, 2010). Up to 17% of the mark can be accounted for by guessing rather than word knowledge – one correct answer out of six alternatives (Webb, 2008), and this may give a misleading impression of how many words a test-taker knows. Therefore, a matching format test like the VLT is still not the one to be used in the present study. However, the frequency levels used in the VLT seems suitable for my study as they can represent vocabulary as a whole and the academic words are necessary for learners at a university level as well.

Multiple-choice format

Some researchers who view receptive vocabulary knowledge as the knowledge of form and meaning also used a multiple-choice vocabulary test (e.g. Amin, 2020; Dizon and Tang, 2017; Hatano, 2008; Johnson et al., 2016; Kotchana and Tongpoon-Patanasorn, 2015; Nirattisai, 2014; Nirattisai and Chiramanee, 2014; Pathan et al., 2019; Shin et al., 2011; Šišková, 2016; Yunus et al., 2016; Zechmeister et al., 1995; Zhao and Macaro, 2016).

A widely-used vocabulary size test with a multiple format is the VST developed by Nation and Beglar (2007). The VST uses a meaning recognition format. It is a proficiency test used to measure how much English vocabulary learners know or how much vocabulary that learners have an idea of its meaning (Nation and Beglar, 2007). The test can be used as a diagnostic test as it can show which group of

vocabulary to focus on (Webb, 2021). It was not "developed and validated for the purpose of predicting reading comprehension" (Webb, 2021, p. 454).

The test requires a test taker to only choose the correct meaning for the test word embedded in a decontextualized sentence. This test format can also test many words at a time. "The frequency levels are based on word families occurring in [Nation's (2006)] British National Corpus according to Bauer and Nation's (1993) levels up to Level 6" (Nation, 2012, p. 2).

There are two versions: based on 14,000 word families and based on 20,000 word families. The 20,000-word version is suitable to test native speakers' receptive vocabulary size, and the 14000-word version is suitable to test EFL learner's receptive vocabulary size (Nation, 2013). It has been translated into many languages so that the test takers do not need more effort to understand the choice, in other words, the test will only test the knowledge of the test words, not the knowledge of the choices (Nation, 2013). Both the monolingual and the bilingual versions can be found on Victoria University of Wellington website (www.wgtn.ac.nz).

In the VST, there are ten items in each frequency level; these ten items represent 1,000 word families. Even though it is criticised by Stoeckel et al. (2020) that ten items for each frequency level is not sufficient to show test takers' vocabulary knowledge, Webb (2021, p. 458) argues that it will not be pedagogically practical because of long "time for test administration, time for grading, and test taker fatigue".

A dichotomous scoring system is also used to grade the test answers, and to interpret the VST results a test taker's score needs to be multiplied by 100 because each item in the test represents 100 word families (Nation and Beglar, 2007). The scores can show the estimates of a learner's receptive vocabulary size at each frequency level.

1. They **saw** it.
   a. cut
   b. waited for
   c. looked at
   d. started

Figure 3.3: Sample of the VST

Yunus et al. (2016) measured and compared written receptive vocabulary size of two groups of university students from two different contexts, Thailand (EFL) and

Malaysia (ESL). The participants were first-year English majors. The researchers measured their vocabulary size to investigate whether their vocabulary knowledge would meet the academic needs at the tertiary level. There were 80 Malaysian students and 86 Thai students. They were purposively selected as they were in their first year and studying in the first semester. The VST (Nation and Beglar, 2007) was used with these two groups of students in their classes at different locations and different time. Only the first 100 items of the test were used, and these 100 items represents the 10,000 frequent word families of English sampled from Nation's (2006) fourteen 1,000 BNC word lists. This is because the size of 10,000 word families was suitable for university students, and it was sufficient for academic purposes in university level (Laufer, 2013).

The findings show that the Malaysian students had more than twice larger receptive vocabulary than the Thai students; on average, the Malaysian students had the knowledge of 4,460 word families (ranging from 3,000 – 8,000 word families) and the Thai students had around 2,090 word families (ranging from 1,000 – 7,000 word families). The reasons why the Malaysian first-year students had larger receptive vocabulary might be because the English is frequently used in Malaysia and its status in Malaysia is the second language, so the students are exposed to English in their everyday life. Another reason put forward by the researchers is that the Malaysian students had learnt the English language longer than the Thai students. Despite that, the researchers commented that the Malaysian students' vocabulary size did not reach an optimal threshold level for EAP courses. However, they would be able to comprehend general texts as the vocabulary size of 5,000 word families can cover 95% of English texts (Hirsh and Nation, 1992; Schmitt et al., 2001).

More opportunity to use the language and the amount of time spent on learning seem to be important factors in vocabulary knowledge or vocabulary size. Therefore, to explore the vocabulary growth in my study, I expect that the learners at higher years of study tend to have larger vocabulary than those at lower years of study because they have learnt the English language longer and then tend to use the language more.

When comparing the vocabulary size of the first-year English-major students in Yunus et al. (2016) study (using the VST) and that in Sungprakul's (2016) study (using the XK_Lex), I found that the vocabulary sizes of these two groups of participants (first-year English-major students) are very different. The participants in Sungprakul's (2016) study had approximately three times larger receptive vocabulary size (around 6000 word families). The difference might probably be because of the number of participants or the difficulty of the test formats.

Firstly, the number of the participants per year cohort in Sungprakul's (2016) study might be too small and this might affect the statistics. Cohen et al. (2017) suggest that there be at least 30 participants per variable. Therefore, I planned to sample around 30 participants per year group for the study validity.

Secondly, these two studies used different test formats – a checklist format (the XK_Lex) and a multiple-choice format (the VST – the monolingual version). The multiple-choice format might be more demanding than the checklist test. It requires test-takers to read the sentential context and the choices in English. This might have affected the participants' lower scores.

However, none of these test formats will be used in my study because the chance of guessing. Despite a correct choice is chosen by a test taker, there are still 25% chance of guessing when using a multiple-choice format (one correct answer out of four choices). Some researchers such as Shin et al. (2011) tried to solve this problem by adding a choice of 'I don't know' in the test in their study to reduce the possibility of guessing. Shin et al. found that there were "relatively high percentage of responses for option 5. This indicates methodologically that the format used in the vocabulary size test was valid for eliciting L2 learners' vocabulary knowledge" (p. 135). Even though adding a choice of 'I don't know' as done in the study by Shin et al. (2011) was found to be useful to reduce the possibility of guessing, we cannot guarantee that the students always choose this choice when they do not know the meaning of a test word.

Regarding the frequency levels, the VST gives the estimates of vocabulary knowledge at more frequency levels than the VLT. The VST shows the scores at each of the 14 frequency levels while the VLT shows only four frequency levels. However, the frequency levels to be used are dependent on the test format to be used as well. Using a more demanding test format might not allow including many frequency levels, like a simple test format does (Read, 2000). Therefore, I will decide about frequency levels to test after I find the most suitable test format.

3.2.2   Research on controlled productive vocabulary

Cued-recall format

In vocabulary assessment, controlled productive vocabulary knowledge refers to the ability to recall an L2 word and the ability to use a given word in a sentence level (Laufer, 1998). To measure controlled productive vocabulary size, many researchers

used a cued recall test format (e.g., Fan, 2000; Nirattisai, 2014; and Zhong and Hirsh, 2009). A few studies used a sentence-writing format (e.g., Oberg, 2012; Zhong, 2016). Some other studies that also used a cued-recall format together with the use of a receptive vocabulary test are reviewed in Section 3.2.4 (e.g., AbManan et al., 2017; Amin, 2020). However, this section reviews studies that investigated only controlled productive vocabulary knowledge (e.g., Alonso and Garcia, 2014; Laufer and Nation, 1999).

A widely-used controlled productive test is the Productive Vocabulary Level Test (PVLT) (Laufer and Nation, 1999). The unmodified version of the test appears in Laufer and Nation's article (1995, p. 320). It has "a sentence with a word missing like a normal gap fill or 'C-Test' (Dörnyei and Katona, 1992) which the subject must complete correctly" (Waring, 1997a, p. 55). It has often been used together with the VLT if the research aim is to compare receptive and productive vocabularies. Figure 3.4 is the sample of the PVLT.

```
1. I'm glad we had this opp_____ to talk.
2. There are a doz_____ eggs in the basket.
3. Every working person must pay income t_____.
4. The pirates buried the trea_____ on a desert island.
5. Her beauty and cha_____ had a powerful effect on men.
```

Figure 3.4: Sample of the PVLT

Like the original version of the VLT (Nation, 1983; Nation, 1990), there are eighteen items in each of the four frequency levels and in the academic word section. Researchers normally choose only some levels to suit their research aims and their subjects. For example, Zhou (2010) only used the section of academic words both the VLT and the PVLT in her study because her participants were university students and not many studies at the time explored EFL learners' knowledge of academic words.

The test items of the PVLT were selected from the same lists as the VLT. A score of 80% or more at each frequency level is taken to show the mastery of retrieving items at such level. The PVLT was designed to reflect the situation where a learner needs to recall an L2 word that matches the concept he/she aims to express and fits in the context he/she will write about (Laufer and Nation, 1999). Some initial letters are given in the blanks so that a test taker can provide target answers.

Laufer and Nation (1999) measured the controlled productive vocabulary levels of four groups of EFL students and investigated their vocabulary growth cross-sectionally. There were twenty-four 10th graders, twenty-three 11th graders, eighteen 12th graders, and fourteen first-year university students in. The test was administered with four groups of EFL students. As the test requires test takers to retrieve and supply target words, the scoring system is slightly different from the VLT. "The grading was in terms of correct/incorrect for each item. Minor spelling mistakes were not marked as incorrect, and grammatical mistakes were also ignored" (Laufer and Nation, 1999, pp. 38-39).

The results showed that the controlled productive vocabulary scores of the participants decreased when the frequency levels become lower, and their scores of the academic words were between those at the 3,000- and the 5,000-level. The students at higher level significantly got higher scores, but the score of the 12th graders was not significantly higher than that of the 11th graders. Overall, frequency levels and years of study seemed to affect the participants' controlled productive vocabulary.

However, not all the four groups were significantly different at each of the frequency levels, except at the UWL level. The scores out of 18 points at the UWL level of the lowest group to the highest group were 14.44%, 29.44%, 41.11%, 70%. At the 2,000-level, there were no differences among the 11th graders (83.33%), the 12th graders (90%), and the university students (94.44%). At the 3,000- and the 5,000-levels, there was no significant difference between the 11th graders (51.67% and 21.67%) and the 12th graders (60% and 26.11%). At the lowest frequency level – the 10,000-word level, the university students' score (21.11%) was significantly higher than the other three groups. This was probably because the three groups' mean scores were approximately zero (0-5%). The insignificant difference between the 11th graders and the 12th graders was probably because the 12th graders focused on lesson revision for their entrance exam. If not, this might be because the small sample size. Laufer and Nation then suggested that future research recruit more participants per year group in their study.

Measuring controlled productive vocabulary of secondary students, Alonso and Garcia (2014) also used the PVLT with thirty-eight Spanish 10th graders. Alonso and Garcia also examined whether gender had an impact on their controlled productive vocabulary knowledge. However, this is not the focus on the present study, so I will not review the findings related to gender.

To measure the participants' controlled productive vocabulary knowledge, they combined the two parallel versions of the PVLT – all eighteen items from versions A

and only twelve items from version C so that the test battery consisted of 30 items. They did not mention how to select some items from the later version, nor did they mention the necessity of making up the 30-item test. They included only the 2,000-level because they argued that words from this level is necessary for communication both in spoken and in written form and it is suitable for secondary-level learners.

The participants sat the tests at their regular school time. Different from Laufer and Nation (1999), Alonso and Garcia (2014) checked the answers with a stricter scoring system, i.e., the answers needed to be grammatically and orthographically correct. Therefore, a verb with a wrong tense form or a word with a spelling mistake was given no point.

The results showed that their scores out of 30 points were relatively low (32.20%), and no one scored above 73.33%, not reaching the mastery level of 80%. The researchers converted the participants' scores into rough estimate of their vocabulary size by using formula by Nation (1990, p. 78). The participants had the controlled productive vocabulary size of around 644 word families. The researcher concluded that the participants' controlled productive vocabulary was relatively small and they might not be able to communicate well by speaking or writing.

Some researchers criticised the frequency levels used in the PVLT, so they adapted the PVLT to suit their study aims (e.g., Shin et al., 2011). Shin et al. (2011) does not agree with the use of only four frequency levels as they cannot provide the vocabulary knowledge level in more detail. Shin et al. then randomly selected more test words from Nation and Beglar's (2007) 14,000 BNC list and included ten frequency levels in their controlled productive test with the same format as the PVLT. I review Shin et al.'s (2011) study in Section 3.2.4, and the frequency levels for a controlled productive vocabulary test was decided later in this chapter.

For a scoring system, if the participants are asked to supply answer, a strict scoring system will not be used for grading the participants' answers. Minor misspelling should be considered as partial productive knowledge. Partial knowledge should be considered as part of someone's vocabulary knowledge; this can be seen from a tip-of-the-tongue phenomenon (Melka, 1982, p. 13; 1997).

In terms of the test format, some scholars criticised providing some initial letters in the PVLT (e.g., Read, 2000; Webb, 2008) that this might affect the validity of the cued-recall format. According to Webb (2008, p. 80), providing some letters of the test words makes the PVLT become "biased toward receptive vocabulary size" as the letters can allow a test taker to recognise the word form more easily.

However, whether or not to use this test format was decided together with using a receptive test format, i.e., when I review the previous studies investigating both receptive and controlled productive vocabulary knowledge in Section 3.2.4.

Additionally, when I searched for studies on receptive and productive vocabulary knowledge, I found some research studies particularly explored the controlled productive vocabulary by using a sentence-writing task alone. They are classroom experiments comparing different reaching methods (e.g., Barcroft, 2004), not with the aim to explore the nature of controlled productive vocabulary, though. Therefore, I do not present them in this section.

3.2.3   Research on free productive vocabulary

Essay-writing task/ collecting writing samples

Previous studies of free productive vocabulary knowledge do not estimate the size of learners' free productive vocabulary; they generally investigate vocabulary used at free will in learners' writing compositions (e.g., Azodi et al., 2014; Djiwandono, 2016; Laufer, 1998; Laufer and Nation, 1995; Laufer and Paribakht, 1998; Šišková, 2016; Walters and Bozkurt, 2009; Zheng, 2012; Zyad, 2017). They normally ask learners to write essays with provided topics (e.g., Laufer and Nation, 1995; Zyad, 2017) or collect learners' writing samples (e.g., Azodi et al., 2014; Djiwandono, 2016; Zheng, 2012) to analyse the vocabulary use. Pictures can be used for writing a short story, that is narrative writing as in Šišková's (2016) study.

This task cannot be used to measure a learner's free vocabulary size. It is difficult to devise a tool to measure the size of every word one can recall or produce freely (Laufer, 1998; Laufer and Paribakht, 1998). What researchers can do with this up until present is only to explore the vocabulary produced and used in learners' compositions. They explored only lexical richness.

As free productive vocabulary knowledge is considered as word use at someone's free will (Laufer, 1998), an essay-writing task seems to be a familiar task for EFL learners and seems appropriate for learners at a university level. Hence, this task was used in the present study. The decision about the research design (which topics, which types of essays, essay length, time duration for the test, which measures to analyse word use) was made while reviewing the previous studies.

To analyse learners' vocabulary use, some researchers employed lexical richness measures (Lexical Density (LD), Lexical Sophistication (LS), Lexical Variation (LV)

(e.g., Šišková, 2016; Zyad, 2017). They considered that learners' proficiency is related to using content words, infrequent words, and a variety of words.

Lexical density is the proportion of content words of the output supplied by a learner, lexical sophistication is the proportion of infrequent or advanced words, and lexical variation or lexical diversity is the proportion of different words in a text – "often called type-token ratio" (Schmitt and Schmitt, 2020, p. 225). Lexical Variation is also called Lexical Diversity by some researchers (e.g., Laufer and Nation, 1995; Schmitt and Schmitt, 2020). Laufer and Nation (1995) present the formulas for these measures in their paper (pp. 309-310).

$$LD = \frac{\text{Number of lexical tokens x 100}}{\text{Total number of tokens}}$$

Equation 3.1: Formula for the Lexical Density index

$$LS = \frac{\text{Number of advanced tokens x 100}}{\text{Total number of lexical tokens}}$$

Equation 3.2: Formula for the Lexical Sophistication index

$$LV = \frac{\text{Number of types x 100}}{\text{Total number of tokens}}$$

Equation 3.3: Formula for the Lexical Variation index (also called Lexical Diversity or TTR)

Laufer and Nation (1995, pp. 309-310) do not view the above indices as reliable or valid measures. They write:

> "The LD measure is influenced by the number of function words, … The same piece of writing may be analysed differently in terms of LS, depending on how 'advanced' vocabulary is defined. This makes the measure unstable. … The type/token ratio [or the LV] has been shown to be unstable for short texts and can be affected by differences in text length; … LV also does not distinguish what kinds of words are used. In a composition of 300 tokens, for example, 200 word types could be used by someone who knows 2,000 words, or someone who knows 5,000 words" (Laufer and Nation, 1995, pp. 309-310).

Also, analysing vocabulary use in terms of lexical richness does not show learners' free productive vocabulary knowledge at different frequency levels as I aimed to in this study. I, therefore, decided not to use these lexical richness measures.

Seeing the problems of the lexical richness measures, Laufer and Nation (1995) then introduced another measure used to analyse vocabulary use in learners' compositions, and this is Lexical Frequency Profile (LFP). Lexical frequency profiles show the number and percentages of used tokens, types, and word families that fall into different frequency levels: the first 1,000 frequent words in the GSL, the second 1,000 frequent words in the GSL, the words from the UWL, and the words that do not fall in these three lists (called as the 'Not-in-list' group). However, Range can now be downloaded on Nation's website (www.wgtn.ac.nz) with the GSL, the AWL or the BNC.

A more recent and simpler lexical-profiling software than the Range (Nation, 2008) called VocabProfile has been devised by Cobb (n.d.). It shows the number and the percentages of tokens, types, and word families into four frequency groups as well.

Figure 3.5: Sample of VocabProfile on Lextutor Website (Cobb, n.d.)

Many research studies also used a vocabulary-profiling programme to analyse learners' vocabulary use (e.g., Azodi et al., 2014; Djiwandono, 2016; Hatano, 2008;

Hsu, 2014; Laufer, 1998; Laufer and Nation, 1995; Laufer and Paribakht, 1998; Lee and Muncie, 2006; Malmström et al., 2018; Muncie, 2002; Treffers-Daller et al., 2016).

Laufer and Nation (1995) used this measure in their research study in order to establish the reliability and the validity of the LFP. They cross-sectionally analysed vocabulary used by three groups of learners with different proficiency levels: 22 EFL students who enrolled on English for Academic Purposes course in New Zealand (Group 1: the lowest proficient group), 20 Israel students majoring in English Language and Literature and studying in the first semester (Group 2), and 23 Israel students in the same major but completing their second semester (Group 3: the highest proficient group).

The participants were asked to write two essays of about 300 words (tokens) each as part of their regular classwork so that they would keep motivated in writing the compositions. The learners wrote these two compositions on different days within one week; the interval needed to be short because the researchers did not want the subjects' productive vocabulary knowledge to change. The topics were of "a general nature and dealt with controversial issues" (p. 314). The first topic for the first composition was the same for all the subjects and for the second composition the subjects could choose one from the three offered topics they were interested in. Figure 3.6 shows the topics for the two compositions that Laufer and Nation (1995) used in their study.

Essay 1

The Common Topic:

"*Should a government be allowed to limit the number of children a family can have?*" **Discuss this idea considering basic human rights and the danger of population explosion.**

Essay 2

One topic to be chosen out of three:

"A person cannot be poor and happy, because money is always needed to gain something that is important to that person." Argue for and against this idea.

> "It is always what you do not have as a child that is important to you as an adult." Agree or disagree with this statement.
>
> "In a free country, industry has the right to develop any product that will sell, and industry can sell it to anyone who can pay for it." Do you agree with this idea or do you think that the government should be able to control what is produced and sold?

Figure 3.6: Essay Topics in Laufer and Nation's (1995) study

The subjects spent one hour on each composition. Then only the first 300 words of each of their compositions were typed into the computer. A word used incorrectly was not typed in because Laufer and Nation did not consider that as part of the subjects' productive vocabulary knowledge. However, they corrected misspelled words before typing them in the LFP and these words were considered as the subjects' partial productive vocabulary knowledge.

The results showed the trend of their vocabulary use. The participants with lower proficiency level tended to use more of the high-frequent words (the first and the second 1,000 frequent words) and fewer of the academic and infrequent words (the UWL words and the 'Not-in-list' words) than the participants with higher proficiency level. The percentages of using words from the 1,000-level in the first compositions for Groups 1, 2, and 3 were 86.50%, 79.70%, and 77.00%. Those from the 2,000-level were 7.10%, 6.70%, and 6.60%. The percentages of academic words in their writings were 3.20%, 8.10%, and 8.10%. Those of 'Not-in-list' words were 3.30%, 5.60%, and 7.50%. The trend for the percentages of their word use in the second compositions were similar.

> "These differences are in accordance with the concept of language proficiency which assumes that richer vocabulary is characteristic of better language knowledge. If the LFP has tapped these differences, this is evidence for its validity. … [and n]one of the differences between the two essays is significant, showing that the LFP is stable between two compositions" (Laufer and Nation, 1995, pp. 316-318).

Therefore, the LFP is a reliable and valid measure for lexical richness. The LFP has been used by more recent studies (e.g., Azodi et al., 2014; Djiwandono, 2016; Lemmough, 2008; Zheng, 2012).

Zheng (2012) also investigated the development of free productive vocabulary of advanced EFL students. Her study was a multiple-case study, including only four

third-year Chinese students majoring in English, because she believed that with this research design she would be able to better track how learners' productive word use change even though the findings might not be able to generalise. Unlike, Laufer and Nation (1995), she did not ask the subjects to write compositions specifically for the research. She aimed to examine their productive vocabulary use in a naturalistic manner, so she asked the subjects to also send her electronic versions of the drafts they wrote in their argumentative writing classes. She collected their writing compositions three times during the academic year.

She analysed their vocabulary use by using VocabProfile software. She deleted the composition titles and proper nouns from the students' free writings before uploading them onto VocabProfile. Only the percentages at the level of beyond-2,000 (the percentages of the AWL group combined with that of the 'Not-in-list' group) were compared to explore the changes because, based on Laufer (1994), the beyond-2,000 level was an indication of "is a clear indication of productive vocabulary development" (Zheng, 2012, p. 109).

The findings showed that not all of the subjects used more difficult or academic words in their drafts later in the academic year but used more of high-frequency words instead. Their average percentages of vocabulary use at the beyond-2,000 level were 16.57% at the first session, 14.45% at the second session, and 10.37% at the third session. Examining only the percentage of the academic words used, they found that it also decreased (5.77%, 5.55%, and 3.81%). Differently, the percentages of their high-frequency words increased (83.43%, 85.55%, and 89.63%).

Zheng (2012) also investigated the influence of their motivation towards this declining profile by using face-to-face semi-structured interview. She found that the learners did not think that using infrequent words was important – only easy words could be used to communicate successfully. Also, one of the subjects used to think that using difficult words could lead to better writing scores. She tried hard and gave up, she then recycled only easy words in her essays instead. Even though my study does not focus on motivation, it is also good to keep in mind that learners' productive vocabular development might be affected by any of these extralinguistic factors – students' thought and motivation.

Azodi et al. (2014) studied free productive vocabulary growth cross-sectionally. They collected 210 writing samples from all the four year cohorts of Iranian university students; these students majored in translation. The reason why they analysed learners' vocabulary knowledge through writing tasks is that writing tasks are similar to the task learners need to do in class, or, as Zheng (2012, p. 107) mentioned, "in a naturalistic manner". After collecting the writing samples, they analysed them by

using Range (Nation, 2008) and compared the lexical profiles of these writing samples produced by the four groups of students. The Range software showed the number, together with percentages, of vocabulary used into four categories: words in the first 1,000 of the GSL, words in the second 1,000 of the GSL, words in the AWL, and words not in the lists.

The results showed that the students from all the four years used almost equal number of words at the 1,000-level. The percentages of types used at this level by the students from Year 1 to 4 were 59.77%, 57.09%, 53.10%, and 59.67%. The second-year students used more words from the 2,000-level, but not many were used by the third-year or the fourth-year students; the percentages of types used by Year 1-4: 13.46%, 18.83%, 6.68%, and 8.51%. There was an increase in using academic words from Year 1 to Year 4 (6.08%, 8.86%, 20.92%, and 15.62%.), with the third-year students using the most of them (a significant increase with p value of less than 0.01).

The researchers write, "to some extent the changes were predictable" (p. 1,846). For example, the students from Year 3 and 4 were likely to read and understand authentic and academic texts, so they used more words from the AWL in their productive tasks. Using a lot of academic words, the proportion of the high-frequency words used then decreased. The researchers concluded that "EFL students not only add to their vocabulary knowledge during the years of study, as they are exposed to different sources of English vocabularies, but also use this knowledge in their productive tasks" (p. 1846).

Djiwandono (2016) investigated the free productive vocabulary used in academic papers or theses by fourth-year students of English Letters Department in an Indonesian university and compared it to that of their lecturers' academic papers. Even though lecturers are not the focus group of my study, they can serve as a group with higher proficiency level. Djiwandono analysed the vocabulary use by using two measures: VocabProfile for lexical profile and TTR for lexical variation. Among 12 theses in the year 2007-2009, nine of these were sampled. He did not mention why he sampled these nine theses. Only some chapters (chapters 1, 4, and 5) were entered into VocabProfile. Djiwandono argued that these chapters were more likely to relate learners' thoughts, not citations of other sources. Nine academic papers of the lecturers were also sampled from 25 papers, but the entire papers were used for the analysis. However, this sampling process seems unbalanced. Also, the citations in some excluded chapters might have been paraphrased, so this must be considered as the learners' ability to use academic words as well.

The comparison showed that the lecturers used significantly more academic words (9.48%) than the students (6.83%) and their TTR was significantly higher (0.28 vs 0.17), with p value of less than 0.05. The students used more of the words from the 2,000-level (83.01% vs 81.38%) and the 'Off-list' words (10.16% vs 9.14%), but these differences were not statistically significant. In Laufer and Nation's (1995) study, higher proficiency writers used more of 'Off-list' words, though. Djiwandono then concluded that "the lecturers use more variety of words and use academic words more frequently than their students" (p. 214).

Using both the lexical richness measure and the LFP, Zyad (2017) examined the free productive vocabulary knowledge in terms of lexical diversity and lexical sophistication. The lexical diversity in their study was analysed by using D-index (Malvern et al., 2004). D-index "models the best TTR curve in texts with different text lengths, hence overcoming the shortcoming of TTR being affected by varying text lengths" (Zyad, 2017, p. 367). The lexical sophistication in this study was analysed by using the LFP (Laufer and Nation, 1995). The participants were 90 participants from Year 1 to Year 3 in a Moroccan university and they were asked to write 30 narrative essays on misjudging people from their appearance. The researchers did not state the length of each essay. The researchers controlled how the participants wrote their essays by providing them with details "such as the target audience, the purpose of writing and a few details about the plot of events" (p. 367) so that the participants could focus more on accuracy and lexical use.

The D-index analysis showed that there was a significant increase in lexical diversity from Year 1 to Year 3, or the students from Year 3 used more of different words than the first-year group. Regarding the lexical profiles or lexical sophistication, the analysis showed that the "LFP did not develop across proficiency levels in a linear, upward manner" (p. 368) at the first and the second 1,000 word families of the GSL. The LFP at the 1,000-level did differentiate the first-year students from the second-year group, with the second-year group producing fewer words at the 1,000-level. At the 2,000-word level, the LFP did differentiate the third-year group from the other two year cohorts. Even though the LFP index did increase at the academic word level (UWL), there was no significant difference among the three year cohorts. Therefore, different participants had their own different vocabulary profiles, and for each frequency level the proportions of used vocabulary at different frequency levels might not show linear, upward trends.

I conclude from these studies that an essay-writing task is suitable for university students as this task requires the participants to produce vocabulary freely, which can serve as their free productive vocabulary. This format was then used in the

present study. As this study was conducted at the beginning of the academic year and by then the first-year students would not have any writing assignments. In the first semester, they only enrolled in general English (English speaking and listening). Accordingly, it was more practical to ask them to write an essay. In this way, I could control the topic for all the four year cohorts.

Since the type of essays, the topics, the time duration to complete the task used in Laufer and Nation (1995) was experimented with multi-cultural students at a university level and with different proficiency levels, I then decided to use argumentative essays because they seem to be able to elicit a variety of words from high- to low-frequency words as well as academic words from the participants. Narrative essays as used by Zyad (2017) was not used since they seemed to mostly elicit high-frequency words. Compositions from all the four year cohorts would be comparable because they would be controlled by the same topics and the participants would take about an hour per essay.

Regarding the measures for vocabulary use, lexical richness indices (lexical variation, lexical sophistication, lexical density) were not used for the main study because they do not analyse how well words are used. Schmitt and Schmitt (2020) argue:

> "[Even though lexical variation, lexical sophistication, and lexical diversity] have the advantage of eliciting spontaneous use of vocabulary in learner compositions…, [but] they rely on the assumption that output with higher lexical variation, lexical sophistication, and lexical density indicates a larger vocabulary and more proficient language use. Unfortunately, this is simplistic, and quite often wrong. What really matters in vocabulary use is that the right word is used accurately and appropriately in a particular context, and automated measures are not good at doing this kind of "goodness of use" analysis" (pp. 225-226).

Accordingly, lexical richness was not used in my study. The LFP or the VocabProfile was not used, either. They group words produced by learners into only a few frequency groups. The software focuses on only two high-frequency levels and the academic words. The words from mid-frequency levels or low-frequency levels are grouped into one big group considered as infrequent words, where they are supposed to be treated as separate groups. Hence, what software I used to analyse the participants' free productive vocabulary was decided after the receptive and controlled productive vocabulary tests were selected. The flowchart of how I analysed the free productive vocabulary are presented in Chapter 4.

3.2.4 Research on receptive and controlled productive vocabularies

Matching format and cued-recall format

Even though results from a matching format can be affected by participants'
guessing and a cued-recall format is biased towards receptive knowledge (word form
recognition), many studies used these two formats together to investigate the
relationship between receptive and controlled productive vocabulary knowledge
(e.g., AbManan et al., 2017; Fan, 2000; Ozturk, 2015; Waring, 1997a; Zhong and
Hirsh, 2009; Zhou, 2010). This is probably because they are based on the idea that
receptive vocabulary and productive vocabulary are separate groups of words or
candidates for threshold effects (Meara, 1990a). I review them to see how they
compared receptive and controlled productive vocabularies and the common findings
about the relationship between the two groups of vocabulary.

AbManan et al. (2017) explored the receptive and controlled productive vocabulary
knowledge of 156 first-year non-English-major university students in Malaysia. They
examined whether these students were ready for their university education because
the English language is used as a medium in teaching and learning in this university.
They remark:

> "It has been established that ESL learners' vocabulary knowledge correlates
> highly to their general proficiency in English. The vocabulary size of ESL
> learners is seen as an essential aspect of readiness of first-year students to
> adapt to the university learning environment especially when English is used
> as the medium of instruction" (p. 53).

Based on the idea of a dichotomy, the researchers used two vocabulary levels tests:
the VLT (Nation, 1990) at the 2,000-, the 3,000-, the 5,000-word levels and the
academic word level (UWL) and the PVLT (Laufer and Nation, 1999) at the 2,000-
and the 3,000-word levels. These frequency levels were selected because "tertiary
(advanced) level ESL learners need to have productive vocabulary knowledge of
around 2,000 to 3,000- word families, and receptive vocabulary of around 3,000-
5,000 word families in order to function effectively as university students" (p. 53) and
academic words are the basic requirement to perform academic tasks university
learning environment (p. 58).

Nevertheless, I do not agree with how the researchers choose the frequency levels
for the two tests. It would be better not to assume the participants did not have the
ability to recall words from the 5,000-word level or from the academic word level.
Also, because the participants are at a university level, it is unavoidable for them to
recall some academic words in their learning. Including the academic word level

might make them see clearer whether the participants are ready for an academic environment.

The researchers did not mention about how to score the answers. The results showed that their vocabulary knowledge scores decreased at the lower frequency levels. The participants knew about 89%, 77.6%, 73%, 74% receptively at the 2,000-, the 3,000-, the 5,000-word levels and the academic word level and had a controlled productive knowledge of around 71.2% and 39.6% at the 2,000-, and the 3,000-word levels. They also found that learners' receptive vocabulary was larger than their productive vocabulary, 89% vs 71.21% at the 2,000-word level and 77.6% vs 39.6% at the 3,000-word level.

Moreover, they found that majority of the participants were not ready for learning at university level. The majority of the participants reached the mastery level (80% of scores) of receptive vocabulary knowledge at the 2,000- and the 3,000-word levels (86% and 54% of the participants, respectively), but only half of them passed the 5,000-level and the academic level (48% and 50% of the participants). For their controlled productive vocabulary knowledge, half of the participants (54%) mastered the level of 2,000 frequent words, and only 3% of them passed the 3,000-word level.

Also working on the assumption that the relationship between receptive and controlled productive vocabularies is dichotomous, Zhou (2010) compared the size of receptive and controlled productive vocabularies of 72 Chinese non-English-major university learners. However, as at the time not many studies focused on academic vocabulary, she was interested to explore only academic vocabulary size. She selected only academic section of the VLT (Schmitt et al., 2001). It contains 30 items. To make a controlled productive test comparable to the receptive vocabulary test, she developed a new productive test using the cued recall format like the PVLT (Laufer and Nation, 1999) and sampled 30 academic words from the AWL (Coxhead, 2000) – the same word list for the receptive test. This is a sensible solution. Equalising the number of the test items for the two tests and sampling words from the same word list make the two tests more comparable.

Zhou scored the receptive test and the controlled productive test slightly differently. While she employed a dichotomous scoring – correct or incorrect answers in the receptive test, she allowed partial knowledge performed in the controlled productive test. That is, she gave one point for a correct answer, half a point for a partially correct answer, and no point for incorrect answer. Partially correct answers refer to misspelled words which can be pronounced the same as the correct answers.

Multi-scale scoring like what Zhou adopted seems to be a suitable method of grading what learners produce or supply since vocabulary knowledge is not an all-or-nothing

phenomenon (Laufer, 1998; Melka, 1997). Partial knowledge should be considered as part of one's knowledge. This is dependent on test formats as well; a dichotomous scoring should not be used if the test is not a matching or a multiple-choice format. Therefore, this scoring system was used to grade the participants' word use in the present study.

Like those in AbManan et al.'s (2017) study, the results showed that participants had a larger receptive vocabulary size than their controlled productive vocabulary. The receptive mean score was 78.00% (23.4 out of 30 points), while the mean score of controlled productive test was 35.33%. The controlled productive vocabulary lagged behind. The correlation between the two was statistically significant, though. This implies that a learner with a larger receptive vocabulary seemed to have a larger controlled productive vocabulary, and vice versa.

Zhou also explored in more depth by dividing the participants into two groups based on proficiency levels or the receptive vocabulary scores. The participants scoring higher than the average score (23.44 points) were called Group 1, and those with lower scores were called Group 2. For both groups, their receptive academic vocabulary size was larger than their controlled academic productive size. The correlation between the receptive and the controlled academic productive scores of Group 2 (0.595) was higher than that of Group 1 (0.449). This implies that when learners acquire more receptive academic words, the gap between their receptive and controlled productive academic words then becomes wider, that is, the correlation becomes lower. She then concluded that receptive academic knowledge develops faster than controlled productive academic knowledge and that the growth of these two types of knowledge go along each other.

Even though the methodology in Zhou's study (the way she used different test formats to explore receptive and controlled productive vocabularies) seems to rely on the idea of dichotomy, that is, receptive and controlled productive vocabularies are separate groups of words competing against each other and growing at a different rate, she suggests that controlled productive vocabulary knowledge develops after receptive vocabulary knowledge. She writes:

> "The first step in learning a word is often to be able to recognize a word by its form and understand what it means. Then gradually, with more practice of this word in reading, listening, writing, or some other activities, learners increase their knowledge of this word and learn to use the word productively in writing or speaking. Therefore, it is apparent that productive vocabulary knowledge builds on receptive knowledge" (p. 16).

This idea is similar to the idea of continuum. The difference of these two beliefs might be that there is no dividing line between receptive vocabulary and productive vocabulary, but the knowledge of a word develops from a receptive stage towards a productive stage. Hence, the shared idea is that reception comes before production.

Unlike AbManan et al. (2017) and Zhou (2010), Fan (2000) focused more on the gap between receptive and productive vocabularies. She designed the research study based on the idea that receptive vocabulary is acquired before productive vocabulary and that words should be recognised before they can be recalled. She measured the receptive vocabulary size of 138 higher-diploma students at a university in Hong Kong. To examine the gap, or to be more specific, the proportion of the recalled words (controlled productive vocabulary) to the recognised words (receptive vocabulary), Fan used the long version of the VLT provided by Nation. This receptive vocabulary level test contained 72 items in each frequency level, instead of 18 items of the original version. Fan used only three sections of this test: the 2,000-word level, the 3,000-word level, and the academic word level. A dichotomous scoring system was employed. Among 138 students, there were 9 groups of them – three sub-groups from three major disciplines: Engineering Surveying (ES), Shipping and Management Studies (SMS), and Building Technology and Management (BTM). Among these three groups, three of the SMS groups had the highest receptive scores or they could recognise the most words (27.30% (59 words out of 72 words), 23.60% (51 words), and 23.6% (51 words)).

After Fan identified which words everyone in each group could recognise, she created a controlled productive test in the format of cued recall by including only words commonly recognised by such group and made the students take the controlled productive tests weeks later. In this way, test words included can be the same words for both receptive and controlled productive tests. Fan expected that those who recognised more words were supposed to be able to recall more words, or the ratio between the controlled productive and receptive vocabularies should be higher.

The calculation showed unexpected results. First, there were no consistent ratios between recognised words and recall words. The ratio ranged from 53% to 81%. Second, being able to recognise more words did not mean being able to recall more words. For example, one group of SMS could recognise 51 words out of 72 words but was able to recall 27 words, while one group of BTM could recognise fewer words (44 words) but could recall more words (34 words). Third, the controlled productive/receptive ratio was not related to learners' English proficiency or the size of their receptive vocabulary. For instance, while the second group of the BTM

recognised and recalled more words (22 and 13 words) than the third group of BTM (15 and 10 words), the controlled productive/receptive ratio of the second group (59.1%) was lower than that the third group (66.7%). Fan concluded that the ratios vary greatly and cannot be roughly estimated from knowing someone's vocabulary size. In vocabulary assessment, it might be difficult to estimate learners' productive vocabulary size from knowing their receptive vocabulary size.

However, there are problems with how she designed the controlled productive test. Selecting only the words that could be recognised by everyone in the group might distort the original findings as some students might be able to recognise more words than their peers in the group. Hence, using two tests with the same test words and asked the participants to do the recalled task first might be better. Then their controlled productive/receptive ratio can be analysed and averaged for each group.

Waring (1997a) also compared the receptive and productive vocabulary sizes of second language learners. The study was done at the beginning of a semester for a diagnostic purpose with 76 Japanese university students who majored in English (the first-year and the second-year students). The participants' proficiency level was at upper elementary. He used the VLT (Nation, 2001) and the PVLT (Laufer and Nation, 1999). He argues that tests that can profile test takers' vocabulary knowledge into different frequency levels are useful for pedagogy.

> "This kind of measure, from a word frequency perspective, is useful to teachers and learners to enable them to determine which words are well known and which need the most attention" (Waring, 1997a, p. 53).

Based on one of his studies conducted the previous year with similar group of participants, he adapted the frequency levels by excluding the 10,000-word level and the UWL level because in the previous study the participants' scores at these two levels were relatively low, close to zero, so their inclusion would not give much information about the learners' vocabulary knowledge. To be able to discriminate the vocabulary knowledge at higher frequency level better, he added one more level – the 1,000-word level. As the test words at the 2,000-word level contained words from the most 1-1,000 and 1,001-2,000 frequent words, he then regrouped the test words and randomly selected more words from the GSL so that each level had 18 items like the other levels. Hence, the receptive test in his study consisted of 72 items and the controlled productive test had 72 items with the same test words.

The participants sat the PVLT before the VLT to avoid learning effects from one test to the other. A dichotomous scoring system was used with the receptive test, but a non-strict scoring method (also offering half a point for a partially correct answer) was used with the productive test. He argues, "In the researcher's opinion this was

necessary as it would seem that a strictly marked productive test would yield lower productive scores and falsely widen the differences" between the receptive and the productive scores (p. 60).

The results showed that their receptive vocabulary was larger than the controlled productive vocabulary at all the levels and for all the participants. On average, the receptive vocabulary was around 2.23 times larger than the productive one. However, Waring (1997a, p. 57) states, "[I]t is too simplistic a statement" because the gap varies from level to level, i.e., the receptive vocabulary was 1.58 times larger at the 1,000-word level, 1.79 times at the 2,000-level, 3.17 times at the 3,000-level, and as much as 6.45 times at the 5,000-level. The gap between the receptive and the productive vocabularies was wider at the lower frequency levels.

Also, when he calculated the controlled productive/receptive ratio, he found that the ratio was high at the high-frequency level and low at the low frequency level. The ratios from the 1,000-level to the 5,000-level were 64.5%, 55.7%, 31.5%, and 15.5%, respectively. He interpreted the data that, at higher frequency levels, there was more chance for receptive vocabulary to become productive vocabulary; the percentage of chance was the percentage of the ratios.

Besides, to investigate whether these ratios were also still the same for people with different vocabulary sizes, Waring divided the participants into three proficiency groups based on the sum of their receptive and productive scores: a lower group, a middle group, and an upper group. He found that the profiles of their receptive and productive vocabularies were roughly the same with the smaller receptive-controlled productive gap for the upper group.

To explore the relationship between receptive and controlled productive vocabularies with the idea of dichotomy in my study, I calculated these ratios but at which frequency levels was decided after the test formats was selected.

Using the same tests as Waring (1997a), Ozturk (2015) focused more on learners' receptive and controlled productive vocabulary growth over years of study and the effect of frequency levels towards the vocabulary development. She conducted two studies, the first one was a cross-sectional study and the other one was the longitudinal study.

In her first study, she compared the receptive and the productive vocabularies of two groups of learners by using the VLT (Nation, 2001) and the PVLT (Nation, 2001). The participants were 55 first-year students and 45 fourth-year students in an English-medium programme of a university in Turkey. They were advanced students. Ozturk assumed that the participants from these two groups had similar vocabulary

knowledge level when entering to the university and the main difference was only the "extra years of exposure to English through the academic study" (p. 97).

In my study, I divided the students into two broad groups, but I did not assume their proficiency levels. Therefore, the students in Year 3 and 4 were considered as the higher year group, and the students in Year 1 and 2 were considered as the lower year group.

All the four frequency levels and one academic word level of the two tests were included. In Ozturk's (2015) study, 90% was considered the mastery level, which seems relatively high compared to other studies using the VLT and the PVLT. The results showed that the frequency levels significantly affected the participants' vocabulary knowledge. Their receptive scores were around 95% at the 2,000-level, 86% at the 3,000-level, 70% at the 5,000-level, 32% at the 10,000-level, and 85% at the academic word level. Their controlled productive scores were around 75%, 43%, 30%, 12%, and 40%, respectively.

However, there was no significant difference (p value of higher than 0.05) between the receptive and controlled productive scores of the two groups; the scores were almost equal for all the levels. This means their vocabulary did not improve during their study in the programme. For example, the receptive scores of the first-year students at the 2,000-word level was 95% and that of the fourth-year was 94%. Their receptive vocabulary scores were higher than their controlled productive ones at all the five levels. Their receptive vocabulary at all the levels was around twice larger than their controlled productive vocabulary. For example, the receptive academic vocabulary score of the first-year students was 80%, while their controlled productive vocabulary score was 40%.

As there was no significant difference in the scores of these two groups of students, Ozturk then conducted another study (Study 2, also presented in the same paper) to verify the first study. It was a longitudinal study tracing the vocabulary development of the first-year students in Study 1 over their three years of study in the programme.

The same tests were handed to the 55 students again when they were in their final year and let them do the tests when they are available. By doing so, it is difficult to guarantee that they relied on their knowledge only and did not look up words in dictionaries. This led to a returning problem as well. Therefore, I asked the participants to sit the tests at one of the faculty's rooms.

In Ozturk's (2015) second study, only seventeen of the participants returned the tests (a return rate of 32%) – a problem of a longitudinal study. Therefore, only the scores of these seventeen students when they were in their first year were compared

to their scores when they were in their fourth year. The ANOVA analysis showed unexpected results. Even though their receptive scores increased but the differences were not significant. Frequency levels still had a significant effect on their vocabulary development, though. Despite being good and motivated learners, the participants did not show higher receptive vocabulary scores over years of study at the four frequency levels but did significantly improve at the academic word level.

Ozturk then assumed that this might be because of the ceiling effect as their receptive scores nearly reached 100% at the 2,000-level (96%) and the 3,000-level (91%). When comparing the scores of these seventeen students with the average scores of the rest of their peers in Study 1, these seventeen students seemed to be better learners than the rest. Therefore, there was little room for improvement. However, there were significant differences in their controlled productive vocabulary scores at all the levels – only around 10% higher in their final year.

Ozturk argues that the possibilities of the insignificance in the learners' receptive vocabulary development over the years of study might be either the disadvantages of the learning context or the drawback of the test itself. First, about their learning context, the learners might have known a lot of high-frequency words and needed to learn a lot of academic words. When reading, they could manage to understand the texts, so they ignored few mid- or low-frequency words they encountered. This leads to less chance of learning them. Second, the VLT might not be sensitive enough to capture their vocabulary gain. The gain might happen at the 4000-level or 6000-level, but the VLT is not evenly spaced from one frequency level to another. Hence, Ozturk suggests using the VST or Y_Lex, a checklist format (Meara and Miralpeix, 2006) for measuring the vocabulary development of advanced students. Nevertheless, it is not certain whether the test takers really know the test words if a multiple-choice or a checklist format is used.

The significant improvement of controlled productive vocabulary in the longitudinal study might be because of the learners' larger receptive vocabulary in Study 2 and opportunities to practice writing. When learners knew more receptive vocabulary, they appeared to know more controlled productive vocabulary. Also, that the programme required learners to write compositions might be conducive to the growth of controlled productive vocabulary. Frequency affects the order of receptive vocabulary development, but productive vocabulary development is probably influenced by learners' need to produce words when expressing themselves. Lastly, Ozturk concluded that there was not much growth in receptive academic vocabulary knowledge, and this might be because of the ceiling effect resulted from hard work

for the entrance exam, while there was room for improvement for their productive vocabulary.

Zhong and Hirsh (2009) longitudinally examined the growth of receptive and controlled productive vocabularies over ten weeks' time by using the VLT (Zhong, 2008) and the PVLT (Zhong, 2008). The researchers approached 83 students high-school students (11th graders) in China, but only 64 of them gave consent to participate in the study, and only 41 students completed all the tests. Therefore, the data of these 41 students was analysed in the study. Both tests were administered twice as a pre-test and a post-test. As the knowledge of 10,000 word level would be beyond the participants' expected vocabulary size, it was excluded from the tests. Only the 2,000-, the 3,000-, the 5,000-word levels and the AWL level were included. The maximum score for each of the VLT was 30, while that of the PVLT was 18.

The results showed that the average scores for both receptive test and the controlled productive test decreased when the frequency level decreased. For example, the receptive pre-test scores at the 2,000-, the 3,000-, and the 5,000-levels were 88.20%, 77.13%, and 54.47%. The mean scores at the academic word level for all the tests were between those at the 3,000- and those at the 5,000-levels; the receptive AWL pre-test score was 70.30%.

The results also showed that the participants' post-test scores were significantly higher than their pre-test scores at all levels (p value of less than 0.05), the receptive test at the 3,000-level and the controlled productive test at the 2,000-level increased but without statistical significance. The researchers assumed that these insignificances might be because the high scores in the pre-tests at that level, so they did not "leave a measurably large enough margin for improvement" (p. 100). Another possible reason assumed by Zhong and Hirsh was about the small sample size – forty-one students – as in a study with a large sample only a small difference would be statistically significant. However, according to Cohen et al. (2017), thirty was an optimal number of participants per variable. Hence, I do not see forty-one participants were a small group.

When they converted the scores into vocabulary size, they found that the participants' receptive vocabulary size was larger than their controlled productive vocabulary size at all levels and for both pre-test and post-test. Different from what Zhou (2010) found, the controlled productive/receptive ratio showed that the ratio at all levels increased in the post-tests. This implied that the gap between these two types of vocabularies became smaller, with the controlled productive vocabulary growing at faster rate during these ten months. For example, the controlled productive/receptive ratio in the pre-tests at the 2,000-level was 96.31%, while that in

the post-test was 97.26%, and the gain of the receptive score from the pre-test to the post-test at the 2,000-level was around 3.12%, but that of the controlled productive score was 4.18%. This might be because there was not much room for receptive knowledge to improve and this might often happen at high-frequency levels.

Finally, also based on the controlled productive/receptive ratios, Zhong and Hirsh found that at higher frequency levels there was more chance that the receptive vocabulary would become the controlled productive vocabulary. In other words, the ratios tended to be high at higher frequency levels. This is consistent with the results in Waring's (1997a) study. For example, in the post-tests, the ratio at the 2,000-level was 97.26%, implying that there was as much as 97.26% chance that the receptive vocabulary at this level would become controlled productive vocabulary. The ratio at the 5,000-level was 64.30%, implying that there was 64.30% chance that the receptive vocabulary at this level would become controlled productive vocabulary.

Multiple-choice format and cued-recall format

The following studies used a multiple-choice format together with a cued-recall format to compare learners' receptive and controlled productive vocabularies (e.g., Amin, 2020; Kotchana and Tongpoon-Patanasorn, 2015; Shin et al., 2011).

Kotchana and Tongpoon-Patanasorn (2015) compared the receptive and productive vocabulary size of 453 Grade 6 students in Thailand. The participants were sampled from 4 primary schools in the North-east of Thailand to represent the population of Grade 6 students in this region. The researchers did not use the available widely-used vocabulary size tests but designed their vocabulary tests themselves. The receptive vocabulary size test was a bilingual multiple-choice format, and the productive vocabulary size test was a fill-in-blank or cued recall format, the sentences of which were embedded in conversations or dialogues. Using these formats, the researchers view productive vocabulary knowledge as the ability to recall words or controlled productive vocabulary knowledge (Laufer, 1998).

As the participants were only primary school and the Ministry of Education (2008) expected Grade 6 students in Thailand to have knowledge of around 1,000 word families, the test words were sampled from only the first 1,000 words in the GSL. There were twenty items in each test representing 1,000 word families in the GSL. They administered the receptive test before the controlled productive test. They did not report whether they used the same test words in both the tests. If so, the test scores might have been affected by a learning effect. I learned from this that I tried to avoid the learning effect in my study.

The raw score was then converted into vocabulary size by using this proportion. The data showed that the participants' receptive vocabulary size (463 word families) was almost twice as large as their controlled productive vocabulary (292 word families).

I learned from the literature that learners' receptive vocabulary is always larger than their controlled productive vocabulary, regardless of whether the tests were newly designed or the available widely-used ones. Second, from all the studies conducted with Thai learners, Thai learners do not seem to know much vocabulary knowledge, regardless their educational levels. This might be because they do not have opportunities to practice it.

With similar purpose, Shin et al. (2011) measured the receptive and controlled productive vocabulary size of 402 Korean high-school students to examine the relationship between these two groups of vocabulary and to check whether the Korean national curriculum was still suitable for the high-school seniors.

They selected and adapted a bilingual version of the VST (Nation, 2010) for measuring the participants' receptive vocabulary size because they argued that some of the test items were culture-specific such as the words 'ruck' and 'lintel'. They then randomly selected 100 words from the first to the tenth levels of Nation and Beglar's (2007) 14,000-words BNC word list by using Random Item Generator v.1 on Lextutor website. Also, they added one more choice 'I don't know' as they believed that it would help reduce the possibility of guessing.

To measure the learners' controlled productive vocabulary size, the researchers adapted the cued recall vocabulary test – PVLT (Laufer and Nation, 1999) by using the same tool to select 100 words from the same word list so that the number of scores would be comparable. Furthermore, in the adapted PVLT, the researchers also provided Korean translations for each sentence in the test with the target word equivalent written in bold. The researchers claimed that, with the translation, test takers would not need more knowledge of a test word beyond their form and meaning, such as the need of knowledge of its collocation. This would maintain the test construct validity. Hence, Shin et al.'s controlled productive test can be viewed as a translation test.

A dichotomous scoring was adopted with the answers in the receptive test, while the three-scale scoring (correct, partial correct and incorrect) was used with the controlled productive test as sometimes test takers supplied misspelled answers that still looked like the correct answers or the answers with minor grammatical mistakes such as missing the plural -s were considered as partially correct and received half a point.

The findings showed that the participants' receptive vocabulary was larger than their controlled productive vocabulary at all the frequency levels. On average, their receptive score was 60 out of 100, and their controlled productive vocabulary score was only 24.40. When multiplied by 100 to convert the scores into vocabulary size out of the 10,000 words, their receptive and controlled productive vocabulary sizes were around 6,000 and 2,440 word families, respectively. The receptive vocabulary was almost three times larger than the controlled productive vocabulary. They calculated the controlled productive/receptive ratio by using their raw scores and it was 39.71%. "[T]he results indicated that the learners were able to produce 39.71% of the words that they knew receptively. However, the percentages cannot be generalised for all word levels" (Shin et al., 2011, pp. 133-134), like those found in Waring's (1997a) study.

When they examined in more detail at each of the frequency level, the controlled productive/receptive ratios varied. Within the ten frequency levels, the ratio was high at the high-frequency level, and it was the opposite at the lower frequency levels. For example, the controlled productive/receptive ratio at the first 1,000 word level was 86.01%, while that at the 10th 1,000 word level was 8.95%. This indicates that "words at the lower frequency levels are less likely to become a part of the learners' productive vocabulary lexicon" (p. 135). Also, the receptive and controlled productive vocabulary sizes tended to decrease at lower frequency levels. For example, the receptive and controlled productive vocabulary sizes at the first 1,000 word level were 935 and 804 word families, while that at the 10th 1,000 word level were 397 and 36 word families.

However, they also found unexpected findings. The participants' receptive scores at the last two frequency levels, the 9th and the 10th low-frequency levels, were higher than that at the 8th level. They then examined this qualitatively and found that some of the words in these levels received more points than the others. The factors for the high scores were the influence of loanwords and frequent exposures to the words in the textbooks. One of them 'carnival' was a loanword meaning 'festival' in the Korean language. Three of the words 'octopus', 'aptitude', and 'ethnic' appeared many times in their textbooks, so the majority of the participants had many encounters to these words, recognised then and knew their meanings. They also assumed that this might be because of "the educational milieu that has put pressure on students to learn and memorise relatively lower frequency or academic words rather than the higher frequency words that are likely to be needed more for reading and writing of English general texts" (p. 136).

A qualitative analysis was also adopted in my study when I found some unexpected findings. This procedure provided more insights about my data and especially on the relationship between receptive and productive vocabulary knowledge.

With participants at a university level, Amin (2020) measured the receptive and controlled productive vocabulary sizes of EFL students in an Afghanistan university and to examine the relationship between these two groups of vocabulary. The participants were 54 fourth-year students majoring in English. He used the VST (Nation and Beglar, 2007) to measure their receptive vocabulary size and the PVLT (Laufer and Nation, 1999) to measure their controlled productive vocabulary size. Only the first ten levels in the VST were selected in his study, so there were 100 items for the participants to complete. As the controlled productive vocabulary test consists of only five levels, Amin "adopted it with some adjustments to the formula by calculating (2 × K2) + K3 + (2 × K5) + (4 × K10) = overall productive vocabulary" (p. 37). However, this formula must have been presented wrong since it seemed to estimate only nine levels, instead of ten levels.

The participants took the VST before the PVLT. Amin does not mention scoring. The results showed that the participants' receptive vocabulary size was around 4300 word families (ranging from 2,300 to 7,400 word families). It was larger than their controlled productive vocabulary size (around 3,000, ranging from 1,700 to 4,360 word families). The scores for both tests at all the frequency levels were converted into 100 points so that they could be compared. The receptive scores were higher than those for controlled productive vocabulary at all the four frequency levels. The receptive-controlled productive scores at each level were 57 vs. 37 at the 2,000-level, 48 vs. 39 at the 3,000-level, 36 vs. 25 at the 5,000-level, 24 vs. 20 at the 10,000-level. The controlled productive score at the UWL was the same as that at the 3,000-word level (39%). The scores for both tests tended to be lower at the lower frequency levels.

As the VST was used with the PVLT, it is not possible to compare their scores at the academic word level because the VST does not include the academic word level. From here, I learned that the two tests should include the same words for comparability.

However, the scores at some lower frequency levels could be higher than the scores at higher levels. For example, the average receptive score at the 4,000-word level was 54 while that at the 3,000-word level was 48, and the controlled productive score at the 3,000-word level (39%) was higher than that at the 2,000-word level (37%). This is the evidence that vocabulary acquisition does not always and completely rely on frequency levels, but the overall trend does show a relationship.

The researcher focused on the main difference between the participants' receptive vocabulary size and the controlled productive vocabulary size. He relates the findings to the idea of vocabulary threshold for performing language skills and states that "the overall receptive vocabulary of the participants is lower than the standard threshold of 5,000 words [to comprehend authentic prose] suggested by scholars in the literature" (p. 44). For the controlled productive vocabulary, he also comments that "the participants can actively participate in everyday conversation, but it would be difficult for them to write, especially academic texts, as their overall productive vocabulary size is below the threshold suggested by scholars" (p. 44). The productive vocabulary size of 2,000 words is for conversational speaking and 5,000 words for writing (Laufer, 1992; Nation, 2006; Schmitt, 2000; Webb, 2008).

L2-to-L1 translation format and L1-to-L2 translation format

An L2-to-L1 translation format

Researchers that consider receptive vocabulary knowledge as the ability to understand a given L2 word and to provide its L1 meaning use a translation format in their studies (e.g., Beaton et al., 2005; Fitzpatrick et al., 2008; Heidari, 2019; Maftoon and Sharif Haratmeh, 2013; Mondria and Wiersma, 2004; Nurweni and Read, 1999; Schmitt, 1998; Schmitt and Meara, 1997; Sydorenko, 2010; Waring, 1997b; Webb, 2005; Webb, 2007a; Webb, 2007b; Webb, 2008; Webb, 2009; Zhang and Lu, 2015; Zhong, 2016).

Like the formats mentioned earlier, this format also focuses on the knowledge of word form and meaning. Because of its simple form, it can be designed by using words from frequency list or words learnt in class. One example of this test format is Webb's (2008) 90-item receptive translation test, which he used to measure the receptive vocabulary size of Japanese students. Figure 3.7 is the sample of Webb's receptive translation test.

1.	Spring _____

Figure 3.7: Sample of Webb's (2008) receptive translation Test

The test requires a test taker to provide an L1 equivalent to the given L2 word. There are many advantages of this test format, as compared to a checklist test, the VST or

the VLT. First, it requires a test taker to provide a translation, so it can verify the test taker's comprehension of the test word. This, therefore, can help avoid guessing. Also, this test format is easier to design than a multiple-choice format or a matching format because choices or distractors should be selected carefully. The test format can also be used to test many words at a time. It also allows different answers that show a learner's knowledge of a word. Webb (2008) indicates that if a test taker supplies a lower-frequency meaning of the test word which is a polysemous word, this shows that he/she has deeper knowledge of the word. However, Webb did not mention how to grade if an L1 translation supplied by a test taker was partially correct such as closely related words but not the target translation.

An L1-to-L2 translation format

In the literature on receptive and productive vocabulary, not many studies employed an L1-to-L2 translation format as a productive test (e.g., Fitzpatrick et al., 2008; Webb, 2008). Focusing on the knowledge of form and meaning, the L1 to L2 translation test format provides L1 translations as prompts and requires a test taker to retrieve L2 word forms that match. As it is a controlled test task and requires a test taker to recall an L2 form, I considered this as another format of a controlled productive test.

To compare receptive and productive vocabulary knowledge, researchers might use the equivalent test formats: an L2-to-L1 translation test (word meaning recall) and an L1-to-L2 translation test (word form recall) (e.g., Webb, 2008). By doing so, there might be learning effects or "transfer of learning from one test to another" (Waring, 1997a, p. 56). However, Webb (2008) avoids this problem by using different test words in both translation tests. Figure 3.8 is the sample of Webb's (2008) controlled productive test; this Japanese translation means 'bubble' in English.

1.　　泡　＿＿＿＿＿＿

Figure 3.8: Sample of Webb's (2008) productive translation test

As a test taker might produce an L2 form that also matches the L1 translation given or produce an L2 word with partially correct written form, a rater might need a

sensitive scoring method which also allows a partially correct answer to get a point. Webb (2008) scored the answers both strictly and sensitively in his study. For the strict scoring system, no point would be given to a misspelled word or an incorrect word, while, with the sensitive scoring system, misspelled words which were still phonologically similar to the target words were acceptable and earned half a point.

Webb's scoring method especially a sensitive scoring system seems reasonable, for it accepts learners' partial knowledge. To know one word partially should be considered as one degree of knowing the word as well and then should not be considered as having no knowledge. What Schmitt (1998) found in his study can support this idea. One of the participants had spelled one target word wrong every time of the test, yet he knew its meaning and its associated words quite well. Also, in the case of native speakers, a young native speaker who does not know how to write a particular word probably knows its meaning, its pronunciation, and its position in a sentence quite well. Therefore, as also mentioned earlier, the sensitive scoring system will be employed in this research study.

In his study, Webb (2008) examined the sizes and relationship between receptive and productive vocabularies of EFL learners and also to examine the relationship between these two for learners with different proficiency levels. This is an experimental and within-subject research design. In his study, the participants were Japanese university students who were in their second year. The English proficiency level of the 48 students majoring in English Literature ranged from intermediate to advance, while that of the 37 students majoring in Commerce ranged from beginner to intermediate.

He used translation tests for both receptive and productive tests so that none of the test would be more demanding than the other and this would not affect the test scores. He selected 90 words for the receptive test and another 90 words for the productive test from three frequency levels of COBUILD Dictionary, which was created out of the Bank of English Corpus with over 200 million running words. Sixty words per level were selected from the 701st-1900th most frequent words (Band 1), the 1901st-3400th (Band 2), and Band 3 from the 3401st-6600th (Band 3). He did not mention the edition of the dictionary, the year of this corpus, or the criteria for frequency levels, though.

The participants' receptive vocabulary was significantly larger than their productive vocabulary for both scoring methods, at all the three levels. For example, the receptive score at Band 1 with a sensitive scoring was 95.57% (28.67 out of 30 points) but the controlled productive score was 90.90%. The differences between the receptive and productive vocabularies were bigger at lower frequency levels. For

example, the controlled productive/receptive ratios from Band 1 to Band 3 (with a strict scoring system) were 88%, 73%, and 65%. This implied that infrequent receptive words were less likely to become productive vocabulary.

The controlled productive/receptive ratios were higher when using the sensitive scoring (93%) than strict scoring (77%), "which indicates that the participants might have partial productive knowledge of L2 forms for almost all of the words known receptively" (p. 85). These high controlled productive/receptive ratios might have been a result of the use of equivalent formats which focused on only the aspects of form and meaning. Webb expected that the differences would be bigger if more aspects of vocabulary knowledge were included in the tests.

However, the controlled productive/receptive ratios between the two groups of students that calculated by using the sensitive scoring method were not statistically significant (p value of more than 0.05) and this was consistent with Waring (1997a). They were significantly different when using the strict one, with the higher ratios of the upper group for all the three levels. For example, at Band 1, the controlled productive/receptive ratio for the upper group was 93%, while that for the lower group was 84%. Webb (2008, p. 90) concludes, "This result gives support to the common assumption that receptive knowledge precedes productive knowledge (Aitchison, 1994; Channell, 1988; Melka, 1997)".

Receptive self-reported scales (VKS (categories I-IV)), cued-recall format, and a sentence writing format

Most syllabus-based or classroom experimental research studies used a receptive self-reported scales together with controlled productive test such as cued-recall format or a sentence-writing format to examine the relationship between receptive and controlled productive vocabularies as affected by different teaching methods or learning techniques (e.g., Vincy, 2020).

Receptive self-reported scales

Similar to Webb's receptive translation test, a self-reported vocabulary knowledge scale like the VKS (Paribakht and Wesche, 1993) also includes an L2-to-L1 translation task or a receptive task in its categories III and IV. The categories require a test taker to supply a translation or a synonym of a target word. The first four categories of the VKS can also be considered as a receptive test task and can be used with a productive task (e.g., Vincy, 2020). They are normally used with a productive task in category V (e.g., File and Adams, 2010; Paribakht and Wesche,

1993; Paribakht and Wesche, 1997). These scales are different from Webb's translation test because they are self-reported steps towards productive vocabulary knowledge with verification, ranging from no knowledge to being able to use the word in a sentence. Figure 3.9 shows the categories III and IV of the VKS.

Category I: I have never seen this word.

Category II: I have seen this word before, but I don't know what it means.

Category III: I have seen this word before, and I think it means _____ . (synonym/translation)

Category IV: I know this word. It means _____ . (synonym/translation)

Figure 3.9: Categories I to IV of the VKS (Paribakht and Wesche, 1993; 1997)

The scales have been used in studies on vocabulary acquisition to track vocabulary progress from a receptive stage to a productive stage (e.g., Tahmasbi and Farvardin, 2017; Vincy, 2020; Zhong, 2016). Previous studies with the idea of dichotomy scored categories III and IV quantitatively. They check whether learners know the meaning of the target word or not. The answer in this receptive translation task can be scored by using a dichotomous scoring (either correct or incorrect) or by a three-scale scoring (correct, partial correct, or incorrect) like Webb's translation test if a researcher believes that vocabulary can be partially known.

Sentence-writing task

Controlled productive vocabulary knowledge can also refer to the ability to use a word in a sentence (Laufer, 1998). Hence, a controlled productive vocabulary test can also be in a sentence-writing format like the category V of the VKS (Paribakht and Wesche, 1993; 1997). Some studies used a sentence-writing task, which requires a test taker to use a given word in a sentence, when investigating learners' controlled productive vocabulary knowledge (e.g., Tahmasbi and Farvardin, 2017; Vincy, 2020; Zhong, 2016). Figure 3.10 is what the category V of the VKS is like.

```
Category V: I can use this word in a sentence.

_____

(Write a sentence and also provide the synonym/translation in Scale IV.)
```

Figure 3.10: Category V of the VKS (Paribakht and Wesche, 1993; 1997)

Test takers write a sentence by using a given word to verify that they do not only know the meaning but also are able to use it. However, this task has a drawback. Zhong (2016) says a test taker might produce a neutral sentence – a sentence, which does not clearly show whether a test taker know the target word or not – in category V. To avoid this problem, Zhong asked the subjects in her study to write approximately 2-3 sentences with connecting words, i.e., 'because', 'so', 'although', and 'that's why'. Zhong found this technique useful and can reduce neutral sentences in her study. The dependent clauses with these connecting words could help clarify the meaning of the target word used.

Another drawback of this task is that not many target words can be tested at a time because the task requires a lot of effort from a test taker in terms of time spent and knowledge involved. Therefore, it has often been used to investigate the controlled productive knowledge of a few words taught or learnt in class, not a large number of general words from frequency lists. Despite the drawbacks, this controlled productive task (a sentence-writing task) seemed to be useful for my study. However, I decided which controlled productive test and free productive test to be used at the end of this chapter.

Using the first four categories of the VKS, the PVLT, and the fifth category (a sentence-writing task) of the VKS, Vincy (2020) examines the learners' vocabulary knowledge as affected by explicit teaching and repeated exposure. According to Vincy, these three test formats can test receptive, controlled productive and free productive vocabulary knowledge. However, the last two test formats in the test battery are controlled productive vocabulary tests, according to Laufer's (1998) explanation.

In Vincy's (2020) experiment, the participants were 62 sixth graders in India. There were equally divided into two groups (controlled and experimental groups). She created a supplementary section for vocabulary to explicitly teach the words from their English textbook to the experimental group. The words taught in class were tested before and after the treatment. As the participants were sixth graders and the

study was a classroom experiments testing words taught in class, I am not reviewing the experiments in detail. I am reviewing only the scoring system she used and the results related to the relationship between receptive and controlled productive vocabularies.

A correct answer for one level was given one point. Vincy did not assume that the correct answer at a higher level (from the controlled productive tests) also refers to the knowledge of the lower levels (from the receptive test). The results showed that receptive vocabulary scores for both the pretest and the posttest of the two groups were higher than those of the controlled productive vocabulary, with the experimental group scoring significantly higher than the control group in the posttest (p value of less than 0.001). For example, the controlled group got 10.7 points in the receptive pre-test and 4.29 points in controlled (cued-recall) productive pre-test. Vincy did not mention the maximum score.

The regression analysis also showed that the participants' receptive vocabulary knowledge influenced their productive vocabulary knowledge, with a 72% significant influence only in the case of the experimental group after the treatment. She concluded that the participants had more receptive vocabulary knowledge than the productive vocabulary knowledge, that these two kinds of knowledge are correlated, that the gap can be decreased by the explicit teaching and repeated exposure, and that using the researcher-created supplementary could improve the learners' vocabulary learning.

Matching format and a sentence-writing format

Oberg (2012) conducted a syllabus-based experimental research study, based on the idea that receptive and productive vocabularies are not the same groups of words in learners' mental lexicon and that receptive vocabulary will become productive vocabulary. He particularly aimed to find out the threshold where receptive vocabulary became controlled productive vocabulary. He writes:

> "[A]t which point – that is, how many sessions – a significant difference may emerge indicating a sudden and noticeable gain in productive acquisition of the material, i.e., a memory threshold being crossed from a receptive knowledge to a productive one" (Oberg, 2012, p. 30).

His study was conducted in Japan. Fifty-eight first-year non-English majors participating in his study. The participants were asked to make their own picture cards for all the 15 target vocabulary items from their textbook. There were two sections of the test: the receptive section (a matching task) and the productive

section (a sentence-writing task). The same test was administered as the pretest and the posttest. The students played the cards and recalled the vocabulary items from the cards and from the context during the treatment.

The test scores showed that the receptive and the productive vocabulary knowledge appeared to increase, with the higher receptive knowledge scores. For example, after the first treatment session, the receptive score was 96.42% but the controlled productive score was only 11.85%. The receptive scores slightly increased and nearly reached the ceiling for all the ten sessions. The receptive score showed a significant different only the first after-treatment session (p value of less than 0.001), while the productive scores had a steady and significant increase for every session, except after the 9th session.

> "However, the results of both analytical tests used did not support the existence of such a threshold…The consistent occurrence of significant differences in the statistical analysis of the data makes it difficult to pinpoint just where a cut-off line could be placed…Moreover, the lack of any sudden and noticeable jump in the mean score themselves appears to support this" (Oberg, 2012, pp. 30-32).

Hence, Oberg concluded that there is no such threshold where receptive vocabulary becomes productive vocabulary and that the productive vocabulary gradually develops.

3.2.5   Research on receptive and free productive vocabularies

In this section, I review how scholars compared learners' receptive and free productive vocabulary knowledge.

Multiple-choice format and essay-writing task

Šišková (2016) studied the relationship between EFL learners' receptive vocabulary size and free productive vocabulary knowledge in terms of lexical richness. Despite that, I review her study because the results of free productive vocabulary are related to frequency levels, a factor of vocabulary knowledge in my study.

Her research study was designed to be similar to that by Laufer (1998) and Laufer and Paribakht (1998) but with more recent tests and measures. The participants were 119 Salvic first-year non-English-major university students, whose English proficiency level was at B2 of CEFR or higher. The tests that were used in the study were Nation and Beglar's (2007) VST – the 140-item version. To examine the

learners' free productive vocabulary, Šišková asked her participants to write a short story from pictures (Plauen, 1996). She decided to use pictures to control the participants' writing was because the participants would not need to think of interesting content but could only focus on the language to be used in their story. However, this was not used in the present study because a narrative story might elicit only high-frequency words that would be limited to only what the test takers see in the pictures. This might be suitable for students at lower educational level.

Lexical richness in Šišková's study covered three areas: lexical diversity, lexical sophistication, and lexical density. The results showed that there was a moderate relationship between the participants' receptive vocabulary size and the lexical diversity in their writings. This implies that when they knew more words receptively, they tended to know more words productively and use various words in their writing. The results also showed that the participants' receptive vocabulary knowledge was not much correlated to the proportion of advanced words used in their writings. This might be a normal distribution of any texts that there is always a small proportion of low-frequency words, so it does not depend on how big one's receptive vocabulary is. Also, learners probably knew many low-frequency words receptively but had partial knowledge of using them; therefore, they would try to avoid them in their writings.

However, Šišková acknowledges that not including a controlled productive test was the limitation of the study; otherwise, the data could have been able to tell whether the learners would use more of low-frequency words if forced to. The findings in Šišková's study showed that there was no relationship between their receptive vocabulary size and lexical density. That is, knowing a lot more words does not guarantee the use of a lot of content words. The learners might have used a lot of pronouns or proper nouns to refer to nouns in the pictures, together with various prepositions or conjunctions.

3.2.6 Research on receptive, controlled productive, and free productive vocabularies

The section reviews how previous studies examined the relationship among the three groups of vocabulary (receptive, controlled productive, and free productive vocabularies) and presents what they found (e.g., Laufer, 1998; Laufer and Paribakht, 1998).

Matching format, cued-recall format, and essay-writing task

Although Laufer (1998) acknowledges that depth is also significant to understand the overall lexical competence of L2 learners, she says that vocabulary breadth is the most important and most correlates learners' language performance. Noting that the ability to recall words and the ability to use words should be distinguished, Laufer (1998) then explored the relationship among receptive, controlled productive, and free productive vocabularies of L2 learners of English and the vocabulary development cross-sectionally. Her study was conducted in Israel with two groups of high-school students (tenth graders and eleventh graders). These two groups of participants were homogeneous as they spoke the same native language (Hebrew), came from similar socioeconomic status and received similar input from the same team of teachers at the school; only their year of study was different.

These two groups of learners sat three tests during three different classes in one week. She used three tests: VLT to measure receptive vocabulary (Nation, 1983; Nation, 1990), PVLT to measure controlled productive vocabulary (Laufer and Nation, 1999), and a composition writing task to measure free productive vocabulary (Laufer and Nation, 1995). Laufer included only three frequency levels (the 2,000-, the 3,000-, the 5,000-word levels) and the academic word level (UWL) in the first two tests (the eighteen items from each of the frequency level represent 1,000 word families). Hence, the total score for all the four groups of test words was 72 points. Regarding the analysis, dichotomous scoring was used for the first two tests and lexical frequency profile was used to analyse words used in the learners' free writings.

The results showed that the participants' receptive vocabulary grew considerably at all the frequency levels over one year of study (from 24 points to 48 points of the total score; from 1900 to 3500 word families out of 4,000 word families), or around 84% (increasing 1,600 word families from 1,900 word families). Their controlled productive vocabulary also developed, approximately 50% larger than those in the lower grade (from 1,700 to 2,550 word families out of 4,000 word families or 850 more word families). The correlation of the receptive and controlled productive vocabulary test scores was high (0.67 for the tenth graders and 0.78 for the eleventh graders).

For their free productive vocabulary, the eleventh graders used more of the second 1,000 GSL words and the academic words than the tenth graders in their compositions. However, the number of words that these two groups of learners produced did not differ much at all the levels analysed by the profiler. The participants' free productive vocabulary knowledge does not correlate with the

receptive and the controlled productive knowledge (r values ranging from 0.07 to 0.25). This indicates that learners who could recognise more words than other learners and produce them if forced to were not necessarily those who would use more infrequent vocabulary in free expression" (p. 265).

For example, while the eleventh graders had larger receptive and controlled productive vocabularies at the 3,000-word level and the 5,000-word level than the tenth graders (e.g. receptive scores at the 3,000-word level: 75.78% vs. 44.17%, and controlled productive scores at the 5,000-word level: 51.67% vs. 34.94%), they tended to use fewer words from 'Not-in-list' (not in the 2,000 or in the UWL) than the tenth graders (approximately 3.5 vs. 4.3 word families per composition, respectively). Therefore, the number of words used at free will does not seem to be related to learners' vocabulary size.

Laufer also commented about the difficulty in measuring free productive vocabulary. She says, "it was impossible to calculate the ratio between free active vocabulary and the other two types (It is unlikely that we can devise a test measuring how many words a person can use at free will, unless his vocabulary is very limited)" (p. 265). Accordingly, finding a suitable way to explore the relationship among these three groups of vocabulary is essential.

Similarly, Laufer and Paribakht (1998) studied the relationship among receptive, controlled productive, and free productive vocabularies, but the participants included both high-school students and university students. They were from an EFL context (Israel) and an ESL context (Canada). The study aimed to compare these three kinds of vocabulary knowledge of the students from the two contexts at all the five vocabulary levels, to compare them across proficiency groups and to examine how the relationship among them changed when the ESL students stayed in Canada longer.

Laufer and Paribakht (1998) used the same tests as Laufer (1998), and the procedure was similar, with the Israeli students spending longer time than those studying in Canada. A dichotomous scoring was used to check the answers in the VLT and the PVLT, while the LFP was used to analyse the proportion of infrequent words (words beyond the 2,000-word level) used in the learners' 300-word compositions.

The results showed that receptive vocabulary was larger than controlled productive vocabulary in both the learning contexts. For example, the average receptive and controlled productive scores were 52.89% and 40.78% in the EFL context, and 74.11% and 46.11% in the ESL context. While the ESL students had larger vocabularies than the EFL students, their controlled productive/receptive ratio was

lower than the EFL students (61.7% vs. 77%). This implies that the receptive – controlled productive vocabularies was wider in the ESL context. This might be because the ESL students had much larger receptive vocabulary and their controlled productive developed more slowly.

When Laufer and Paribakht examined the data in more detail (into five frequency levels), they found that receptive and controlled productive vocabularies were affected by frequency levels. Learners could recall or understand fewer words when the word frequency levels were lower, and the receptive-productive gap was smaller at higher frequency levels. This implies that receptive vocabulary at higher frequency levels is likely to pass to productive vocabulary. This might be because it has been encountered and used often by the learners.

Also, the ESL students' receptive vocabulary was larger than that of the EFL students at all the five levels, and their controlled productive vocabulary was larger at only lower frequency levels and the receptive-productive gap was wider at all the levels. When analysing their correlation, Laufer and Paribakht (1998) showed that these vocabularies were intercorrelated. This implies that the growth of one vocabulary leads to the growth of the other.

The researchers acknowledged that it was impossible to compare the scores of these three vocabularies. They write, "[W]e measured the free active (FA) vocabulary in terms of relative proportions of frequent and infrequent vocabulary produced in essays, not in terms of knowledge of specific items, we cannot calculate the ratio between FA vocabulary and the other 2 types of knowledge" (p. 378).

When dividing the students into three proficiency groups based on their receptive scores, the researchers found that the advanced students had more controlled productive and free productive vocabulary knowledge than the low-proficiency group. For example, the average controlled productive scores for the advanced group and the low-proficiency group were 57.33% and 23.33%, the average proportion of beyond 2,000 words for these two groups were 15.80% and 6.80%.

The controlled productive/receptive ratio was lower when the students were more proficient or when they acquired more receptive words. For example, the ratio of the advanced EFL students was 71.40% but that of the low-proficiency group was 91.10%. This implied the number of controlled productive words went up regularly as students became more proficient but that the receptive vocabulary went up more quickly, so the receptive-controlled productive gap became wider.

From reviewing the studies comparing receptive with free productive vocabularies or studies comparing receptive, controlled productive, with free productive

vocabularies, I learnt that it seems impossible to measure the size of someone's free vocabulary. Hence, the proportion of the participants' free vocabulary used in compositions was used instead. I also learnt that the data on free productive vocabulary of these studies was not parallel to the data on receptive and controlled productive vocabularies. Sometimes it was analysed in terms of lexical richness, and sometimes it was profiled into fewer groups than the data on receptive vocabulary or on controlled productive vocabularies. For instance, the VLT and the PVLT involve vocabularies from the 2,000-, the 3,000-, the 5,000-, and the 10,000-word levels and the academic word level, while free vocabulary from participants' compositions were profiled into four groups (the 1,000- and the 2,000-word levels, the academic word level, and the not-in-list words). Words at the 5,000- and the 10,000-word levels cannot be compared. I then decided not to use Cobb's VocabProfile, but I decided which programme to be used and which frequency levels to be included at the end of this chapter.

3.2.7   Research on the relationship among receptive and productive vocabulary aspects and their acquisition

The focal receptive vocabulary aspects examined by the studies that have been reviewed earlier from Section 3.2.1 to Section 3.2.6 are the aspects of form and meaning. This section reviews studies that examined the relationship among receptive aspects and productive vocabulary use (e.g., Zhong, 2016) and how vocabulary aspects develop over time (e.g., Amy, 2006; Schmitt, 1998; Schmitt and Meara, 1997) or after some treatments (e.g., Min, 2008; Paribakht and Wesche, 1997; Webb, 2005; Webb, 2007a; Webb, 2007b; Webb, 2009). I review these, focusing on methodology and findings.

Self-reported scales (VKS (Categories I-V)), tasks for aspects

Based on Nation's (2001) multi-aspect framework, Zhong (2016) studied the relationship among five receptive aspects (meaning, form, word class, collocation, and association) and the relationship between these receptive aspects and controlled productive vocabulary use. The participants were a large group of secondary Chinese students (620 of 8th graders) from two schools. The target words consisted of 26 words (nouns, verbs, and adjectives) selected from the textbook in conjunction with the first and the second 1,000 levels in the GSL and the AWL.

Zhong revised many test tasks and included them into a test battery. The multiple-choice format of form recognition test (Webb, 2008) was employed; it requires a test taker to choose a correct spelling. The format of Read's (1998) WAT test (a test of collocation and word association recognition) was also used. The test takers were supposed to choose two correct collocates and two correct associated words among the eight choices. The collocates and the semantic associates were sampled for the test items from the (Nation, 2004) and a thesaurus. The fill-in-the-table format (Ishii and Schmitt, 2009) was used to test the knowledge of word class.

One point was awarded for one correct answer in all these tests. Based on the idea of receptive-productive continuum, all the five categories of the VKS (Wesche and Paribakht, 1996) were used by Zhong (2016). The format of the first four categories of the VKS was used as the receptive meaning test, requiring the test takers to rate how much they knew the target word and to supply a correct meaning if they believed they knew it. The first four categories were adapted in terms of wording. For example, the category III "I have seen this word before, and I think it means _____ . (synonym or translation)" (Paribakht and Wesche, 1997, p. 180) was revised into "I have seen or heard this word before, and know its meaning a little. What I know about it is_____ (in Chinese)" (Zhong, 2016, p. 16). The scoring of the VKS was also adapted by including partial knowledge score and the maximum score was three points: three points for a correct translation, two points for a partially correct translation, one for a wrong translation but being familiar with the word, zero point for not being familiar with the target word.

The category V of the VKS was used separately as the test for controlled productive vocabulary use. To avoid a neutral sentence, Zhong asked the participants to write two to three coherent sentences by using the target word and one of the conjunctions: 'because', 'so', 'although', 'that's why'. Three points are for using the word correctly in the sentences, two points for the word use with minor grammatical mistakes, one point for the word still used with correct meaning, and no point for the word use in the wrong context.

The participants performed the form recognition test, the meaning comprehension test and the controlled productive vocabulary use test on Day 1, and the rest on Day 2 to avoid the test fatigue effect. Each of the 26 test words was tested with these test tasks. Another rater was trained before rating the answers of the VKS (categories I-V). The inter-rater reliability was analysed, and the statistics showed that it was strongly reliable (r_s = 0.975 for VKS (categories I-IV) and r_s = 0.951 for VKS (category V)).

The results showed that the participants' scored the highest for the form recognition test (81.71%), followed by the meaning test (64.44%), and the lowest for the sentence writing (controlled productive vocabulary use) (43.86%). Zhong explored the relationship among the receptive aspects by using Pearson correlations. She found that all the receptive aspects tested in her study were positively correlated and that the relationship between the knowledge of collocation and that of association was the strongest (r = 0.760), followed by that between the knowledge of form and meaning (r = 0.639), and the weakest one was between the knowledge of form and collocation (r = 0.314).

When examining the relationship between these five receptive aspects and the controlled productive vocabulary use, she found that the knowledge of meaning, word class and form were strongly correlated with the controlled productive vocabulary use (r = 0.858, 0.630, 0.605, respectively). The receptive knowledge aspect that had the weakest correlation with the controlled productive use was collocation (r = 0.459). This shows that the knowledge of form and the knowledge of meaning are the fundamentals of word knowledge for controlled productive use of a word. The degrees at which meaning, word class, form, association and collocation were correlated to the controlled productive vocabulary use somehow show the order of the knowledge aspects to be used in writing, consistent with what Jiang (2000) suggested. That is, learners think of the concept/meaning before choosing the right form (and word class) before combining the form to its collocates (and/or associates).

She investigated further the predictive power of form and meaning to the controlled productive use by using hierarchical multiple regression. The results showed that form and meaning could predict 74.1% of the controlled productive ability. Zhong concluded that these aspects were interrelated, but the knowledge of form and meaning is the fundamental aspects for productive use. She writes:

> "Capturing form and meaning only was able gauge approximately two thirds of comprehensive productive vocabulary knowledge. In other words, assessing meaning and form only could be an effective method for estimating EFL learners' productive vocabulary ability" (Zhong, 2016, p. 11).

I view that the VKS (category V) seems valid to test learners' ability to use words in a sentence level because it has verification. The participants can really produce it (not too easy for university students) in the form that they need to regularly do as English-major students. Moreover, it can be used with VKS (categories I-IV) – a receptive test (the test of form and meaning) because the same words can be tested with both formats and the knowledge of form and meaning is the most fundamental aspect for

productive use. I acknowledge that receptive knowledge is not only related to the knowledge of form and meaning, though. Using the same target words for all the tests makes the scores more comparable. Also, using this test format would allow me to score both with the continuum and the dichotomy approaches in order to investigate the relationship between receptive and productive vocabulary knowledge without presuming it is either continuum or dichotomous at first. Hence, the VKS was chosen in my study.

In the dichotomy approach, I can score the VKS (categories I-IV) separately from the VKS (category V) and compare the receptive scores to the controlled productive scores. I decided to use a multi-scale scoring for both the receptive answers and the controlled productive answers, similar to Zhong (2016), because it allows partial vocabulary knowledge. However, for comparability, the same three-scale scoring system will be used. Two points for a correct answer, one point for a partially correct answer, and no point for an incorrect answer.

In the continuum approach, I decided to use the 5 scores applied by Paribakht and Wesche (1993; 1997) as it was tested as a valid and reliable test in their studies. Like Zhong (2016), I chose to adapt the category V by asking the participants to use a conjunction such as 'because' and 'so' when producing a sentence as well to avoid neutral sentences. Two English native speakers rated the sentences, and inter-rater reliability was analysed. However, the number of sentences to produce for each target word, the number of target words and the number of frequency levels to be included in the test battery were mentioned in the conclusion, and these needed to be piloted.

Interviews about aspects

Schmitt (1998) studied the acquisition of vocabulary. Based on the idea that vocabulary knowledge consists of many aspects, Schmitt investigated how individual words are acquired in terms of the four aspects of word knowledge (written form, word association, grammatical information (parts of speech and derived forms) and meaning (senses)). He focused on how these aspects change over time but not on the factors of the change.

To explore vocabulary acquisition in more detail, instead of using paper-based test battery, Schmitt (1998) used an interview method three times with six-month intervals in his study with a longitudinal multi-case research design. The first interview was conducted at the beginning of an academic year; the second and the third interviews were conducted with a half-term interval. The participants in his study

were four postgraduates who attended a pre-sessional course in a British university. They were from different majors and from four different countries: Lithuania, Korea, Taiwan, and India. However, the Korean participant dropped off, so only the data from the other three were analysed.

Since many aspects were involved and the focus was on individual words, he did not include many target words. He selected 18 words from the UWL and the Brown corpus. After piloted with EFL learners similar to the participants, only eleven words were included in the main study, two unknown words, four relatively well-known words, and five words in between. These words were selected because they were likely to be exposed to by EFL university learners and the words were polysemous. An interview method that deals with many vocabulary aspects at a time was not used in the present study because not many test words can be included.

Also, as it is widely accepted that vocabulary acquisition is incremental and moves from no knowledge to full native-like mastery, Schmitt then scored each aspect of the vocabulary knowledge with a multi-scale scoring system. For the knowledge of spelling, no point for no knowledge of how a target word was spelled, one point for knowing its initial letter, two points for the answer that was phonological correct, and three points for correct spelling.

For the knowledge of association, Schmitt asked the participants to give three responds and compared the participants' responds to a native speaker's norming list (NS norming list – a list showing words associates produced by native speakers). No point was given when the participants did not supply any associated words to a stimulus, one point to a word matching an infrequent word in the NS norming list, two points to a typical word in the norming list, and three points to a respond matching any words in the top half in the norming list.

For the knowledge of grammatical knowledge, Schmitt asked his participants whether they knew the part of speech of the target word and asked them to give him the form of such word in the other three parts pf speech. The target words were content words, so the parts of speech were noun, verb, adjective, and adverb. Knowing one part of speech was worth one point.

For the meaning, Schmitt asked his participants to give all the meaning senses they knew of a target word. If the participant were able to give a meaning without being hinted or hearing a prompt, two points were given. This ability is considered by Schmitt as the productive knowledge of meaning. If the participant could give a word meaning because of a prompt, then one point would be given. If the target word was unknown, then no point was given. Schmitt checked all the meanings by checking them against three dictionaries: the Longman Dictionary of English Language and

Culture (1992), the Oxford Advanced Learner's Dictionary (1995), and Webster's Ninth New Collegiate Dictionary (1987).

The results on the learners' knowledge of meaning showed that it was not easy for the participants to pick up many senses of words longitudinally. Their knowledge of most of the test words remained static over one academic year. The knowledge of some words developed, and that of fewer words decreased. This implied that once a learner learns word meaning senses it is unlikely for them to forget.

Regarding the knowledge of written form, the participants did not seem to have trouble spelling. Nevertheless, Schmitt noticed that one participant kept supplying the wrong spelling 'ellumninate' or 'eluminate' for the target word 'illuminate', even though for this word she had a good knowledge of its associated words, grammar, and meaning senses. According to the scoring system that Schmitt adopted, this misspelled word 'eluminate' should have been considered as a partial knowledge of form and been given two points. Therefore, "[h]aving a good understanding of other types of word knowledge does not necessarily mean that one will know how to spell a word correctly" (Schmitt, 1998, p. 303). This is consistent with what Webb (2007b) suggests about the development of aspects; different aspects develop at different rate.

Next, the learners' knowledge of association seemed to progress upward over time. Schmitt noticed that "learners are very unlikely to be able to give [native-like] associations unless they know at least one meaning sense of a word – unless that word is part of a known phrase" (p. 303). Therefore, the knowledge of a meaning sense is directly related to the knowledge of association.

The participants' knowledge of grammatical information (part of speech and derived forms) was rather unpredictable. The trend was all different for these three learners. However, the knowledge of grammar of one of the participants who looked up the target words in dictionaries during this academic year had a steady increase. Hence, explicitly learning by using dictionaries must have helped improve his knowledge.

Schmitt also investigated the relationship among the four aspects and found that they seemed to increase when the other increased. As Schmitt hypothesised at first that there might be a developmental hierarchy of vocabulary knowledge aspects, he then conducted the Guttman procedure (Hatch and Lazaraton, 1991) to explore the possibility of the implication scale, he found "no evidence of a developmental hierarchy for word knowledge types" (p. 309). It can be concluded from Schmitt's study that vocabulary aspects are interrelated; however, their increases do not need to be at the same pace. Also, one aspect does not always develop before the others. These findings seem to support Melka's (1997) argument about the acquisition of

vocabulary knowledge aspects. She writes, "Knowing a word is not an all-or-nothing proposition; some aspects may have become productive, while others remain at the receptive level" (Melka, 1997, p. 87).

Pellicer-Sánchez (2019) calls for the replications of a longitudinal study that focusses on the depth of vocabulary knowledge like Schmitt's (1998) study. Some of the modifications she suggested include 1) conducting the research with a larger group of participants for both descriptive and inferential statistical analysis, 2) the addition of a second rater for the objectivity of scoring methods, and 3) the inclusion of a structured interview for the information about the participants' vocabulary learning strategies to verify what Schmitt observed. For my study, I conducted the present study with a large group of participants, included two raters and included information about vocabulary learning in an out-of-classroom context, which might be able to explain some unexpected findings.

Vocabulary Level Test, recognition and recall tests for different aspects

Another example of multicomponent research studies was conducted by González-Fernández and Schmitt (2020). Their aim was to explore whether components of vocabulary knowledge are interrelated, how knowledge of individual words develops and how to best conceptualise the relationships between the components. According to Nation's (2013) taxonomy, the aspects or the components (as called by the authors) that were explored in this research were the knowledge of form and meaning, word parts (derivatives and parts of speech), concept and referent (multiple meanings or senses), and collocation. The participants were 144 Spanish students with different proficiency levels. They took the VLT (the 2,000-, the 3,000-, the 5,000-, and the 10,000-word levels) for their vocabulary size. They also sat 8 tests (four aspects x recognition and recall tests) for the 20 target words carefully selected by the authors. To avoid the learning effect among the tests, the authors placed the tests in the following order: "form-meaning link form recall → VLT 5K/3K → form-meaning link meaning recognition → derivatives form recall → derivatives form recognition → multiple-meanings recall → collocate form recall → VLT 10K/2K → multiple-meanings recognition → collocate form recognition" (González-Fernández and Schmitt, 2020, p. 490). The correlation analysis showed that the components are interrelated and related to vocabulary size. The implicational scaling showed the difficulty order of the components in general (from easier to more difficult):

> "Form-Meaning link meaning recognition > Collocate form recognition >
> Multiple-Meanings meaning recognition > Derivative form recognition >

> Collocate form recall > Form-Meaning link form recall > Derivative form recall
> > Multiple-Meanings recall" (González-Fernández and Schmitt, 2020, p. 493).

González-Fernández and Schmitt (2020, p. 500) added that recognition comes before recall and "this does not mean that they are always strictly learned in sequence" because it can depend on different learners, different contexts, different individual words, etc., and. these aspects only include four components. Therefore, they recommended future research to explore the order of other aspects and with EFL learners in other contexts.

They suggested that the major distinction of vocabulary knowledge should be between recognition at the receptive end on the continuum and recall at the productive end. Recognition of all the aspects explored in the study seemed to happen before the aspects recall. Hence, "this conceptualisation suggests that perhaps the focus of pedagogy should be shifted towards pushing learners' knowledge from receptive towards productive mastery" (González-Fernández and Schmitt, 2020, p. 501).

Multiple-choice format, translation format, recalling/supplying aspects, sentence-writing task

Adopting Nation's (2001) taxonomy, Webb (2007b) also studied the acquisition of receptive and productive vocabulary aspects over time (the number of encounters: 1, 3, 7, and 10 encounters) to also find threshold or at which encounter time where receptive knowledge would become productive knowledge.

The participants were 121 second-year university students in Japan, who passed 80% of the 2,000-level of the VLT. Their major was not mentioned. They were divided into five groups and assigned as one control group and four experimental groups (groups with different encounters called as E1, E3, E7 and E10). Only ten high-frequency words were the target words (6 nouns and 4 verbs – as proportional frequency of occurrence (Kučera and Francis, 1967)), but they were replaced with similar non-words to make sure that the participants did not know the words at the beginning of the study. The participants in the experimental groups needed to read ten sentences with ten different target words (non-words). The sentences were selected from the Oxford Bookword Series and similar in length and included only high-frequency words. All the sentences were rated on their degree of informativity by Webb and another native speaker. The less informative sentences were shown in later encounters. After each of the encounters, the participants sat ten test tasks. The control group did not read these ten sentences, nor did they sat any tests, while

the E10 group needed to read these sentences ten times and took the tests ten times (every time after each encounter). They could spend time as much as they need to finish the test battery each time.

In my study, the participants were given as much time as they needed to finish the task so that the participants could supply the best answers without being affected by time limitation. However, the tests needed to be piloted to ensure that they would not be too long for them to complete. Otherwise, fatigue would also affect the data.

Webb (2007b) designed ten test tasks to measure both the receptive and productive knowledge of the following five aspects: orthography (written form), meaning and form, grammatical function, association, and syntax (collocation). The receptive tests for the five aspects were all in multiple-choice formats to investigate the ability to recognise each aspect. For example, the receptive test for association required a test taker to choose the correct associate of the target word. In the productive test tasks, the participants were asked to supply a correct spelling, a grammatical sentence using the target word, a correct associate, a word with a syntactic relationship with the target word. However, a productive test for meaning and form was not used, and a receptive recall meaning test (or a translation test) was used instead. Therefore, there are two tasks for the receptive knowledge of meaning.

To avoid the learning effect, Webb was careful about the order of the tests. The first test to be completed by the participants was the productive orthography test (a spelling test), which the participants heard the target non-words twice and had ten seconds to write it with a correct spelling. The second test was the receptive orthography test (a task of choosing a correct spelling), followed by the receptive recall meaning test (a translation test), the productive grammatical functions test (a sentence writing task), the productive syntax test (a task of supplying one collocation), the productive association test (a task of supplying a correct associate), the receptive grammatical functions (a task of choosing a correct grammatical sentence), the receptive syntax task (a task of choosing a correct collocation), the receptive association test (a task of choosing a correct associate), and the receptive meaning and form test (a task of choosing a correct meaning).

The statistical results showed that "gains in all aspects of knowledge tended to increase as the number of presentations increased" (p. 59). Some of the gains were significant and some were not, though. However, the gains would be all significant or sizeable after the tenth encounters even though the sentences they read was less informative.

After the first encounter with the words, the participants gained some knowledge of the words, at different degrees for different aspects. They scored the most in the

receptive orthography test (M = 6.7/10), followed by the receptive meaning test (M = 5.78/10), the receptive grammatical functions test (M = 5.65/10), the receptive associations test (M = 4.78/10), the receptive syntax test (M = 4/10), the productive grammatical functions (M = 1.09), the productive syntax test (M = 0.91/10), and the receptive recall meaning test (M = 0.35). The gain in receptive knowledge of orthography was also significant after the encounter 3 and its score was the highest when tested after 3, 7, and 10 encounters, so "spelling is likely to be the first knowledge type acquired" (p. 60).

While the average score of the receptive meaning and form was high after all encounters; unexpectedly, the score of the receptive recall meaning test was relatively low. Webb then concluded:

> "Since partial gains were made for all aspects despite relatively little knowledge of meaning, acquiring knowledge of association, syntax, spelling, and grammatical functions may not be dependent on the acquisition of meaning. This is certainly the case in L1. Adult native speakers occasionally demonstrate productive knowledge of a word with grammatical, orthographic, and syntactic accuracy but without accuracy of meaning" (Webb, 2007b, p. 61).

Since the gain in knowledge of different aspects became significant at different encounters, Webb also concluded that it could not guarantee at which encounter would ensure vocabulary learning as acquiring a word seemed to be influenced by other variables.

3.2.8 Research on vocabulary acquisition from no knowledge to ability to use the word (using self-reported knowledge scales/states)

Many other researchers also explored vocabulary acquisition by using a classroom experiment to compare two or more different teaching methods or learning techniques, using a pre-posttest design to examine the vocabulary gain and focusing on vocabulary learnt in class. Previous studies of word acquisition employed the test formats like the VLT and the PVLT (e.g., Aizawa et al., 2003), a sentence-writing task (e.g., Walters and Bozkurt, 2009), Paribakht and Wesche's (1997) VKS (e.g., Bahramlou and Esmaeili, 2019; Bakla and Çekic, 2017; File and Adams, 2010), the adapted VKS (e.g., Bao, 2015; Min, 2008), or the State Rating Task (SRT) – a self-reported states without knowledge verification (e.g., Dabaghi and Rafiee, 2012; Waring, 2000). As the aim of my study is not comparing different teaching methods nor focusing on vocabulary learnt in class, I review only a few studies for the idea of

how the researchers analyse the vocabulary scores using these scales/states but do not address much on the study results because it cannot be compared to that of my study.

Self-reported vocabulary knowledge scales with verification

Based on the idea that the relationship between receptive and productive vocabularies moves along a continuum, the scales range from no knowledge to being able to use it in a sentence. The VKS was designed and validated by Paribakht and Wesche (1993).

> "[It] uses a 5-point scale combining self-report and performance items to elicit self-perceived and demonstrated knowledge of specific words in written form. The scale ratings range from total unfamiliarity, through recognition of the word and some idea of its meaning, to the ability to use the word with grammatical and semantic accuracy in a sentence. … The VKS scale is not intended to go beyond the ability to use the words in initial contextualized production, for example, to tap knowledge of additional word meanings, or derivational, paradigmatic, semantic and other relationships and networks" (Paribakht and Wesche, 1993, p. 179).

Figure 3.11 is the sample of the VKS (categories I-V). The maximum score is five and the scoring system is shown in Figure 3.12.

Many studies used Paribakht and Wesche's (1993; 1997) VKS to track learners' vocabulary development (e.g., Bahramlou and Esmaeili, 2019; Bakla and Çekic, 2017; Bao, 2015; Dubiner, 2017; File and Adams, 2010; Hashemi and Gowdasiaei, 2005; Helms-Park and Perhan, 2016; Mahdavy, 2011; Min, 2008; Mondria and Wiersma, 2004; Mousavi and Gholami, 2014; Rashtchi and Aghili, 2014; Smith et al., 2013; Sun, 2017; Tahmasbi and Farvardin, 2017; Varandi and Faezi, 2013; Vincy, 2020; Wesche and Paribakht, 1996; Yang et al., 2017; Zhao and Macaro, 2016).

Self-report
categories

I   I don't remember having seen this word
before.

II   I have seen this word before, but I don't
know what it means.

III   I have seen this word before, and I think
it means _____. (synonym or
translation)

IV   I know this word. It means _____.
(synonym or translation)

V   I can use this word in a sentence: ____.
(Write a sentence.)
(If you do this section, please also do
Section IV.)

Figure 3.11: Paribakht and Wesche's (1993; 1997) VKS (Categories I to V)

Figure 3.12: Scoring system of the VKS by Paribakht and Wesche (1993; 1997)

Paribakht and Wesche (1993; 1997) claim that the VKS can track only a small
change in vocabulary knowledge. This test has been used as an achievement test in

a pre-posttest research design within a short period of time such as within one term or one academic year to investigate whether a particular teaching/learning method is effective in increasing learners' vocabulary knowledge or not. It can also be used when researchers aim to compare two or more of teaching methods or learning techniques. Hence, the target words that have been tested were words taught or learnt in class. This means not many words were tested in one study. Concerning scoring, Paribakht and Wesche (1997) state that this can be done both quantitatively and qualitatively.

Quantitatively, Paribakht and Wesche suggest grouping the answers with scoring categories I and II as "unknown vocabulary" and those with scoring categories III, IV, and V as "known vocabulary". The two scores can then be compared to see how many words become known.

Alternatively, the scores of correct translations/synonyms can be considered as the receptive scores and can be compared with those of the correct word usage in the sentence task – the controlled productive scores, as presented earlier in Section 0.

Qualitatively, the percentages of each score were compared and plotted into graphs to examine how the knowledge of each word had changed after the treatments. This follows the idea of vocabulary knowledge moving along a continuum and tries to capture a small change of vocabulary knowledge.

However, some scholars do not think that the VKS is suitable to track small changes in vocabulary development (e.g., Meara, 1996b; Schmitt, 2010; Waring, 1999; 2000). Meara (1996b) does not agree with using the VKS to capture learners' vocabulary knowledge since he believes that the VKS is suitable for exploring how individual words become fully integrated in learners' mental lexicon but does not provide deeper understanding of lexicons as a whole, which he thinks is more important in the area of second language vocabulary acquisition. However, the VKS format seems flexible, that is, it can be used with different groups of vocabulary such as (words from high-, mid-, low-frequency levels) for a holistic picture of vocabulary knowledge development as well.

Paribakht and Wesche (1997) examined the acquisition of some words learnt in class to compare the effectiveness of two instructional methods: reading comprehension plus vocabulary exercises (RP – the "Reading Plus" method) versus reading additional texts (RO – the "Reading Only" method). The participants were 38 students attending the four-semester English course in a university in Canada. Their first languages were different such as French, Arabic, and Chinese. The same group of participants studied in both ways, so the texts and the target words for each instructional method needed to be different. The pre-posttreatment design was

adopted. To be able to track small progress of vocabulary acquisition, Paribakht and Wesche used the VKS.

The themes were selected based on the teachers' experience, the students' background knowledge, interest, and text difficulty. The target words chosen included content words (nouns and verbs), and discourse connectives. Two themes (two texts for each theme) were chosen for the classroom experiment. In the RP method, the participants read selected texts, answered comprehension questions, and did vocabulary exercises (ranging from controlled exercises to free/production exercises). In the RO method, the participants did not do the vocabulary exercises but read more texts with the same themes to be exposed to the same target words again.

The researchers examined the growth both quantitatively and qualitatively. Quantitatively, the scores of unknown words and known words were compared. The quantitative analysis and the t-tests showed that the scores of the target words taught in the RP instruction was significantly greater than those taught in the RO; the RP instruction was more effective.

For example, in pre-test, the scores of the nouns taught by the two treatments were roughly equal; the scores of the RP nouns were 75% (unknown categories) and 25% (known categories), and those of the RO nouns were 74% and 26%. After the treatment, more nouns taught by the RP method moved in known categories. The post-test scores of the RP nouns were 35% (unknown categories) and 65% (known categories), while those of the RO nouns were 50% and 50%, respectively.

Qualitatively, after the treatment, they found that the scores in scoring category I drastically decreased for both treatments, many of the unknown words in the RO method move to category II (word form recognition) and much more words appeared to move to higher scoring categories for the RP method. The pre-test scores for the RP nouns in the categories I, II, III, IV, and V were 21%, 54%, 10%, 5%, and 10%, and after the treatment the scores were 3%, 32%, 19%, 11%, and 35%. The scores moved to later categories and the highest score (35%) was at category V (being able to use the word in a sentence level). The pre-test scores for the RO nouns were 29%, 45%, 10%, 3%, and 13%, while the post-test scores were 4%, 46%, 14%, 13%, 23%. The scores also moved to later categories, but most of them (46%) were still at the category II (having seen the words but having no idea what they mean). This implied that the reading plus vocabulary exercise could help enhance learners' vocabulary acquisition better than reading texts only.

File and Adams (2010) also used the VKS (Paribakht and Wesche, 1997) to track the participants' learning gains and retention after reading treatments, so the VKS

was used as a pre-test (18 days before the first treatment), an immediate posttest, and a delayed posttest (16 days after the final treatment). They studied the effects of isolated vocabulary learning (learning words before reading), integrated vocabulary learning (learning words explicitly while reading), and incidental learning (implicitly learning words though reading without teaching those words). Because my study does not focus on different teaching/learning techniques, I review this study and focus on how the researchers used the VKS, how they analysed the VKS scores and reported the results.

The participants were 20 ESL students from a university preparation academic English course in New Zealand. Their first language varied: Korean, Chinese, Malay, Indonesian, and Thai. The participants were divided into two groups, but each group did these two reading treatments. That is, Group 1 did isolated reading treatment, but Group 2 did integrated reading treatment for the first reading article, and they did the opposite for the second reading article.

The target words were selected from the two articles used in class; 18 words were selected from the article 1 (12 target taught and six words as incidentally learned), 18 words from the article 2 (12 target taught and six words as incidentally learned), and 12 nonsense words (to be included only in the pretest and the delayed posttest to ensure that the participants did not overestimate their word knowledge). However, I consider that using nonwords in the VKS is not necessary, for the VKS requires test takers to verify their knowledge of the words through giving translations or sentences. In my study, I, therefore, did not include nonwords in the test battery.

The researchers explore the participants' knowledge of 36 words and the maximum score for each word was five. For each word learning treatment, the maximum scores were 60 (twelve words, five points each). The average pretest scores for the three groups of words (isolated, integrated, incidental) were relatively low (17.95-19.60 out of 60 points). This implied that the students had little knowledge about the target words. After the treatments, the participants had gained more knowledge of the words. The mean scores for the isolated, integrated, and incidental words were 37.10, 33.55, and 21.90, respectively. The participants seemed to learn the target words through isolated word learning best and both the isolated and the integrated word learnings were much better than learning words incidentally. The delayed posttest scores decreased to 25.10, 24.60, and 22.95, respectively. The isolated word learning helped retention the most.

In addition to qualitatively comparing the scores among the three learning techniques, the researchers also examined the scores quantitatively for all the participants and for all the words tested (20 participants x 36 words = 780 items).

They examined how the scores of these items distributed over the five VKS scores. They grouped the VKS scores into two groups: VKS scores of 1-2 representing 'no and little knowledge of the target words', and VKS scores of 3-5 representing 'having learnt the target words', like Paribakht and Wesche (1997).

They found that before the reading treatments most of the words were at VKS score one and two, or "almost uniformly unknown" (p. 235). By the posttest, more than half of the words learned through isolation (139 items or 58%) were at VKS scores three to five, while more than half of the words from the other two groups were still at the unknown stages. The analysis also showed that learning words through either isolation or integration could lead to better learning and retention than incidental learning. Even though the isolated word learning was better than the integrated word learning, the difference was not statistically significant (p value of more than 0.05).

For my study, I decided to use the VKS and compared the receptive score (answers from categories III and IV) with the controlled productive score (answers from category V). These answers were graded with a multi-scale scoring system. I also decided to use the scoring system used by Paribakht and Wesche (1993; 1997) because this was proved valid to track small increase in vocabulary knowledge.

Self-reported vocabulary knowledge scales without verification

Meara (1996b) and Waring (1999; 2000) comment that the progressive stages in the VKS implies that once the knowledge of a word move from one state to another the knowledge will be permanent at that particular state which does not agree with what happen in real life. Waring argues that vocabulary knowledge should not move neatly on scale (from no knowledge to know its form/meaning and to the ability to use it). Therefore, Waring (1999) devised a new test for vocabulary acquisition called State Rating Tasks (SRT) – a rating task which allow test takers to rate their knowledge of a word by starting at any of the states presented.

Some previous studies used Waring's (1999; 2000) State Rating Tasks (e.g., Aizawa et al., 2003; Dabaghi and Rafiee, 2012; Sydorenko, 2010; Waring, 1999; Waring, 2000) to track learners' vocabulary development.

> "When doing this task, the subject is presented with a list of words against which she has to assign a letter, A, B, C, D or E. All words must be rated and blanks are not acceptable" (Waring, 2000, p. 139).

The numbers of words falling into each state are compared to show the change of vocabulary knowledge. Figure 3.13 shows the SRT presented in Waring's (2000) paper.
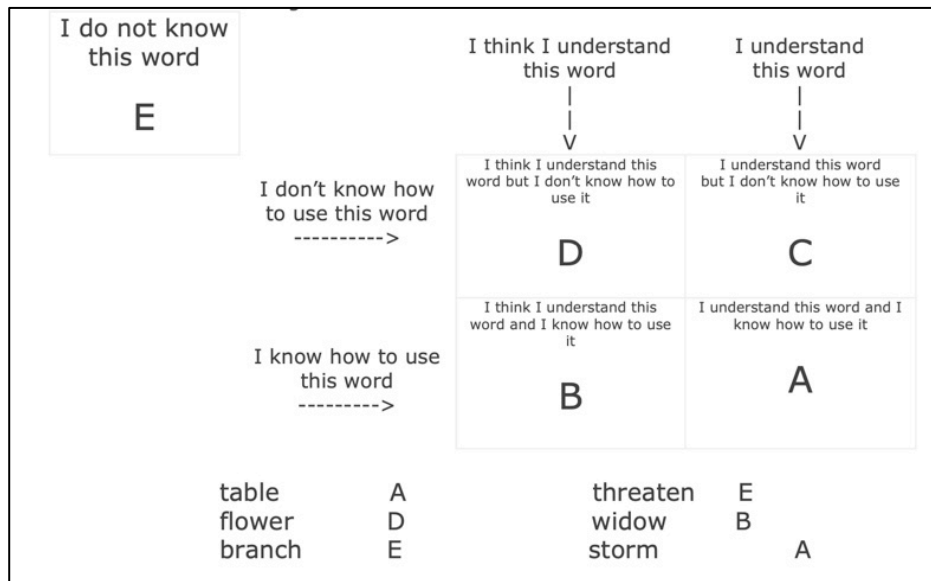
Figure 3.13: Waring's (2000) State Rating Tasks (SRT)

Dabaghi and Rafiee (2012) used the Waring's (2000) SRT to investigate changes in vocabulary knowledge as affected by learning words through two types of glosses (L1-to-L2 gloss vs L2-to-L1 gloss). L1 gloss showed Persian translations of the target words and L2 gloss showed English definitions. There were 56 Iranian university students in the study, and they were randomly assigned to Group 1 (reading and using L1-to-L2 gloss) and Group 2 (reading and using L2-to-L1 gloss). Twenty words from a passage taught in class were the target words and these words were academic words, checked by using Nation's LFP.

The SRT was used as a pretest (two weeks before the main study) and as an immediate posttest. The students were not told about the posttest to ensure the incidental vocabulary learning. The number of the states E, D, C, B, and A that were chosen by the subjects for each of the target words were converted into percentages. The percentages of the data time 1 (pretest) was compared to Data time 2 (posttest). They found that the Group 1 subjects' knowledge of most of the target words move to better states. The scores decreased a lot at State E and increased at State A. A few of them moving to the near state or staying at the same state. For example, the scores of Item 1 for Group 1 in Time 1 were 70%, 10%, 5%, 5%, and 10% (from E to A), but the scores in Time 2 were 20%, 10%, 5%, 5%, 60%. This showed that learning words using glosses increased the participants' state knowledge. Even though the participants in Group 2 had similar scoring patterns like those in Group 1, the percentages of State A in Time 2 tended to be slightly smaller

than that of Group 1. For example, the Time 2 scores of Item 1 and Item 2 for Group 1 were 60% and 60%, those for Group 2 were 50% and 35%. This indicated that learning words with L2-to L1 gloss was not as effective as with L1-to-L2 gloss.

Although Waring (1999; 2000) argues that the SRT does not show vocabulary knowledge in a progressive scale, how the researchers capture small increases in vocabulary knowledge by using the SRT does not differ from how they do by using the VKS, though. For example, when Dabaghi and Rafiee reported the scores, they arranged the scattering states into a linear progression ranging from no knowledge (State E), having receptive knowledge of the word (States D and C), having productive knowledge of the word (State B and A). They are just arranged in different way from the VKS.

Moreover, in Dabaghi and Rafiee's experiment, the glosses seem to help the participants improve the knowledge of meaning and the knowledge of L2 form, but the scores in State A (knowing meaning and being able to use it) also increased a lot. This cannot guarantee whether the participants had really developed the knowledge of using the words. Therefore, without knowledge demonstration, State B (not sure about the word meaning but being able to use it) can be really a problematic choice. This emphasises the main drawback of a self-reported rating scale without knowledge verification, and I decided not to use this test format. The summary of previous studies with details about methodology is presented in Table 3.1.

3.3   What I have learned from the literature review

After reviewing the previous studies, I have learned two major things, which are how they conducted their research (what tests, what types of vocabulary, what frequency levels) and what they found in common about receptive and productive vocabulary knowledge.

3.3.1   underlying ideas of research on receptive and productive vocabulary knowledge

Firstly, I have learned that researchers with the idea of the dichotomous framework see receptive and productive vocabularies as separate groups. They explored the size or the level of either receptive vocabulary or productive vocabulary and compared the vocabulary size or the vocabulary level of students with different

backgrounds and different native languages or to explore the nature and the gap between these two types of vocabularies. They also compared learners' vocabulary level to the vocabulary level of the textbooks they need to use in their classroom to check whether the learners would be able to cope with the teaching materials or the real language use. Some also examined the effects of factors such as years of study, different teaching methods or task types and time spent in different learning contexts on the growth or the changes of both vocabularies, either longitudinally or cross-sectionally.

However, the researchers who believe that the relationship between receptive and productive vocabularies is on a continuum do not focus on the vocabulary size but on how much vocabulary items taught in class are acquired by learners. The test they used are designed to capture small changes of vocabulary knowledge, from the state of no knowledge, through recognising word form and having knowledge of meaning, to being able to recall the words or being able to use them in writing.

3.3.2 Test formats

I have learnt that different test formats were designed to suit research with different aims and designs. Importantly, no test fits all. They all have both advantages and drawbacks. However, when considering the research aim, research design, target participants, focused groups of vocabulary, theory about the relationship between receptive and productive vocabulary knowledge, I came to the conclusion that the VKS would be used to explore receptive and controlled productive vocabulary knowledge. An essay-writing task would be used to explore free vocabulary knowledge in this study for the following reasons.

1) The VKS is also a flexible format that allows me to include many target words (both general and academic words) to suit the research aim (to understand vocabulary knowledge as a whole) and the participants (university learners).
2) The VKS and the essay-writing task certainly require learners' knowledge demonstration. Therefore, the possibility of guessing can be avoided, and this can estimate the knowledge of words more accurately.
3) The VKS include all the main features of vocabulary knowledge (form, meaning, and use) as presented in Nation's (2013) taxonomy. It does not only focus on only form or on only form and meaning.
4) The VKS and the essay-writing task can help me understand more about the relationship among receptive, controlled productive, and free productive vocabulary knowledge.

5) The VKS is used to help confirm how sure the participants are with their answers. It is used as a guide to the participants. If they know how to use the test word in a sentence, they choose a score of 5 and need to provide a sentence using the test word afterwards.

6) I can score what the participants supply in the VKS and in the essay writing tasks with a sensitive scoring system or a multi-scale scoring system because it is normal for a learner to have partial knowledge of a word as Read (2000) suggests. Therefore, two points is given to a correct answer, one point is given to a partial correct answer and no point is given to a wrong answer.

7) The final and the most important reason is that the VKS allows me to adopt both of major ideas of how receptive and productive vocabularies are related. This is because up until now the relationship between the two groups of vocabulary has still been debatable. Based on the idea of dichotomous relationship, I can separately analyse the participants' answers. That is, categories III to IV can be grouped and analysed as the receptive knowledge (having the knowledge of word meaning), while categories V can be analysed as the controlled productive vocabulary knowledge (being able to use the word in a sentence level).

   If based on the idea of a continuum, the participants' answers can be scored in more detail. I would examine how it moves from one category to another category. Even though Meara (1996b) and Waring (1999; 2000) argue that the VKS has a progressive scale, the VKS can also show the increase, decrease, and no change of the participants' vocabulary knowledge. For all these reasons, the VKS and the essay-writing task are then adopted in this research study.

3.3.3 Test words and frequency levels

The next issue to be taken into account before the main study is which group of words to be tested and which frequency levels to be included. Content words have been the focus when exploring receptive and productive vocabularies, so content words would be used as test words, not function words.

According to the literature, frequency level is strongly related to the likelihood of a word being known" (Nation and Beglar, 2007, p. 11). That is, how one acquires words is affected by word frequency. High-frequency words tended to be acquired before the low-frequency ones. Also, it might be useful to know whether the participants reach the mastery level of particular frequency levels. Frequency levels have been used as thresholds for learners' language abilities.

To investigate vocabulary knowledge, it might be ideal to include as many frequency levels as possible. The aim of my study is to explore the vocabulary knowledge in general for university students, so words from many frequency levels and from academic words seem to be most appropriate. Also, I decided to use words tested in widely-used vocabulary tests such as the VST or the VLT because they had been carefully selected by the test designers and tested with many multi-cultural ESL and EFL students, ranging from the primary level to the university level.

Even though Nation and Beglar's (2007) VST contains many more frequency levels than the VLT, the VST does not include academic words. Words from another widely-used test – Schmitt et al.'s (2001) VLT – does not only include word from high-, mid-, and low-frequency levels but it also includes academic words (the 2,000-, the 3,000-, the 5,000-, and the 10,000-word levels, and the academic word level). Even though these four frequency levels were sampled from old word lists: Thorndike and Lorge's (1944) 30,000-word list, West's (1953) GSL and Coxhead's (2000) AWL, they have been tested worldwide with different groups of students by many of recent studies, and they seem to suffice to represent words in general. However, the test was piloted so that I knew how many test words to be included.

As word frequency is involved, the research questions of the present study from Section 2.3 was revised and they are presented in Section 4.2.

3.3.4   Analysing data on free productive vocabulary knowledge

From the literature, I found that the data on receptive vocabulary could be compared to that on controlled productive vocabulary in terms of frequency levels or groups of words. However, the data on free vocabulary knowledge could be hardly compared to the first two. This is because the widely-used vocabulary-profiling programmes such as Cobb's (n.d.) VocabProfile can profile words from compositions into limited groups (the 1,000- and the 2,000-word levels from the GSL, the academic words (the AWL), and 'not-in-list' words). It cannot show words that fall into the 3,000-, the 5,000-, and the 10,000-word lists. Therefore, the VocabProfile was not used to analyse the participants' free vocabulary in my study.

Instead, I decided to use Text Lex Compare, which is also available on Lextutor Website. Figure 3.14 shows Text Lex Compare software on Lextutor Website.

This was because it allowed me to compare words supplied by the participants and falling in the same word lists I used in the VKS such as the 3,000-, the 5,000-, and the 10,000-word list. Then these words were graded whether they were used by the participants correctly or not. A multi-scale scoring was adopted. In this way, the data

on free productive vocabulary was not only about how many words from the GSL or the AWL were used.

With these methods, the data on free productive vocabulary was comparable to that on receptive and controlled productive vocabulary. How I analysed the free productive vocabulary is presented in Section 4.6.4.
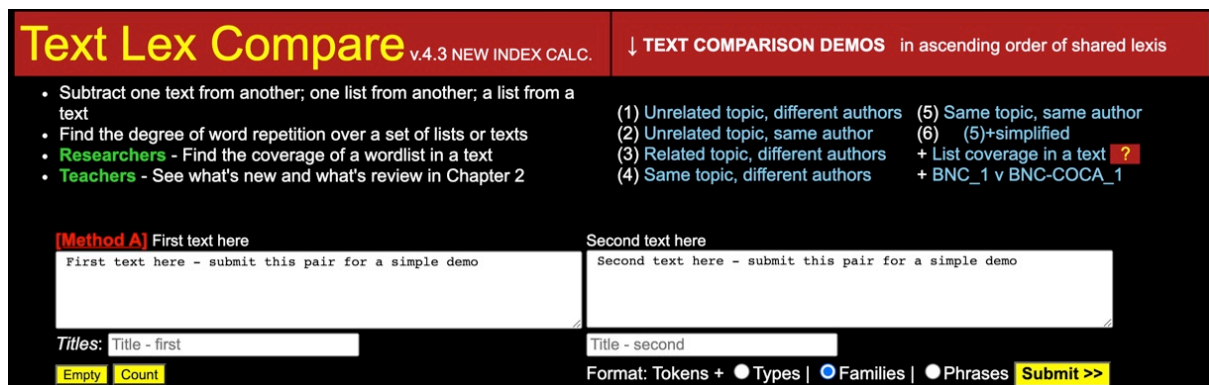
Figure 3.14: Text Lex Compare software on Lextutor Website

3.3.5   issue about word families or lemmas

In order to make all the tests comparable, another important issue is whether to compare words families or to compare lemmas. Word families have been used in studies on receptive and productive vocabularies (e.g., AbManan et al., 2017; Al-Masrai, 2009; Al-Masrai and Milton, 2012; Barrow et al., 1999; Fan, 2000; Laufer, 1998; Laufer and Goldstein, 2004; Laufer and Paribakht, 1998; Meara and Jones, 1990; Sahiruddin, 2008; Sungprakul, 2016; Ward, 2009a; Ward, 2009b; Zhou, 2010). This is probably because the available vocabulary tests are based on lists of word families.

Word families have been claimed by Bauer and Nation (1993) and Nation (2013) to be the most suitable when testing learners' vocabulary knowledge because when knowing one base form learners tend to know its inflections and its derivatives. However, there is evidence showing that word families might not be suitable for EFL students as they lack knowledge of affixes (e.g., Schmitt, 1998; Ward and Chuenjundaeng, 2009; Stoeckel et al., 2020). Also, "correct responses [to a base form should] not assume knowledge of derivational forms" (Stoeckel et al., 2020, p. 11).

Ward and Chuenjundaeng (2009) studied the engineering students' knowledge of English affixes and found that they had little knowledge of derivational affixes. They

concluded that "[o]ur results contradict the assumption that knowledge of headwords implies knowledge of word families, at least with lower-level students from non-Latinate L1 backgrounds" (Ward and Chuenjundaeng, 2009, p. 465).

This is because their participants, engineering students with low-proficiency English level, failed to recognise derived forms of the base words they knew. Ward (2009a; 2009b) found that his engineering students in Thailand had relatively weak knowledge of Latinate suffixes. When he created an engineering word list for them, he even included 'word types' in the list.

Similarly, interviewing the participants to capture their knowledge of derived form, Schmitt (1998) reported in his study that even the advanced EFL participants, who were able to enter a British university for their postgraduate programmes, did not seem to master the knowledge of different derived forms of the target words.

> "This definite gap in these advanced learners' morphological knowledge called in question the assumption that a learner who knows one member of a word family can easily learn the others" Schmitt (1998, p. 307).

Schmitt (2010) also suggests that lemmas should be used when investigating learners' productive vocabulary knowledge and word families could be used for exploring receptive vocabulary knowledge. However, Schmitt and Schmitt (2020) write:

> "[F]or most purposes, and for most L2 learners, the lemma [is] likely to be the most suitable unit. This is especially true in testing. Research has shown that many learners do not know the full range of derivative forms of words, either receptive or productively (e.g., González-Fernández and Schmitt, 2020; McLean, 2018; Schmitt and Zimmerman, 2002). … Because of this, several scholars (e.g., Kremmel, 2016) have argued that the lemma is a better counting unit for L2 tests, because learners are much more likely to know the inflected members of lemmas (e.g., boy → boys; deploy → deployed, deploying, deploys), leading to more accurate test estimates" (Schmitt and Schmitt, 2020, p. 212).

Considering the evidence mentioned above and the participants of the study (English-major EFL students), lemmas were the best counting unit for this study. The participants were not low-level learners, so they tend to have good knowledge of inflections. Their knowledge of derivational affixes cannot be guaranteed, though. Also, the aim of this study is to compare receptive and productive vocabulary. Consequently, lemmas were used in all the tests for comparability. Even though the

findings cannot be completely compared to other studies in terms of vocabulary size, the trends or the developmental patterns can be compared.

3.3.6   Research design

Regarding the research design, a longitudinal research design and a cross-sectional research design have their own advantages and disadvantages. Longitudinal studies (e.g., Dabaghi and Rafiee, 2012; Tahmasbi and Farvardin, 2017; Zheng, 2012; Zhong, 2016) normally track learners' vocabulary development within one academic term or one academic year. They test the same group of students twice (before and after a treatment or different treatments using a pre-posttest design), or many times during a course or an academic year, for example, at the beginning, at the middle, and at the end. The pre-test and the post-test include the same group of words learned in class, and the tests are employed to check whether the learners know more of the taught words. This design is better to track the change of a few words, especially words focussed during syllabus-based treatments within a short period of time.

Even though a longitudinal method allows the researchers to track vocabulary growth of the same participants (e.g., Pellicer-Sánchez, 2019; Schmitt, 1998; Webb and Chang, 2012) , it would be time-consuming and difficult to track the vocabulary used by the same large group of participants. If this study longitudinally investigates whether the participants have developed the receptive and productive vocabulary knowledge over their years of study as the English-major students, it might take four years. Because of the time limitation, the research design for my study was then not a longitudinal study, but a cross-sectional study (Azodi et al., 2014; Zyad, 2017).

A cross-sectional study allows the comparisons of the vocabulary knowledge development of all the four year groups with the same test at the same period of time, e.g., during the same month in a semester.

> "Another advantage [of a cross-sectional study] is that the researcher does not have to maintain long-term cooperation with the participants, which has been known to be demanding in terms of time and effort. However, one important disadvantage is that the researcher can't say how the participants developed from one given point in time to another and thus the results should be treated with caution" (Zyad, 2017, pp. 366-367).

The purpose of this study is not to examine the efficiency of any particular treatment but to examine the knowledge of vocabulary as a whole (both general and academic vocabulary) of the four year groups of English-major students in Thailand in general

and to investigate the relationship among their receptive, controlled productive, and free productive vocabulary knowledge over their years of study. Additionally, the time duration for the data collection should be less than one year, so a longitudinal design was not a suitable one for this study. A cross-sectional study was more suitable and more practical for the present study. I acknowledge that the higher scores or the lower scores of the vocabulary knowledge of the participants in my study cannot completely represent the vocabulary development over years of study.

3.3.7   Number of participants

The next issue to consider before the main study is the number of participants. As the research design is a cross-sectional study and the aim is to explore the vocabulary growth over time, the participants would be selected from each of the four year cohorts. According to Cohen et al. (2011; 2017), the optimum number of participants should be 30 per variable, so I planned to recruit 30 students from each year group. Therefore, the first sub-questions for each of the research questions (presented in Section 2.3) should be slightly changed. That is, the phrase 'over time' is changed into 'at different years of study'. The revised research questions were presented Section 4.2.

3.3.8   Common findings from previous studies

In addition to the methodology, what I learn from reviewing previous studies on receptive, controlled productive, and free productive vocabularies is some common findings, which I later can compare to those from my study. The follows are the patterns of vocabulary development, the factors affecting vocabulary knowledge, the proportion of receptive and productive vocabularies when comparing them and some unexpected results of some common findings.

1)   Receptive vocabulary is always larger than controlled productive vocabulary, regardless of tests used, frequency levels, and groups of participants (e.g., Amin, 2020; Kotchana and Tongpoon-Patanasorn, 2015; Ozturk, 2015; Shin et al., 2011; Webb, 2008; Zhou, 2010).

2)   Both receptive and controlled productive vocabularies tend to be larger at higher frequency levels; they correlated to each other (e.g., AbManan et al., 2017; Li and MacGregor, 2010; Ozturk, 2015; Webb, 2008). However, Amin's (2020) showed some unexpected results. While most of vocabulary was affected by frequency levels, the average receptive score at the 4,000-word

level was higher than that at the 3,000-word level, and the average controlled productive score at the 3,000-word level was higher than that at the 2,000-word level. Amin did not explain the reason. This is unlikely, but it made me aware that it can happen in research. The major trends can be reported, and the exceptions can be investigated further qualitatively.

3) The scores of both receptive and controlled productive vocabulary knowledge at the 3,000-word level are likely to be similar to those at academic word level (e.g., Amin, 2020; Li and MacGregor, 2010; Ozturk, 2015).

4) Years of study is not always a dominant factor of vocabulary growth both in cross-sectional studies and longitudinal studies. In some studies, it did affect vocabulary development (e.g., Alonso and Fontecha, 2014; Azodi et al., 2014; Sungprakul, 2016) and for some studies it did affect but not significantly (e.g., Al-Masrai and Milton, 2012; Gallego and Llach, 2009; Li and MacGregor, 2010; Ozturk, 2015; Zyad, 2017). In Ozturk's (2015) studies, the scores of the participants in both a cross-sectional study and a longitudinal study did not show vocabulary improvement over years of study. Ozturk assumed that this was probably because of the ceiling effect as the participants' receptive and controlled productive vocabulary scores were relatively high, for example, around 96% for the receptive vocabulary and 80% for the controlled productive vocabulary at the 2,000-word level.

5) Controlled productive/receptive ratios were influenced by word frequency levels. The controlled productive/receptive ratio at high-frequency levels tends to be high. This implies that the receptive-productive gap at high frequency levels were smaller and that words with higher frequency seemed to become part of learners' productive vocabulary more easily than words with lower frequency (e.g., Shin et al., 2011; Waring, 1997a; Webb, 2008).

6) The ratios are also affected by learners' proficiency level. Those who have larger receptive tended to have larger controlled productive vocabulary (e.g., Waring, 1997a; Webb, 2008; Zhou, 2010). Receptive vocabulary grew at a faster rate (Laufer and Paribakht, 1998; Waring, 1997a; Zhou, 2010). Differently, the controlled productive vocabulary in a study by Zhong and Hirsh (2009) grew at faster rate. This was the growth at the 2,000-word level, at which the receptive vocabulary scores were affected by the ceiling effect. However, the controlled productive and receptive ratios in Fan's (2000) study are not consistent with this finding. The ratios varied and were unpredictable. This is probably because of the methodology, especially how to design the controlled productive vocabulary test, in Fan's study.

7) Vocabulary knowledge aspects are interrelated. The change of one aspect leads to the change in the other aspects. There is no evidence of hierarchical development of aspects (e.g., Schmitt, 1998; Zhong, 2016). When using more complex statistics, González-Fernández and Schmitt (2020) found that there seems to be the order of the acquisition of vocabulary knowledge aspects and that aspects recognition comes before aspects recall on the receptive-productive continuum. However, there is no measurable threshold found in productive vocabulary development (e.g., Webb, 2008). The knowledge of form and meaning is best correlated the productive knowledge (e.g., Webb, 2008).

8) Learners' vocabulary knowledge appears to develop incrementally in an upward trend over time. However, it can also stay at the same knowledge stage or even decrease if it is not used or exposed to (e.g., Dabaghi and Rafiee, 2012; Paribakht and Wesche, 1997).

Through reviewing, I learnt about the methodology usually adopted in the field and I decided which methods should be the most suitable for my study. I learnt the common findings the common findings and some exceptions in the end. All these are useful for my methodology and the findings of my study. A summary of the studies reviewed is given in Table 3.1.

No.	Author (Year)	Aim: To examine	Research design	Test	Scoring/ Analysis	Word list	Word group	Context	Participant	Major
1	Ward (2009b)	receptive vocabulary size	devising tests and assessment	an EWL and a GSL checklist tests	dichotomous	EWL (Ward, 2009a) and GSL (West, 1953)	engineering and general words	Thailand	250 first-year	engineering
2	Al-Masrai and Milton (2012)	receptive vocabulary size and growth	assessment, cross-sectional	EVST (Re1) (Meara and Jones, 1990) and XK_Lex (sets A (Re2) and B (Re3)) (Al-Masrai, 2009)	dichotomous	Thorndike and Lorge (1944), Nation (1984), (Kilgarriff, 2006)	general	Saudi Arabia	55 first-year and 37 fifth-year	English
3	Sungprakul (2016)	receptive vocabulary size and growth	assessment, cross-sectional	XK_Lex (sets A (Re1) and B (Re2)) (Al-Masrai, 2009)	dichotomous	Thorndike and Lorge (1944), Nation (1984), (Kilgarriff, 2006)	general	Thailand	10 first-year, 10 second-year, third-year, 10 fourth-year	English
4	Yunus et al. (2016)	receptive vocabulary size	assessment, comparative	VST (Nation and Beglar, 2007) (Re)	dichotomous	Nation's (2006) BNC	general	Thailand and Malaysia	86 Thai first-year and 80 Malay first-year	English
5	Li and MacGregor (2010)	receptive vocabulary level and growth	assessment, cross-sectional	VLT (Schmitt et al., 2001) (Re)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953), AWL (Coxhead, 2000)	general and academic	Hong Kong, China	81 first-year BA and 47 MA	English

6	Alonso and Fontecha (2014)	receptive vocabulary level and growth	assessment, longitudinal	VLT (Schmitt et al., 2001) (Re)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953)	general	Spain	30 adolescent and adult leaners	English language course
7	Gallego and Llach (2009)	receptive vocabulary level and growth	assessment, longitudinal	VLT (Schmitt et al., 2001) (Re)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953)	general	Spain	283 4th graders	-
8	Zhang and Lu (2013)	receptive vocabulary level and growth	assessment, longitudinal	VLT (Schmitt et al., 2001) (Re)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953), AWL (Coxhead, 2000)	general and academic	China	300 first-year	English and non-English
9	Laufer and Nation (1999)	controlled productive vocabulary level and growth	assessment, cross-sectional	PVLT (Laufer and Nation, 1999) (CP)	dichotomous but sensitive	Thorndike and Lorge (1944), GSL (West, 1953), UWL (Xue and Nation, 1984)	general and academic	EFL	24 10th graders, 23 11th graders, 18 12th graders, and 14 first-year	-
10	Alonso and Garcia (2014)	controlled productive vocabulary size	assessment	PVLT (Laufer and Nation, 1999) (CP)	dichotomous but strict	Thorndike and Lorge (1944), GSL (West, 1953), UWL (Xue and Nation, 1984)	general and academic	Spain	38 10th graders	-

11	Zyad (2017)	free productive vocabulary: lexical diversity and lexical sophistication	assessment, cross-sectional	30 narrative essays (FP)	D-index (Malvern et al., 2004) and LFP (Laufer and Nation, 1995)	GSL (West, 1953), UWL (Xue and Nation, 1984)	general and academic	Morocco	30 first-year, 30 second-year, and 30 third-year	Faculty of Letters and Humanities
12	Laufer and Nation (1995)	free productive vocabulary: lexical profile	developing a measure, cross-sectional	2 argumentative essays (Laufer and Nation, 1995) (FP1 and FP2)	LFP (Laufer and Nation, 1995)	GSL (West, 1953), UWL (Xue and Nation, 1984)	general and academic	Israel and EFL in New Zealand (NZ)	22 EFL students in NZ (low-proficiency), 20 Israeli (English majors – Term 1)(medium proficiency), 23 Israeli (English majors – Term 2) (high-proficiency)	English and non-English
13	Zheng (2012)	free productive vocabulary development: lexical profile	assessment, longitudinal, multi-case study	collecting classroom drafts – argumentative essays (FP)	LFP (Laufer and Nation, 1995)	GSL (West, 1953), UWL (Xue and Nation, 1984)	general and academic	China	4 third-year	English

14	Azodi et al. (2014)	free productive vocabulary development: lexical profile	assessment, cross-sectional	collecting 210 writing samples (FP)	Range (Nation, 2008), number of word types	GSL (West, 1953), AWL (Coxhead, 2000)	general and academic	Iran	4 year cohorts	English translation
15	Djiwandono (2016)	free productive vocabulary: lexical profile and lexical variation	assessment, comparative	collecting academic papers (FP)	VocabProfile (Cobb, n.d.)	GSL (West, 1953), AWL (Coxhead, 2000)	general and academic	Indonesia	fourth-year students and teachers	English Letters Department
16	AbManan et al. (2017)	receptive and controlled productive vocabulary knowledge	assessment, comparative	VLT (Schmitt et al., 2001) (Re) and PVLT (Laufer and Nation, 1999) (CP)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953), AWL (Coxhead, 2000), UWL (Xue and Nation, 1984)	general and academic	Malaysia	156 first-year	non-English
17	Zhou (2010)	receptive and controlled productive vocabulary knowledge	assessment, comparative	VLT (Schmitt et al., 2001) (Re) and a revised PVLT (Zhou, 2010) (CP)	dichotomous	AWL (Coxhead, 2000)	academic	China	72 university students	non-English

18	Fan (2000)	receptive and controlled productive vocabulary knowledge	assessment, comparative	VLT (72-item) (Nation, 1990) (Re) and a revised PVLT (Fan, 2000) (CP)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953), UWL (Xue and Nation, 1984)	general and academic	Hong Kong, China	138 first-year	non-English
19	Waring (1997a)	receptive and controlled productive vocabulary knowledge	assessment, comparative	an adapted VLT (Nation, 1990) (Re) an adapted PVLT (Laufer and Nation, 1999) (CP)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953), UWL (Xue and Nation, 1984; Waring, 1997a)	general and academic	Japan	76 first- and second-year	English
20	Ozturk (2015) Stage 1	receptive and controlled productive vocabulary development	assessment, comparative, cross-sectional	VLT (Nation, 2001): 30 items each but 36 items at AWL(Re) and PVLT (Nation, 2001): 18 items each level (CP)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953), AWL (Coxhead, 2000) UWL (Xue and Nation, 1984)	general and academic	Turkey	55 first-year and 45 fourth-year	ELT programme
21	Ozturk (2015) Stage 2	receptive and controlled productive vocabulary development	assessment, comparative, longitudinal	VLT (Nation, 2001: 30 items each but 36 items at AWL(Re) and PVLT (Nation, 2001: 18 items each level (CP)	dichotomous	Thorndike and Lorge (1944), GSL (West, 1953), AWL (Coxhead, 2000), UWL (Xue and Nation, 1984)	general and academic	Turkey	17 first-year	ELT programme

22	Zhong and Hirsh (2009)	receptive and controlled productive vocabulary development	assessment, comparative, longitudinal	an adapted VLT (Zhong, 2008) (Re) and an adapted PVLT (Zhong, 2008) (CP)	dichotomous	general word lists (mentioned in unpublished work by Zhong (2008)), AWL (Coxhead, 2000)	general and academic	China	44 11th graders	-
23	Kotchana and Tongpoon-Patanasorn (2015)	receptive and controlled productive vocabulary sizes	assessment, comparative	a devised multiple-choice test (Re) 20 items and a devised cued-recall test (CP) 20 items	dichotomous	GSL (West, 1953)	general	Thailand	453 6th graders	-
24	Shin et al. (2011)	receptive and controlled productive vocabulary sizes	assessment, comparative	an adapted VST (Nation, 2010) 100 items (Re) and an adapted PVLT (Laufer and Nation, 1999) 100 items	dichotomous for Re and multi-scale scoring for CP	14000 BNC list (Nation and Beglar, 2007)	general	Korea	402 11th graders	-
25	Amin (2020)	receptive and controlled productive vocabulary sizes	assessment, comparative	VST (Nation and Beglar, 2007) 100 items, PVLT (Laufer and Nation, 1999)	dichotomous	14000 BNC list (Nation and Beglar, 2007, Thorndike and Lorge (1944), GSL (West, 1953), AWL (Coxhead, 2000), UWL (Xue and Nation, 1984)	general and academic	Afghanistan	54 fourth-year	English

26	Webb (2008) Stage 1	receptive and controlled productive vocabulary	assessment, comparative	Translation tests (Re) 90 items and (CP) 90 items	dichotomous (sensitive)	180 words sampled from 3 bands of COBUILD	general	Japan	83 second-year	English and non-English
27	Webb (2008) Stage 2	receptive and controlled productive vocabulary	assessment, comparative	Translation tests (Re) 90 items and (CP) 90 items	dichotomous (sensitive and strict)	180 words sampled from 3 bands of COBUILD	general	Japan	25 upper group and 25 lower group	English and non-English
28	Vincy (2020)	receptive and controlled productive vocabulary	classroom experiment, pre-posttreatment (T1 and T2), comparative	VKS (Wesche and Paribakht, 1996) and PVLT (Laufer and Nation, 1999)	dichotomous	words sampled from a textbook	general	India	62 6th graders (31 experimental group and 31 control group)	-
29	Oberg (2012)	receptive and controlled productive vocabulary	classroom experiment, pre-posttreatment (T1 to T11), comparative	a matching test (Re) and a sentence-writing task (CP)	dichotomous	15 words from a textbook	general	Japan	58 first-year	non-English

30	Laufer (1998)	relationship and development of receptive, controlled productive and free productive vocabularies	assessment, comparative, cross-sectional	VLT (Nation, 1983; 1990), PVLT (Laufer and Nation, 1999), and essays (Laufer and Nation, 1995)	dichotomous and LFP (Laufer and Nation, 1995)	Thorndike and Lorge (1944), GSL (West, 1953), UWL (Xue and Nation, 1984)	general and academic	Israel	26 10th graders and 22 11th graders	-
31	Laufer and Paribakht (1998)	relationship and development of receptive, controlled productive and free productive vocabularies	assessment, comparative, cross-sectional	VLT (Nation, 1983; 1990), PVLT (Laufer and Nation, 1999), and essays (Laufer and Nation, 1995)	dichotomous and LFP (Laufer and Nation, 1995)	Thorndike and Lorge (1944), GSL (West, 1953), UWL (Xue and Nation, 1984)	general and academic	EFL – Israel and ESL in Canada	52 10th and 11th graders, 27 English majors in Israel, and 103 ESL non-major in Canada	English and non-English
32	Zhong (2016)	relationship of the five aspects: meaning, form, word class, collocation, and association	assessment and comparative	spelling test (Webb, 2008), WAT (Read, 1998), Fill-in-blank (Ishii and Schmitt, 2009), VKS (categories I-IV) (Paribakht and Wesche, 1997)	dichotomous and three-scale scoring system for the VKS	26 words from textbooks	general words	China	620 8th graders	-

33	Schmitt (1998)	acquisition of vocabulary aspects: form, meaning, association, and grammar	interview, multi-case, longitudinal	supplying a correct spelling, an associate, all derived forms, all meaning senses	a multi-scale scoring system	11 words from UWL (Xue and Nation, 1984) and Brown corpus (Francis and Kučera, 1979)	general and academic	ESL in England	3 students in pre-sessional course	non-English
34	González-Fernández and Schmitt (2020)	acquisition of vocabulary aspects: form and meaning, collocation, concept and referent (multiple meanings), word parts (derivatives or parts of speech)	devising tests and assessment	VLT, recognition and recall tests for the four aspects	A dichotomous scoring system	20 words based on these criteria: -1K-9K -Different parts of speech -Multiple-meanings (at least 3 students) -At least 3 derivatives form for one of the meaning senses -A percentage of Spanish-English cognates (34-37%)	general	Spanish-speaking learners of English (in Spain and in the UK)	144 students (18-65 years old)	Not mentioned, but from beginners to advanced

35	Webb (2007b)	acquisition of receptive and productive vocabulary aspects over time (1, 3, 7, and 10 encounters); aspects include form, meaning, grammar, association, and collocation	experiment, comparative	(Re) multiple-choice tests for all the five aspects; (CP) supplying a form, an associate, a collocate, and a sentence-writing task	dichotomous	10 words from a textbook (replaced with ten nonwords)	general words	Japan	121 second-year students	not mentioned
36	Paribakht and Wesche (1997)	effect of different teaching methods on vocabulary knowledge	classroom experiment	VKS (Paribakht and Wesche, 1993; 1997)	a developmental scoring system	28 content words and 10 connectives from a textbook	general	ESL in Canada	38 students in English course in Canada	not mentioned
37	File and Adams (2010)	effect of different learning techniques on vocabulary gain and retention	classroom experiment, pretest-posttest-delayed posttest	VKS (Paribakht and Wesche, 1993; 1997)	a developmental scoring system	36 words from a textbook	general	ESL in New Zealand	20 students in a pre-sessional course	non-English

38	Dabaghi and Rafiee (2012)	effect of different teaching methods on vocabulary acquisition	classroom experiment, pre-posttreatment	SRT (Waring, 2000)	five states: A, B, C, D, E (Waring, 2000)	20 words from a textbook	academic	Iran	59 university students	not mentioned

Table 3.1: Summary of previous studies on receptive and productive vocabularies

Chapter 4 Research methodology

4.1   Introduction

Reviewing the literature, I found that there is no consensus whether the relationship between receptive vocabulary and productive vocabulary is dichotomous or continuum. I also found that the scoring systems to check participants' answers in the receptive, controlled productive, and the free productive vocabulary tests were not consistent, especially that for the free productive vocabulary test. Therefore, I aimed to explore these three types of vocabulary by not taking sides but would rather explore them within both the dichotomous and the developmental approaches and attempts to find how they are related in the end. I also aimed to present the innovation of the research methodology – how to make the analysis of the free productive vocabulary consistent to that of the receptive and the controlled productive vocabulary. All the analysing methods are presented in the form of flow charts in Chapter 6.

Based on these ideas, Chapter 4 presents the research questions, all the research tools used, and how data is analysed. It also presents the context of the study and the selection of the participants, the preparation of the research tools, arranging test words, the selection of the essay topics, as well as rating the participants' answers.

4.2   Research questions, research tools, and analysis methods

This section presents a table showing the revised research questions, the sub-questions, the research tools required to answer them, and the analysis methods (scoring, references and statistics used). These are shown in Table 4.1.

Research questions	Research tools	Analysis methods
1. What is the participants' receptive vocabulary knowledge?	Receptive test (Supplying a translation/synonym to the lemma given) (*10 lemmas per one frequency level; 5 frequency levels)	**Coding:** Classifying the translations/synonyms into groups Deciding which group deserves which score **Dichotomous approach (3-scale scoring system):** Score of 2 for a correct translation/synonym Score of 1 for a partially correct translation/synonym Score of 0 for an incorrect translation/synonym **Developmental approach (5-scale scoring system):** Score of 1 for no knowledge of the lemmas at such frequency level Score of 2 for recognising the written form of the lemma Score of 3 for recognising the written form of the lemma and knowing the lemma's meaning Score of 4 for recognising the written form, knowing the lemma's meaning, and being able to use it with semantic appropriateness in a sentence Score of 5 for recognising the written form, knowing the lemma's meaning, and being able to use it with semantic appropriateness and grammatical accuracy in a sentence **References:** Cambridge English-Thai Dictionary Thai-English Dictionary by NECTEC Oxford Advance Learners' Dictionary **Statistics:** A descriptive statistics, an independent samples t-test, ANOVA, MANOVA, effect size, graphs
1.1 What is the participants' receptive vocabulary knowledge at different frequency levels?		
1.2 What is the participants' receptive vocabulary at different years of study?		
2. What is the participants' controlled productive vocabulary knowledge?	Controlled productive test (Supplying a sentence using the lemma given) (*10 lemmas per one frequency level; 5 frequency levels)	**Dichotomous approach (3-scale scoring system):** Score of 2 for a correct lemma use (with appropriate meaning and correct grammar) Score of 1 for a partially correct lemma use (with appropriate meaning but incorrect grammar) Score of 0 for an incorrect lemma use (with inappropriate meaning) **Developmental approach (5-scale scoring system):**
2.1 What is the participants' controlled productive vocabulary knowledge at different frequency levels?		

2.2 What is the participants' controlled productive vocabulary at different years of study?		Checked together with the receptive test **References:** Oxford Advance Learners' Dictionary Sketch Engine (Kilgarriff et al., 2014) Two British raters **Statistics:** A descriptive statistics, an independent samples t-test, ANOVA, MANOVA, effect size, graphs, inter-rater reliability
3. What is the participants' free productive vocabulary knowledge? 3.1 What is the participants' free productive vocabulary knowledge at different frequency levels? 3.2 What is the participants' free productive vocabulary at different years of study?	the free productive test (*a free writing task - an argumentative essay of around 300 running words)	**Lemmatising:** Lemmatising the words produced in essays Classifying the lemmas into five groups by the five frequency levels using Text Lex Compare on Lextutor Website (Cobb, n.d.) **Dichotomous approach (3-scale scoring system):** Score of 2 for a correct lemma use (with appropriate meaning and correct grammar) Score of 1 for a partially correct lemma use (with appropriate meaning but incorrect grammar) Score of 0 for an incorrect lemma use (with inappropriate meaning) **Developmental approach (5-scale scoring system):** N/A **References:** Oxford Advance Learners' Dictionary Sketch Engine (Kilgarriff et al., 2014) Two British raters **Statistics:** A descriptive statistics, an independent samples t-test, ANOVA, MANOVA, effect size, graphs, inter-rater reliability
4. What is the relationship among the participants' receptive, controlled productive and free productive vocabulary knowledge?	All the three tests above	**Statistics:** A descriptive statistics, an independent samples t-test, ANOVA, MANOVA, effect size, graphs **Analyses:**

4.1 What is the relationship among the participants' receptive, controlled productive and free productive vocabulary knowledge at different frequency levels?		**Quantitative analysis within the dichotomous approach:** quantitatively comparing the scores of the three groups of vocabulary **Quantitative analysis within the developmental approach:** Paribakht and Wesche's (1997) scoring system – quantitatively examining the changes of vocabulary knowledge along the five scores
4.2 What is the relationship among the participants' receptive, controlled productive and free productive vocabulary knowledge change at different years of study?		****Qualitative analysis:** Qualitatively analysing answers of the three tests supplied by some participants for more insights of how receptive and productive vocabulary knowledge is related.
General information of the participants	Questionnaire	**Statistics:** A descriptive statistics
Note: ** Before piloting, the receptive test and the controlled productive test included the same 30 lemmas per frequency level, and the free productive test included two essays.* ***I decided to add the qualitative analysis after I had the findings from the first two approaches of the quantitative analysis.*		

Table 4.1: Research questions, research tools, and analysis methods

4.3   Context for data collection

The context used in the current study was the English programme, the Faculty of Liberal Arts in a public university in Thailand. Each year the programme recruits 50-60 students, through both university central admission examinations and the examinations arranged by the faculty. There are approximately 200-240 English-major students (Year 1 to Year 4) in each academic year. Throughout their university studies, the English-major students need to practice a lot of reading and writing skills.

4.4   Selection of the participants

The four year groups of the English-major students were chosen purposefully as they could represent English-major students in Thailand because 1) they were from every region in Thailand, 2) their mother tongue was Thai, 3) their English language

proficiency varied, and 4) they learnt similar English subjects throughout the four years.

Before collecting the data, asking for permission from the faculty was essential. I wrote a formal letter to the faculty, and it was about the overall information of the research such as the purpose of conducting this research with the English-major students from all the four year cohorts and the procedure of the data collection. The approvals of the ethical review (Appendix A) from both the University of Leeds and the university in Thailand, along with the research tools, the information sheets, and the consent forms were shown to the faculty. I had an informal meeting with the executive team who had the authority of this permission. After receiving the permission from the faculty, I met the students at the end of their regular classes so that I could explain the details of the research study and provided them the information sheets and the consent forms. It was a week before the day of collecting the data so that the students could have time to decide whether to participate in the study. I also asked the staff to help arrange computer rooms at the faculty for administering the tests. The information sheet and the consent form for the participants are shown in Appendices B and C.

4.5   Preparing research tools

I prepared four research tools, which are a questionnaire, the receptive test (VKS, categories III and IV), the controlled productive test (VKS, category V), and the free productive test (essay-writing tasks).

4.5.1   Questionnaire

A questionnaire was used for the information about their student ID so that the questionnaire is identifiable. Another section asks the participants whether they had an experience of studying or living abroad.

Those who were below 18 years old or those who lived abroad for more than one year would be excluded from the study since they did not represent the majority of English-major students in Thailand. Normally, only a few of Thai high-school students in Thailand would have a one-year experience abroad as exchange students. The last part of the questionnaire is about activities that the participants usually did and the amount of time spent practicing English outside class.

The purpose of including a questionnaire in the study was only to know more about their English learning outside class and about their background. Even though the research questions were not related to experience of studying, living abroad, or different ways in practicing or learning English, these to some extent had an influence on vocabulary knowledge development, according to some research studies in the field (e.g., Laufer and Paribakht, 1998; Dizon and Tang, 2017; Min, 2008). Therefore, these were included in the questionnaire, and they were used to explain some of the unexpected findings in the discussion chapter too. The questionnaire is shown in Appendix G.

4.5.2 Receptive test (categories III and IV of VKS)

The second research tool I prepared was the receptive test, which was in the format of the categories III and IV of the VKS. All the five groups of words used in the VLT were used in this test. However, I used a different unit of counting a word from the VLT. As explained in Chapter 4, I used the lemma to count the items in the test instead.

There were 30 lemmas at each of the frequency level, so there are 150 items altogether for the receptive test. The test words in Schmitt et al.'s (2001) VLT has "the 3:2:1 (noun: verb: adjective) ratio" (Schmitt et al., 2001, p. 72). Hence, the test words in the receptive test (Version 1) included nouns, verbs, and adjectives. However, after the pilot study, only the first ten items were selected (see Section 5.5.4) and the data supplied by the participants in the receptive and the controlled productive vocabulary tests represented the participants' knowledge of nouns.

The test was divided into two equal parts (seventy-five items each) – the first half to be tested on Day 1 and the second half for Day 2 to avoid the fatigue effects from the participants. However, this was piloted to see whether it was too long for the participants or not.

I reordered a sequence of the test lemmas. As Nation and Coxhead (2014) point out, "The order of items seems to affect the amount of effort put into each item". This happened in a previous study by Nation and Coxhead (2014) using the computerised VST. Without the mixture of frequencies, the participants randomly guessed the answers and took shorter time than expected to complete the items in lower word-frequency levels.

This shows that when facing a lot of difficult words (low-frequent words) in a row, test takers might be too exhausted or demotivated to complete the test. Therefore, the order of the test lemmas from item 1 to item 5 were one lemma from the 2,000-word

level, one from the 3,000-word level, one from the 5,000-word level, one from the 10,000-word level, and one from the AWL. This pattern was adopted throughout the test.

In terms of types of answers, as there would not always be one-to-one equivalent translations for some lemmas between two languages (Nation, 2013), the three following options were accepted as answers to the receptive test.

1) a single-word Thai translation,
2) a multi-word Thai translation as in Zhong (2016)
3) an English synonym as in the study by Paribakht and Wesche (1997)

The use of Thai translations and English synonyms were advantageous to both the test takers and the graders. Test takers did not need to provide a long clause or sentence to define any word that did not have an equivalent in the Thai language. If English synonyms were what they were familiar with or what they can recall, they could use the English synonyms as the answers to show their understanding of the lemmas. The test takers could also save their time while doing the test and the raters could save their time while checking and scoring the test. The receptive test is presented in Appendix H.

4.5.3   Controlled productive test

The third research tool I prepared was the controlled productive test. It was in a sentence-writing format (the category V of the VKS). The test lemmas in this test were the same as those in the receptive test. The participants would be asked to use the given lemmas in sentences. However, a test taker might write a neutral sentence. To avoid neutral sentences, they were required to use conjunctions. This technique was used in Zhong's (2016) study. Only 'so' and because' were used as they are easy to use, and Thai students were familiar with them. These two conjunctions link cause-effect clauses, so the participants would carefully think through and produce meaningful sentences. When the participants supplied a sentence, they were asked to also provide a translation/ synonym for that target word as well to ensure that they knew the meaning of the target word. If the students knew only the meaning of a lemma given but did not know how to use it in a sentence, they could leave the box for a sentence blank.

Both the receptive test and the controlled productive test were combined in one spreadsheet. The test takers need to finish them both before starting to do the free productive test. Moreover, these 150 lemmas in both the receptive test and the controlled productive test had never been trialled in the format of sentence writing by

any previous studies. One hour was then roughly set as the time duration for both tests. The time duration would be adjusted after the pilot study. The test is presented in Appendix H, together with the receptive test.

4.5.4   Free productive test

The last research tool that I prepared was the free productive test. It consists of two tasks: Essay 1 for Day 1 and Essay 2 for Day 2. The subjects would be asked to write two 300-word essays (300 running words). Two topics from the study by Laufer and Nation (1999) were adopted.

Essay 1: "*Should a government be allowed to limit the number of children a family can have?*"

Instruction: Write an essay of 300 words on the topic provided. Discuss this idea considering basic human rights and the danger of population explosion.

Essay 2: "*A person cannot be poor and happy, because money is always needed to gain something that is important to that person.*"

Instruction: Write an essay of 300 words on the topic provided. Argue for or against this idea.

These two essay topics used in the study by Laufer and Nation (1995) were selected. According to Laufer and Nation (1995), these topics are controversial topics, which encourage the test takers to provide different groups of words but not technical words, which were not the focus of the study. The test takers do not need to have special knowledge to write about them. The suggested topics have been tested with multicultural participants (Chinese, Japanese, Thai, Samoan, Polish, Malaysian, Russian, and Israeli) in the study by Laufer and Nation (1995). Therefore, these two controversial topics were adopted. The time duration for each essay was set for one hour, equal to that in the study by Laufer and Nation (1995). The free productive test (both Essay 1 and 2) is shown in Appendix I.

4.6   Plans to analyse the research tools and the rationale

This section presents how I planned to analyse each of the research tools before the pilot study and the main study. This includes scoring systems, references, and statistics to be used. However, I described how I analysed data of the main study in Section 6.3.

4.6.1   Plans to analyse the questionnaire

The information from the questionnaires would be converted into numbers and analysed by using descriptive statistics. It was piloted. The participants' background information is shown in Section 7.2.

4.6.2   Plans to analyse the receptive test

Coding is needed before grading the participants' answers. The flow chart of how I analysed the participants' answers on the test is shown in Section 6.3.2, and the findings are presented in Section 7.3.

When scoring, I used reliable and widely-used online dictionaries as references:

1. Cambridge English-Thai Dictionary Online (*Cambridge Dictionary (English-Thai Dictionary)*, 2020)
2. NECTEC's Lexitron Thai-English Dictionary Online (*Thai-English Electronic Dictionary Lexitron*, 2020)
3. Oxford Advanced Learners' Dictionary Online (*Oxford Advanced Learners' Dictionary*, 2020)

The Cambridge English-Thai Dictionary Online was chosen because it is one of the reliable and well-known dictionaries available online and has a bilingual (English-Thai) version. It shows many senses of a lemma with an example. If such translation did not match the translations provided in this dictionary, I would check it against another dictionary – a Thai-English dictionary.

NECTEC's Lexitron Thai-English Dictionary Online was also chosen to grade the translations supplied by the participants because it is a reliable online dictionary developed by a Thai government organisations and leading universities in Thailand such as National Electronics and Computer Technology Centre (NECTEC), National Science and Technology Development Agency (NSTDA), Sirindhorn International Institute of Technology (SIIT), Chulalongkorn University, Kasetsart University, etc. The dictionary also presents English synonyms and some other Thai compound words of such lemma, together with its English equivalents.

Oxford Advanced Learner's Dictionary Online would be used to check when a test taker provided an English translation or synonyms. The dictionary was chosen because it provides English synonyms and definitions of a lemma with some examples of sentences. It was chosen because it can be used as a reference when I checked the answers in the controlled productive test. The reasons I chose this dictionary to check the data in the controlled productive test are written in Section 4.6.3

Stewart et al.'s (2012) three-scale scoring system was adopted: correct, partially correct and incorrect. This is because learners might have partial receptive knowledge of the test lemmas. Two points would given to a correct Thai translation or an English synonym. One point would be given to a partially correct answer or to the one with a related sense. Zero point would be given to a wrong answer (not relevant to the meaning of the tested lemma at all) or when no translation or synonym was provided.

4.6.3   Plans to analyse the controlled productive test

Before the participants took the test, I would not inform them of the part of speech of the test lemmas. This allows them to provide what they can first think of when seeing the lemmas. I acknowledged that different senses of a lemma can fall into different frequency levels, but this could show their real vocabulary knowledge.

Only the content words were the focus, and the grammar points to be checked depended on the word classes of the lemmas they provided.

- If they used a tested lemma as a noun, I would check its countable/uncountable form, singular/plural form, its modifiers, article, verb, adjective, pronoun, and context.
- If they used a tested lemma as a verb, I would check its subject, object, adverb, and context.
- If they used a tested lemma as an adjective, I would check its noun, adverb, and context.
- If they used a tested lemma as an adverb, I would check its verb, sentence, and context.

I used two tools to check the meaning and the correctness of the lemmas. They are:

1. Oxford Advanced Learner's Dictionary Online (*Oxford Advanced Learners' Dictionary*, 2020)

2. Text analysing software called the Sketch Engine (Kilgarriff et al., 2004; 2014) and English Web 2013 (enTenTen13) corpus.

Oxford Advanced Learner's Dictionary Online was chosen because it shows a lemma's part of speech and meaning senses. It provides many examples and grammatical patterns or collocational behaviours for each meaning sense. The presentation of the website is easy to use, and one page contains the information of one part of speech only, so not too much information is presented on the same page. Even though other well-known online dictionaries like Longman, Collin, and Cambridge dictionaries also show various meaning senses with a lot of examples, they show all the parts of speech for a word (not a lemma) in one page. A word's grammatical patterns or collocational behaviours are presented after all the meaning senses, not specific to a particular sense. Hence, I decided to use Oxford Advanced Learner' Dictionary Online to grade the lemma use in sentences supplied by the participants.

Another tool I used when grading the participants' answers in the controlled productive and the free productive vocabulary tests was the Sketch Engine (Kilgarriff et al., 2014).

The Sketch Engine was chosen because it "is a leading corpus tool. … The Sketch Engine website offers many ready-to-use corpora" (Kilgarriff et al., 2014, p. 7). It has been widely used by lexicographers, teachers, students, researchers in many fields such as linguistics, language, computing, translation, and discourse analysis, and it has been used in a variety of countries such as "China, the Czech Republic, Germany, Italy, Japan, Spain and Taiwan as well as the UK" (pp. 15-16).

> "[The function word sketch is] a one-page summary of a word's grammatical and collocational behaviour. … The word sketch can be seen as a draft dictionary entry. The system has worked its way through the corpus to find all the recurring patterns for the word and has organised them." (Kilgarriff et al., 2014, pp. 9-10).

One of the preloaded corpora that was chosen to check the lemma use against was 'English Web 2013 (enTenTen13) corpus'. The English Web 2013 (enTenTen13) corpus is a large corpus which consists of approximately 20 billion words (or nearly 23 billion tokens) collected from texts on the Internet from various domains where English is the official language such as UK domain, Australian domain, Canadian domain, US domain, and Wikipedia.

They "are built using technology specialized in collecting only linguistically valuable web content. … [T]exts were checked manually and content with poor quality text and spam was removed" (Sketch Engine Website).

I then decided to use this corpus when I graded vocabulary used by the participants in the controlled productive and the free productive vocabulary tests.

However, I did not choose the English Web 2015 (enTenTen15) corpus at the time of data analysis because it was smaller than the English Web 2013 (enTenTen13) corpus. I did not use a larger and a more recent corpus like the English Web 2018 (enTenTen18) corpus or the English Web 2020 (enTenTen20) corpus because it had not been made when I analysed the data. I did not use the Timestamped JSI web corpus 2014-2020 English, which contains over 57 billion words (or nearly 67 billion tokens) because its sources are news articles, which is too specific to the news genre.

When these two tools could not be used to check the correctness of the lemma use, the problematic sentences would be checked by Two native speakers – British teachers (hereafter called as 'raters'). The information sheet for the raters is presented in Appendix F. They would be given a scoring sheet with the same set of data and would be explained on how to grade the sentences. They needed to score the first 20 sentences together and the rest on their own. Their scores would be calculated for inter-rater reliability. After that, they would consult and compromise the scores for the sentences that they provide different scores.

Like the receptive test, two points would be given to the sentence in which the tested lemma is used correctly both in semantics and grammar. One point was given when the tested lemma was used semantically correctly in the sentence. No point is given when the tested lemma was used incorrectly in terms of their meanings and/or grammar or when no sentence was provided. After I scored the participants' answers, all the scores were analysed by using descriptive statistics, an independent samples t-test, ANOVA, and MANOVA.

Apart from the dichotomous approach with the three-scale scoring, I also planned to analyse the vocabulary development with the continuum approach. Therefore, Paribakht and Wesche's (1997) five-scale scoring system was used to track vocabulary changes at different frequency levels and at different years of study as well.

Score of 1 for no knowledge of the lemmas at such frequency level

Score of 2 for recognising the written form of the lemma

Score of 3 for recognising the written form of the lemma and knowing the lemma's meaning

Score of 4 for recognising the written form, knowing the lemma's meaning, and being able to use it with semantic appropriateness in a sentence

Score of 5 for recognising the written form, knowing the lemma's meaning, and being able to use it with semantic appropriateness and grammatical accuracy in a sentence

4.6.4   Plans to analyse the free productive test

For the consistency to the receptive and the controlled productive tests, I would follow these procedures:

1) A three-scale scoring system was used.
2) Content words were the focus.
3) Grammar was checked by using Oxford Advanced Learner's Dictionary and the Sketch Engine (Kilgarriff et al., 2014).
4) The two raters graded some sentences that could not be checked by using the references.

However, two more steps that needed to be completed before grading the participants' lemma usage are:

1) lemmatising the participants' free writings
2) classifying the lemmas produced by the participants into five frequency levels.

The details are presented in flow charts in Section 6.3.4. After that, I analyse all the scores by conducting descriptive statistics, an independent samples t-test, ANOVA, and MANOVA on SPSS. Then the statistics from the three tests were compared to explore the relationship among receptive, controlled productive, and free productive vocabulary knowledge. This would be done with both dichotomous and continuum approaches to answer the fourth research question.

To summarise, this study used four research tools. The questionnaire was used for the general background information of the participants. The receptive test was in the format of self-reported scale with a translation task. The controlled productive test was in a sentence-writing format. These two tests included the same test words from the VLT (Schmitt et al., 2001). The sequence of the test lemmas was reorganised; low-frequency lemmas were mixed with high-frequency lemmas. The free productive test was an essay-writing task. The topics were chosen from the study by Laufer and Nation (1995). All the words produced by the participants were lemmatised and

classified into the five frequency levels. The three-scale scoring system as in the study by Stewart et al. (2012) was used with all the tests. Paribakht and Wesche's (1997) five-scale scoring system was used with the receptive and the controlled productive tests. The results were analysed in relation to frequency levels (also including academic word level) and to years of study. However, only the questionnaire and the first two tests were piloted for necessary adjustment.

Chapter 5 Pilot study

5.1 Introduction

This chapter presents the objective of the pilot study, its procedures such as approaching the participants, administering the pilot study, how I adjusted the research tools and what I learnt from the pilot study.

5.2 Objectives of the pilot study

The pilot study was done in November 2018 with the attempts to ensure the effectiveness of the receptive test and the controlled productive test in terms of the time duration, the instructions, the sequence of the test lemmas, and chances of producing neutral sentences.

The time to complete 75-item receptive and controlled productive test for Day 1 was set as one hour. The pilot was done to check whether one hour was an appropriate time duration for the participants to complete all the test items. The instructions with the five self-reported categories were checked whether it was clear to the participants or not. The word sequencing of the test words was checked to ensure that it was effective to obtain the participants' attention throughout the tests. The controlled productive test was piloted to check whether including conjunctions would help reduce the chance of producing neutral sentences or not. Finally, the questionnaire was also tested during the pilot to ensure the clearness of all the questions before it was used in the main study.

5.3 Sample size of the pilot study

In a previous study piloting a vocabulary test by Bayazidi and Saeb (2017), four participants (approximately 5%) were sampled from the same educational environment as the projected participants (75 participants). Their pilot study was conducted with the same objective as the current study, which is to check the overall test design. If this percentage was applied to the present study, 5% of the 200-240 target English-major students were 10-12 participants. This number is also the minimum number of the samples for a pilot study suggested by Isaac and Michael (1995). Therefore, approximately ten Thai students and graduates were recruited. The convenience sample was used because:

1) The subjects were easily accessible.
2) The subjects were willing to participate
3) It was inexpensive
4) It allowed me to obtain some basic data on how to adjust the receptive test, the controlled productive test, and the questionnaire.

The participants of the pilot study were ten Thai students and graduates in the UK and in Thailand. They were easy to approach because I contacted them through my contact lists on social media platforms (Line and Facebook). I approached them and those who consented to participate were recruited. I used video calls during the pilot study, so it was inexpensive and fast. I did not need to travel to Thailand to meet them in person to conduct this pilot study. The pilot study gave me a basic idea of whether the research tools were effective or not.

Although convenience sampling has some criticisms of selection bias and not being representative of the entire population, the participants of this pilot study were carefully approached. I tried to approach some English majors and graduates from English major first. Some of them had just graduated, but some had graduated for three years. There were two non-English-major undergraduates: one studying in the UK and one never lived or studied in an English-speaking country. Therefore, the one who did not major in English and had never lived in an English-speaking country could represent low-proficiency English-major students and those who majored in English and/or used to live an English-speaking country could represent the high-proficiency ones. Besides, their mother tongue was Thai. They were from different regions throughout Thailand, and they tended to use a lot of English in their class or when they were in their classes.

These ten students of the pilot study were not included in the main study because they had seen the test words and this would affect their scores in the main study. Moreover, the information sheet and the consent form for this pilot were prepared and given to the participants.

5.4   Procedure during the pilot study

5.4.1   Approaching the participants

After the ethics application was approved by ESSL, Environment and LUBS (AREA) Faculty Research Ethics Committee University of Leeds in October 2018 (Appendix A), I approached some Thai students in the UK and some graduates who majored in

English in Thailand through one social media application called 'Line'. I explained to them about my research study such as the rationale of this pilot study, the reasons why they could represent the projected participants, the aim of the main study, and that their participation would be kept confidential. I gave them an information sheet (Appendix B) and a consent form (Appendix C). They all had one week before making their decision whether they would participate in the pilot study.

After one week of the approach, I asked them again through Line one by one whether they would participate in the study. They signed the consent forms and sent back to me. After that, the date was set for the pilot study. The Thai students in the UK met me in person and they used my laptop to complete the tests, while those living in Thailand completed the tests on their computer through video calls.

5.4.2   Administration of the pilot study

All the participants completed the questionnaires and all the tests, except the free productive test, by using a computer. The pilot study was arranged to be as similar as possible to the main study. All the soft files of the blank questionnaire and the tests were uploaded on my personal Google Drive. The shareable links of these files were given to the participants. After the participants downloaded all the files, I explained the details of the pilot study once again. I also explained to them in Thai about how to fill out the questionnaire. Then I explained how to do the tests. The full version of the receptive test and the controlled productive test contained 75 items, and the time was set as one hour per set.

However, they were able to continue after one hour until they finished two sets (150 items) of the tests. I explained all the self-reported scales to them. I emphasised that if they knew how to use any test lemmas in a sentence level they should select number 5 and write a sentence using the test word with a conjunction either 'so' or 'because'. I also explained that the translation could be a single-word or a multi-word translation or a synonym and that they could ask me anytime during the tests. Before they started doing the tests, they were told that this would be timed so that the time duration could be adjusted in the main study.

5.5   Phases of the pilot study

There were three phases of the pilot study piloted with the ten participants. All the tests in each phase were timed so that they were adjusted for the next phases. The details of the tests and the numbers of the participants are shown in Table 5.1.

Phase	Research tools	Number of Items	Number of Participants
1	- Version 1 (Full versions) - Questionnaire	75 items each set	4 participants
2	- Version 2 (10 words from each word level) - Questionnaire	50 items each set	3 participants
3	- Version 3 (5 words from the 2,000-level and 10,000-level) - Questionnaire	55 items each set	3 participants

Table 5.1: Three phases of the pilot study

5.5.1 First phase

In the first phase, the full versions of the receptive test and the controlled productive test, together with the questionnaire, were piloted. There were 75 test lemmas for each set of the tests; the two sets included 150 lemmas.

The participants could have ten minutes or longer for the break before they took the second set of the tests. The first four participants took the tests on 5th November 2018. Pseudonyms are used as the participants' names. The average of the time durations they spent on the tests are shown in Table 5.2.

Participants	VDO call/ In person	Date	Time for Set 1 (minutes)	Time for Set 2 (minutes)
1. Wendy	In person	5 Nov 18	75	63
2. May	Video call	5 Nov 18	200	210
3. Kacee	Video call	5 Nov 18	100	75
4. Beca	Video call	5 Nov 18	65	70
Average time duration			110	105

Table 5.2: Average time duration for phase 1 of the pilot study (Full versions of the receptive test and the productive test)

As shown in Table 5.2, the time that the first four participants spent doing both sets of the tests exceeded the time that I expected them to finish the tests. The average time for the first set of the tests was 110 minutes or nearly two hours, and that for the second set was similar, 105 minutes. One participant, May, spent the longest time – 400 minutes for both sets. This might be because she completed almost all of the items and provided long sentences for each test word, so she spent almost seven hours completing these tests. Therefore, 75 items appeared to be too many for the projected participants to complete.

5.5.2 Second phase

The average time spent by the first group of the participants was too much, so I removed some test lemmas in the second phase of the pilot study. I decreased the number of items for each set to 50 items each, or 100 items for two sets. The first ten test words at each of the frequency level in Schmitt et al.'s (2001) VLT were selected. As most of the test words in Schmitt et al.'s (2001) VLT are nouns, the ten words were all nouns and the findings from the tests would represent the knowledge of nouns.

Participants	VDO call/ In person	Date	Time for Set 1 (minutes)	Time for Set 2 (minutes)
1. Caden	Video call	6 Nov 18	66.5	50
2. Brian	In person	8 Nov 18	45	50
3. Tony	In person	8 Nov 18	45	54
Average time duration			52	51

Table 5.3: Average time duration for phase 2 of the pilot study (Version 2 of the receptive test and the controlled productive test – 10 items for each word level)

Table 5.3 shows the average time that another three participants in the pilot study took to complete the first set and the second set of the adapted versions of the tests. On average they spent 52 minutes for the first set and 51 minutes for the second set of both tests.

These time durations were what I expected the projected participants to spend in completing the receptive and the productive test. However, I tried adjusting the tests one more time as Phase 3 in order to know which phase would be the most appropriate for the main study.

5.5.3   Third phase

In Phase 3 of the pilot study, ten words of the 2,000-word level and ten lemmas of the 10,000-word level were taken out from each set of the tests. Therefore, there were five lemmas left for the 2,000-word level, five lemmas left for the 10,000-word level, but still 15 lemmas for the rest of the levels (3,000 level, 5,000 level, and the academic word level). The total number of the lemmas for each set of the tests in Phase 3 was 55 lemmas. The 15 nouns from each of the frequency level were selected as well.

Participants	VDO call/ In person	Date	Time for Set 1 (minutes)	Time for Set 2 (minutes)
1. Tiana	Video call	10 Nov 18	52	36
2. Paul	Video call	10 Nov 18	37	31
3. Yasmine	Video call	10 Nov 18	40	35
Average time duration			43	34

Table 5.4: Average time duration for phase 3 of the pilot study (Version 3 of the receptive test and the controlled productive test – 5 words from the 2,000-word level and the 10,000-word level)

The average time the last three participants spent doing the tests was less than I expected although there were more test lemmas than those of Phase 2. As shown in Table 5.4, the average time for the first set of Version 3 of the tests was 43 minutes and 34 minutes for the second set. This is probably because most of the difficult lemmas at the 10,000-word level were removed and the participants did not need to spend too much time thinking. It is probably because most of the words at the 2,000-word level were removed, and they did not need to write many sentences of these easy test lemmas.

5.5.4   Decision after the pilot study

After the last phase of the pilot study, I decided not to use the full versions of the tests as there were too many test lemmas and the participants would take too long to finish the tests.

- Most of the participants took a lot of time to finish the tests. The longest time that one of the participants took was up to seven hours as shown in Table 5.2. It was almost four times longer than the time expected. It was quite difficult to complete all the 150 items within two hours or 120 minutes. Also, they all complained that there were too many items.
- I decided to use only one set of the receptive test and one set of the controlled productive test in the main study to avoid fatigue effect. When I considered the issue of consistency, the second version of the tests (10 items for each of the five frequency levels) seemed to be the best version since they had the same number of test words at each of the five frequency levels. The total scores for each frequency level would represent the whole vocabulary at such level. Also, the findings would represent the receptive and the controlled productive knowledge of nouns at the 2,000-, 3,000-, 5,000-, 10,000-word levels, and the academic word level.
- Regarding the answers supplied, the participants did quite similarly in both sets of the tests. That is to say, the results of the two sets of the tests were not much different. Hence, either one of the sets could be chosen and used in the main study. Randomly chosen, the first set of Version 2 was to be used in the main study.

5.6   Revising the research tools

This section presents the problems when the participants in the pilot study completed the research tools. The problems lead to how I revised them for the main study so that I could avoid these problems in the main study.

5.6.1   Revising the questionnaire

I revised the questionnaire in several points:

- In the section asking about the participants' experience abroad, the question about 'an exchange student' was too specific. Some participants had never

been exchange students but used to live in the English-speaking countries. Therefore, I used a wider term instead. Therefore, the question "Have you ever been an exchange student in any English-speaking country?" was changed into "Have you ever lived or studied in any English-speaking country?"

- In the section of outside-class activity to practice English, one of the participants in the pilot study could not estimate how often and how much time she spent on one on her outside-class activities. She wrote 'one book' to answer the question, instead of giving numbers of hours. The other participants did not have a problem with it. To avoid different types of answers or any mistakes, validating every cell in the excel file was necessary. Hence, I validated the cell and only numbers can be typed in. I also did this to the data cells for their student ID. Only seven number digits can be typed in the cell. In case that other information or not seven-digit number was typed in, there would be an error pop-up message informing that the participants had typed in the wrong information.

- To ensure the participants understand all the questions in the questionnaire, there were Thai translations below each question. The validating pop-up boxes were also written in Thai so that it was easily understood by the Thai participants.

- To know more about the participants' educational background such as the programme they studied in their secondary schools was added.

- I adjusted the format – the position of all the data cells – for the convenience in transferring what the participants supplied into the SPSS more easily.

- I added the section of the information such as their name, their student ID, their year of study, their email addresses, and their telephone number in the questionnaire so that the questionnaires would be identifiable to me and that I could contact them to inform them of their test scores.

- I shortened some sections so that the questionnaire looked clearer and easy to read for the participants. For example, the number of the English-speaking countries the participants have lived or studied in was decreased to three countries. I also rewrote some questions. The question "How often per month?" was changed into "How many times per month?", and the question "How long each time?" was changed into "How many minutes each time?". Finally, I added one more option "Prefer not to say" in the section asking their gender. This was for the ethical issue.

In conclusion, learning the problems from the pilot study and anticipating what would happen in the main study helped me adjust the questionnaires. The overall

adjustment was to validate the data cells and to make the questionnaire identifiable and understandable for the participants.

5.6.2   Revising the receptive test and the controlled productive test

This section presents how the participants supplied the answers to the receptive test and the controlled productive test, how I learnt from this, and how I revised the tests.

- In the column of translation or synonym in the receptive test, the participants provided one-word Thai translation, multi-word Thai translation, synonym as expected. Many of them explained the meanings in English phrases as well. Hence, I also included an English or Thai phrasal explanation as another option.
- Some participants in the pilot study did not provide a sentence even though they chose Number 5 on the knowledge scale and provided a correct translation or synonym. Therefore, in the main study I needed to emphasise this to the participants before the tests.
- Despite being informed to use either of the two conjunctions 'so' or 'because', one participant used other connecting words throughout the whole tests, and some did not use 'so' or 'because' in a few of their sentences. For example, for the tested lemma 'belt', one of them supplied the sentence '*Fasten your seat **belt** <u>before</u> the plane take off*.' Moreover, some of them also produced simple sentences or did not use the conjunctions given. Even though these sentences were not neutral sentences, it was important for me to remind the participants to use one of the conjunctions designated in their sentences before administering the tests. I also created a pop-up message (with a Thai translation) to remind them when they clicked the sentence box.
- One of the participants in the pilot study pointed to me that there was a typo in the tests. Hence, I checked and corrected the typo before giving it to the projected participants.

5.7   Revising the free productive test

Even though the free productive test was not piloted, what I learnt from the pilot study gave me some ideas of how I should do to adjust the free productive test to make sure that the projected participants would produce the essay as I expect. Here are how I adjusted it.

- Similarly, instructions in Thai would also be provided with the English instructions so that it was clearer for the participants. I also needed to emphasise how to complete the test to the participants before the tests as well.
- Concerning the time duration, it seemed necessary that only one set of the free productive test be used to avoid a fatigue effect.
- The topics of the two sets were quite different. The topic of the first set was related to government and family planning, while that of the second set was related to money and happiness. The second topic seemed to be more suitable for English-major students in Thailand as it was more relevant to them. It could encourage them to write more about it. The first topic was not relevant to the students at this age, and the policy of limiting the number of children in a family was not implemented in the Thai context. Therefore, the first topic would not be used in the main study.

5.8   Lessons learnt from the pilot study

I have learnt some lessons from the pilot study. First, it is often that things do not go as planned. For example, I expected that the participants on average could finish one set of the receptive test and the controlled productive test within one hour, but it turned out that they need more time than that. Hence, it is ideal to pilot the research tools before using them in the main study.

Second, clear instructions of the tests are vital. I should also provide the translated instructions and emphasise how to complete the tests before administering them. In this way, I could get the answer types I needed so that I can answer the research h questions. I also needed to prepare a PowerPoint slide showing how to do the test. The slide needed to be easy and clear for the participants to follow.

Finally, I decided to use only one set of all the tests – the first set of the receptive test and the controlled productive test and the second set of the free productive test. It was to avoid a fatigue effect. All the questions and the test instructions were presented with Thai translations so that the participants could understand them clearly and complete them correctly.

Chapter 6 Main study

6.1   Introduction

Chapter 6 presents the data collection procedures, covering how I contacted the university where the projected participants were studying, how I prepared the research tools for the main study, how I approached the participants, and how I administered the research tools. All these are presented in Section 6.2. The analysing methods and the flow charts of these processes is presented in Section 6.3.

6.2   Data collection procedure

6.2.1   Contacting the university in Thailand

After I had been formally granted permission from the dean of the Faculty of Liberal Arts of a university in Thailand, I had an informal meeting with the dean and deputy deans to confirm when I would collect my data and explained to them that I needed to use some facilities such as computer rooms. I also informed them that this data collection would last approximately one month. The executive team were all in agreement and gave me a formal permission to collect the data from the students. After that, I contacted the relevant staff to confirm the class schedule, to ask for the contact details of the teachers who taught all the English-major students, and to check the number of the functioning computers and the available computer rooms.

6.2.2   Preparing the research tools for the main study

The second thing I needed to do was preparing all of the revised research tools in different forms for the main study. I printed the information sheets and the consent forms so that I could give them to all the students when I approached them the first time. The tests used in this study were computer-based tests. I saved the files of the tests in one folder, named it 'Test (Peeraya)' and saved it in 'Documents' folder of every computer so that the tests were ready for the participants in the following weeks. I also backed up the test files in my external disk, uploaded all the tests on my personal Google Drive and saved the shortened access link. If there was no

internet connection on the days of the tests or that the test files were lost from any computers, I would still have the tests ready for all the participants.

6.2.3 Approaching the participants

Before I approached the English-major students, I checked the class schedule to find the appropriate time to approach them in their classes. I began by emailing to all the teachers teaching the English-major students in this semester. I explained to them about my project and asked them for their permission to approach the students in their classes at a particular time. The teachers all allowed me to do so: some at the beginning of the classes, some in the middle, and some at the end of the classes.

In the classes, I introduced myself to the students, gave them the information sheets and explained to them briefly about the research study such as the purposes of the study, the reason why I approached them, what they needed to do if participating, and the expected benefits of the study to the faculty, to themselves, and to teaching and learning English. I gave them the information sheets and the consent forms. In the following week, I went to their classes again to get consent forms from the students who were willing to participate in this study. Then I informed them of the first two time slots for the tests.

In the first two time slots, there were not many students participating – only 24 students (15 first-year students, 5 second-year students, and 4 fourth-year students). This was because the rest of the students were not available at the first two time slots I arranged. To recruit more participants, I consulted with four students from different four year groups about the appropriate time for the students. Then I prepared eight more time slots over four days in the following two weeks so that the rest of the students with consent could sign up at the slots they were available.

Finally, out of the 225 English-major students from the four year groups in that academic year, 111 students participated in the main study. There were 34 first-year students, 24 second-year students, 29 third-year students and 24 fourth-year students. The number of the participants did not reach the optimal number for the participants per variable as suggested by Cohen et al. (2017). When comparing their data, I then grouped the first-year and the second-year students as the lower year group (58 students) and the third-year and the fourth-year students as the higher year group (53 students).

6.2.4 Administration of the tests

On the days of the tests, I presented the details of the research study to the participants by using a PowerPoint slide. I informed them of the purposes of the study, what kind of tests and research tools they needed to complete, how to complete them, the estimated time duration for completing each of them, what they needed to do when they did not understand, what they needed to do at the end of the tests, and where the tests were located in the computers.

There were four research tools for them to complete: the questionnaire, the receptive test, the controlled productive test, and the free productive test. However, the second and the third ones were combined in one spreadsheet. The estimated time durations for each research tool were the following: five minutes for the questionnaire, one hour for the receptive test and the controlled productive test, and another one hour for the free productive test. However, these durations were only the expected durations based on the time spent by the participants in the pilot study.

In the main study, any participants who finished the questionnaire, the receptive test and the controlled productive test could take a break and have some refreshments outside the computer room. Once they finished having their refreshments and when they were ready to start doing the last test – the free productive test, they could start at any time. They could spend as much time as they needed to finish all the tests.

There were three to four invigilators including me in a computer room at each time slot. If any students did not understand anything in the tests, they could ask for help or for more explanation. In case of any technical problems, one invigilator who was an IT staff member in the faculty could help solving it.

Once the participants finished the questionnaire and the tests, they needed to inform one of the invigilators. The invigilators checked whether every file was completed before saving the files into an external disk and asking the participants to send all the files to my university email address (edpu@leeds.ac.uk) in order to back up the data. After the files were sent to my email address, I opened the files and checked them the second time. When there was nothing to be added or edited, the participants were asked to delete the test files they had completed. To thank them for participating in my research, I gave them some stationery such as pens, pencils, and highlighters before they left the test room.

Because the time duration for completing the tests were not limited, the participants did not seem stressed. They could do the tests at their own pace. Some left the computer room after the first hour, while some spent almost three hours to finish everything. When every one of the participants left the computer room in each test time slot, all the invigilators checked whether the test files on the desktop

completed by the participants were permanently deleted in order to prevent other groups of the participants from seeing other students' answers. The invigilators also checked that the blank test files were still in the folder 'Documents' before turning off the computers.

After analysing the answers of all the participants, I gave feedback to them separately via email to tell them their scores of the receptive and the controlled productive tests at each of the five frequency levels. I explained to them that the score of 80% or above means that they knew or were able to recall most of the words at such frequency level. I also recommended links of word lists for them to practice more in case they got low scores at some levels.

6.3   Data analysis

6.3.1   Analysing the questionnaire

The data from the questionnaires had been already typed by each participant in a computer-based questionnaire. The data contained:

- their personal data (the participants' full names, student ID numbers, years of study, ages, genders, and contact details – their email addresses and mobile numbers)
- their educational background (the types of their secondary school and the learning programmes they studied in their secondary schools)
- their exposure to the English language (numbers of years they have been studying English in their life, their experiences studying or living abroad, the activities they usually do to practice their English outside class and numbers of minutes they spend each time of doing those activities)

This data was analysed by using descriptive statistics. For some open-end questions such as their experiences studying or living abroad, the numbers of months they spent in those countries were summed for each participant so that the data was suitable for the statistics programme and can be easily understood or interpreted.

The data on their activities they did to practice their English outside class was classified based on the types of the activities such as listening, watching, speaking, reading, writing, and others. The numbers of minutes for each type of activities were summed and converted into numbers of hours for each participant. The data is presented in tables and graphs in Section 7.2.

6.3.2   Analysing the receptive test

I began analysing the receptive test by unmixing all the test lemmas so that the lemmas from the same frequency word levels could be grouped together. The scores were then imported to the SPSS more easily.

Next, I coded the participants' answers by reading through all the translations/ synonyms/ phrases provided by the participants, checking them against the references (Cambridge English-Thai Dictionary and Thai-English Dictionary by NECTEC and Oxford Advanced Learner's Dictionary), coding them, classifying them into three categories and giving scores to them.

- Category A for correct answers (two points given)
- Category B for partially correct answers (one point given)
- Category C for incorrect answers (zero point given)

The following are the details of the three categories and how I coded the participants' answers.

The answers fell into **Category A** (two points given) when:

- their meanings were correct (regardless of grammatical mistakes, senses, parts of speech of the translation/ synonym/ phrase) – coded as 'CMN'.
    - Example: For the tested lemma 'sport', the translation 'กีฬา' (ki-la, meaning 'sport') was coded as 'CMN'.
    - Example: For the tested lemma 'palm', both senses of the tested lemma 'ต้นปาล์ม' (ton-pam, meaning 'a palm tree') and 'ฝ่ามือ' (fa-mue, meaning 'an inner part of a hand') were coded as 'CMN'.
- they were hypernyms of the test words – coded as 'Hypernym'.
    - Example: For the tested lemma 'salary', the translation 'เงิน' (ngoen, meaning 'money') or 'รายได้' (rai-dai, meaning 'income') was coded as 'Hypernym'.
- they were English loan words in Thai – coded as 'Loan word'.
    - Example: For the tested lemma 'balloon', the translation 'บอลลูน' (bon-lun, meaning 'balloon) was coded as 'loan word'.

The answers fell into **Category B** (one point given) when:

- their meanings were related to those of the test lemmas – coded as 'RSense'.
    - Example: For the tested lemma 'victory', the English answer 'winner' then was coded as 'RSense' because its meaning was related.
- they included the test lemmas as compound nouns – coded as 'CmpndW'.

- o Example: For the tested lemma 'birth', the translation 'วันเกิด' (wan-koet, meaning 'birthday') showed that the participant had partial knowledge of its meaning. It was coded 'RSense'.
- they were meronyms of the test lemmas – coded as 'Meronym'.
  - o Example: For the tested lemma 'palm', the translation 'ลายนิ้วมือ' (lai-niw-mue, meaning 'fingerprint') was partially correct and coded as 'Meronym'.
- they shared the same superordinate with the test lemmas – coded as 'Same Superordinate'.
  - o Example: For the tested lemma 'lieutenant', the translation 'จ่า' (cha, meaning 'a sergeant') showed that the participant had some idea that this was a rank in the army or the air force. Hence, it was coded as 'Same Superordinate'.

The answers fell into **Category C** (Zero points given) when:

- their meanings or senses were totally different from those of the test lemmas – coded as 'XMN'.
  - o Example: For the lemma 'philosophy', the translation 'ความรัก' (kwam-rak, meaning 'love') was incorrect and not related at all. It was coded as 'XMN'. No point was given to this translation.
- they were homophones of the test lemmas – coded as 'Homophones'.
  - o Example: For the lemma 'principle', the translation 'ผู้อำนวยการ' (phu-am-nuay-kan, meaning 'a school principal') was coded as 'Homophone'.
- they were synforms of the test lemmas – coded as 'Synform'.
  - o Example: For the tested lemma 'apparatus', the translation 'หน่อไม้ฝรั่ง' (no-mai-fa-rang, meaning 'asparagus') was wrong. The written forms were similar, but they were different lemmas. It was coded as 'Synform'.
- no answers were provided – no coding and no point was given.

When the participants provided more than one answer to each tested lemma in the receptive test, only the first two answers to each tested lemma were graded for practicality. These five scenarios were used when grading two receptive answers are shown in Table 6.1, and the flow chart showing how I scored the participants' answers in the receptive test is shown in Figure 6.1.

Scenario	The Accuracy of the Two Answers	Score
1	Two answers are correct.	2
2	One answer is correct, but the other is partially correct.	2
3	Two answers are partially correct.	1
4	One answer is partially correct, but the other is incorrect.	1
5	Two answers are incorrect.	0

Table 6.1: Five scenarios when grading two receptive answers
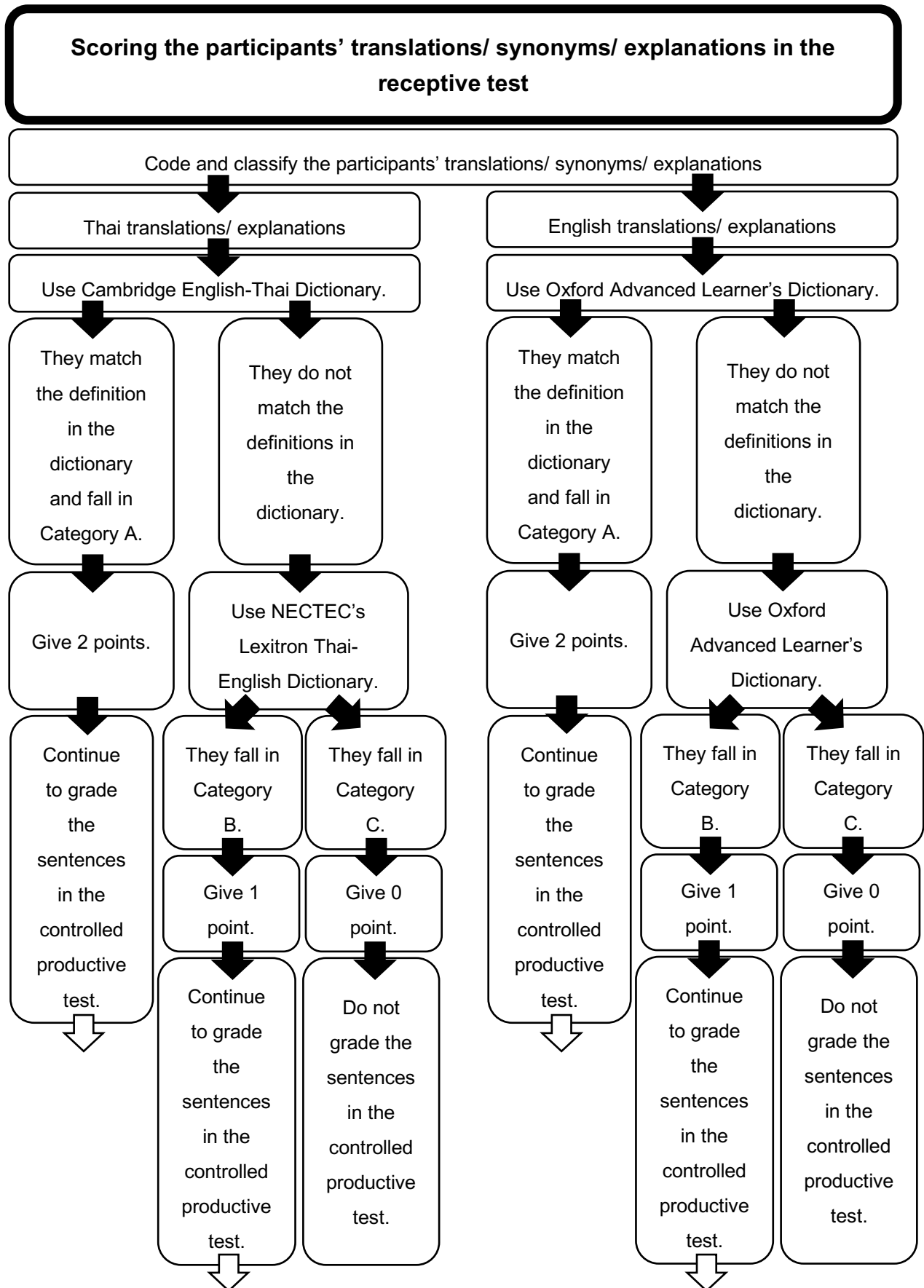
Figure 6.1: Flow chart of how to score the participants' answers in the receptive test

Examples of how I scored a participant's answers in the receptive test

These are some examples of how I scored the participants' answers in the receptive test. The answers of the first participant in Year 1 would be used to exemplify how I scored the answers in the receptive test. These five examples are Thai translations, English synonyms, and a Thai explanation of the given lemmas from the five frequency levels.

Example 1: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**sport**' (a lemma in the 2,000-word level)

The Thai translation: 'กีฬา' (ki-la, meaning 'sport')

I first checked the meaning of this lemma by using Cambridge English-Thai Dictionary. This lemma means 'กีฬา' (ki-la, meaning 'sport') in Thai and 'games or competitions involving physical activity' in English, according to the dictionary. Hence, this translation was coded as 'CMN' and given two points as it matched the translation in the dictionary.

Example 2: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**bench**' (a lemma in the 3,000-word level)

The Thai explanation: 'เก้าอี้นั่งในสวนสาธารณะ' (kao-i-nang-nai-suan-sa-tha-ra-na, meaning 'a seat in a public park')

I used Cambridge English-Thai Dictionary to check this translation. In the dictionary, this lemma means 'ม้านั่ง' (ma-nang, also meaning 'a bench') in Thai and 'a long (usually wooden) seat' as an English definition. This Thai explanation was coded as CMN and was given two points as it matched the translation in the dictionary.

Example 3: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**apparatus**' (a lemma in the 5,000-word level)

The English synonyms: 'tools, instrument'

The participant supplied two English synonyms, so I checked the accuracy of them both by looking them up in Oxford Advanced Learner's Dictionary. In the dictionary, this lemma means 'the tools or other pieces of equipment that are needed for a particular activity or task'. The words 'tools' and 'equipment' show that what the participant supplied were correct and I gave two points for these synonyms.

Example 4: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**antics**' (a lemma in the 10,000-word level)

The English synonym: 'ridiculous'

The participant supplied an English word 'ridiculous'. When I looked it up in Oxford Advanced Learner's Dictionary, I found that what the participant supplied only matched some part of the definition given in the dictionary. The definition is 'behaviour that is silly and funny in a way that people usually like'. Therefore, the word 'ridiculous' was partially correct and coded as 'RSense', and I gave one point to this answer.

Example 5: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**decade**' (a lemma in the academic word level)

The Thai translation: 'ศตวรรษ' (sa-ta-wat, meaning 'a century')

The Thai translation provided by Cambridge English-Thai Dictionary is 'ทศวรรษ' (tot-sa-wat, meaning 'ten years') and the English definition 'a period of ten years'. Therefore, this answer was incorrect and coded as 'XMN', and I gave one point to this answer.

These are five examples of how I scored the translations/ synonyms/ explanations supplied by the first participant in the receptive test. In the next section (Section 6.3.3), I explain how I analysed the controlled productive test and five sentences supplied by the first participant in Year 1 are also used as examples.

6.3.3 Analysing the controlled productive test

The second research question of the present study focuses on the participants'
controlled productive vocabulary knowledge in relation to different word frequency
levels and years of study. I analysed the participants' lemma use in the sentences
they provided in the controlled productive test after I graded their translations/
synonyms or explanations in the receptive test.

I checked whether the tested lemma was used with correct meaning and grammar
by referring to the Oxford Advanced Learner's Dictionary and Sketch Engine
(Kilgarriff et al., 2014). I would consult the two raters when the references could not
confirm the correctness of the sentences. The flow chart of how I graded the
answers in the controlled productive test is presented in Figure 6.2.

Scoring the participants' lemma use in sentences produced in the controlled productive test

Check the **meaning** of the tested lemma in the sentence. ⟹ Check the **grammar** of the tested lemma in the sentence.

Use the references: Oxford Advanced Learner's Dictionary.

Use the references: Oxford Advanced Learner's Dictionary and Sketch Engine.

The lemma is used with correct meaning.

The references cannot show whether the lemma is used with correct meaning.

The lemma is used with correct grammar.

The references cannot show whether the lemma is used with correct grammar.

Give 1 point.

Give 1 point.

Continue to check the grammar.

Consult two raters (native speakers).

Consult two raters (native speakers).

The lemma meaning is correct.

The lemma meaning is partially correct.

The lemma meaning is incorrect.

The grammar is correct.

The grammar is partially correct.

The grammar is incorrect.

Give 1 point.

Give 0 point.

Give 1 point.

Give 0 point.

Continue to check the grammar.

Do not check the grammar.

Figure 6.2: Flow chart of how to score the participants' answers in the controlled productive test

Examples of how I scored a participant's lemma uses in the controlled productive test

This section shows five examples of how I graded the accuracy of lemma uses in the controlled productive test.

Example 1: Checking the accuracy of the lemma '**sport**' in a sentence (a lemma in the 2,000-word level)

The Thai translation: 'กีฬา' (ki-la, meaning 'sport')

The sentence: "I think you'd better work out by playing some of your favourite **sports** because you are getting fatter."

I checked the meaning of this lemma in the sentence first. The words 'work out', 'play', 'fatter' showed that this lemma meaning fitted in this sentence. One point was given.

Then I checked its grammar by looking up the lemma in Oxford Advanced Learner's Dictionary. This lemma can be used in a plural form for American English and be used with the verb 'play'. I used the Sketch Engine to search English Web 2013 corpus in order to identify the modifiers of the lemma 'sport' and found that 'favourite' was one of them. Another point was given, so two points were given for the use of this lemma.

Example 2: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**bench**' (a lemma in the 3,000-word level)

The Thai explanation: 'เก้าอี้นั่งในสวนสาธารณะ' (kao-i-nang-nai-suan-sa-tha-ra-na, meaning 'a seat in a public park')

The sentence: "Sometimes, I just want to sit on a **bench** in a park watching the world moving because I have been through a really tough week and I just want to have a rest."

In terms of the accuracy of its meaning, the surrounding words such as 'sit on' and 'a park' ensured that the meaning fit in this context.

For grammar, Oxford Advanced Learners' Dictionary Online showed that the lemma 'bench' is a countable noun and the structure that is usually used with the lemma is 'sit (down) on'. Therefore, this lemma was used in the sentence correctly both in meaning and in grammar. Two points were given.

Example 3: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**apparatus**' (a lemma in the 5,000-word level)

The English synonyms: 'tools, instrument'

The sentence: "Firefighters need to use firefighting **apparatus** while working because it is risky and they need to have enough safety."

The surrounding words such as 'use' and 'while working' show that the lemma 'apparatus' was used with the correct meaning. Hence, one point was given.

For the grammar, I checked it in Oxford Advanced Learner's Dictionary. It shows that the lemma is an uncountable noun and also provides a sample sentence with similar context 'Firefighters needed breathing apparatus to enter the burning house'. However, when I used the Sketch Engine to search English Web 2013 corpus in order to identify whether 'use' can be used with 'apparatus', The Sketch Engine showed no results. Therefore, I then put this sentence in the scoring sheet for the two raters to grade it.

Example 4: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**antics**' (a lemma in the 10,000-word level)

The English synonym: 'ridiculous'

The sentence: "Sometimes, the novel needs to have some **antics** situations in order to make the readers giggle a bit."

The participant supplied a partially correct answer. Hence, I did not need to check the lemma use in the sentence as shown in Figure 6.2.

Example 5: Checking the accuracy of the translation/ synonym/ explanation of the lemma '**decade**' (a lemma in the academic word level)

The Thai translation: 'ศตวรรษ' (sa-ta-wat, meaning 'century)

The sentence: "Although **decades** have gone by, the Genocide done by Red Khmer is still in many Cambodian people's head because it was a tragedy and too hard to forget."

Again, the translation was incorrect. I then did not grade the sentence. The score for the lemma use was zero even though it looked grammatically correct.

6.3.4   Analysing the free productive test

In order to make the analysis of the free productive vocabulary knowledge consistent with that of the participants' receptive and controlled productive vocabulary knowledge, I followed these two processes. The first process was lemmatising the participants' free writings and classifying the lemmas into five groups of different frequency levels, and the second process was checking the accuracy (in meaning and grammar) of all the classified lemmas.

Process 1: Lemmatising all of the participants' free writings and classifying the lemmas into five groups of different frequency levels

The first main process I went through when analysing the participants' free writings was lemmatising all of the participants' free writings and classifying the lemmas into five groups of different frequency levels. This consists of the following four steps:

Step 1.   Lemmatising the free writings of all the participants (111 participants at a time) and removing all the function words or the mistakes generated by the lemmatising software, which I will explain later with some examples.

Step 2.   Classifying the lemmas produced by the 111 participants into five different frequency levels by comparing them against the words in the word lists used in the word sampling process in Schmitt et al.'s (2001) VLT

Step 3.   Lemmatising the free writing of the individual participants (one participant at a time)

Step 4.   Classifying the lemmas produced by the first participant into five groups of different frequency levels by comparing them against those from Step 2 and repeating this step for the rest of the participants

I explain steps 3 and 4 in more detail by using one of the participants' free writing as an example.

Step 1: Lemmatising all the participants' free writings and removing all the function words and mistakes

Step 1.1: Lemmatising all the participants' free writings

I present how I lemmatised all of the participants' free writings. First of all, I copied the free writings of all the participants and pasted them in one text file. I then uploaded this text file in a programme called Lemmatiser, which is available on Lextutor Website (Cobb, 2017), to lemmatise these running words produced by the 111 participants. Lemmatiser then presented a list of lemmas as the text file. I named this file 'All the participants' lemmas (with function words)'. I then removed function words from the list and named the list without function words 'All the participants' lemmas'. As there was too much data to present here, I use a sentence from the first participant (of the first year cohort)'s original text as an example of how I lemmatised and removed function words from it.

Example 1: This sentence represents the free writings of all the participants.

"Happiness can come to you in many different forms which can be from the feeling caused by loving someone or being loved by someone, or else it can be from being successful or getting what you want."

This sentence contained 26 lemmas.

be	cause	feeling	happiness	loving	successful	which
being	come	form	in	many	to	you
by	different	from	it	or	want	
can	else	getting	loved	someone	what	

The 111 pieces of free writing had 38,387 running words and 2,148 different lemmas.

Step 1.2: Removing all the function words

Next, I checked whether the lemmas in the list belonged to the function word class by looking them up in Oxford Advanced Learner's Dictionary Online . The word classes of function words together with some examples are shown in Table 6.2.

Word classes of function words	Examples
articles	a, an, the
auxiliaries	must, can, will, be, get, have
demonstratives	this, that, these, those
quantifiers	many, few, little, some
prepositions	on, of, with, to
pronouns (also possessive adjectives)	he, she, they (his, her, their)
conjunctions	and, but, or, so

Table 6.2: Function words removed from the lemma list

From the 26 lemmas in Example 1, I removed the following function words.

linking verb: be, being

preposition: by, from, in, to,

modal verb: can

pronoun: it, someone, what, which, you

determiner: many

conjunction: or

Categorizing words as either function or content presented some problems. According to the dictionary, the lemma 'else' is an adverb (content word), but together with the lemma 'or', it is a conjunction (function word). The lemma 'else' would not be removed from the list at this stage because it could appear in the other essays as an adverb. When checking the accuracy of using it in a sentence level for

the individual participants, I would remove it from the final lemma list if it was used as a conjunction.

Hence, this is the lemma list after the function words had been removed.

cause	different	feeling	getting	loved	successful
come	else	form	happiness	loving	want

Some of the lemmas are polysemous and belong to both the content word class and function word class. For example, the word 'concerning', which can be a lemma headword as a preposition (a function word), a lemma headword as an adjective (a content word), and a member of the lemma headword 'concern' as a verb (a content word). For these words, I checked whether that lemma was used as a function word or a content word in the individual participants' free writings. If it was used as a function word, I would remove it from the list. If not, I would keep it so that I could check the accuracy of using it in a sentence level further.

When I checked it (as shown in Figure 6.3), I found that 'concerning' was used as a content word as a member of the lemma headword 'concern' (a verb). As it was not a function word, I did not remove it from the lemma list.

why their parents do this might involve with 'money' because being able to attend famous and high competition university guarantees to have more chance to work in a great company. There is no parents who want their child to be poor, they all want them to live happily and be able to eat or do what they want without concerning about money. Moreover, they all know that money can buy happiness. Whereas some people might say that "my happiness is to stay at my home, so I don't

Figure 6.3: Sentence with the lemma member 'concerning'

Step 1.3: Removing mistakes generated by the lemmatiser

After removing function words, I removed some lemmas that did not appear in the participants' free writings but were lemmatised by the software. This was because of the limitation – the software was unable to distinguish word classes. For instance, the lemma 'bite' (a verb) was in the generated lemma list. However, when I searched for this lemma in the participants' free writings, the lemma 'bite' did not appear in the free writings, but the lemma 'bit' (noun) as in 'a bit' did instead. In this case, I then removed the lemma 'bite' from the list since it was not the lemma written by the participants.

These are the limitations of the lemmatising software. The software cannot distinguish the word classes or whether a word is a lemma headword or a lemma member, so I needed to check the lemma list generated by the software again carefully.

Once I had removed all the function words and the lemmas that were not written by the participants from the list, I then checked the frequency of the rest of the lemmas and classified them into five groups of different frequency levels as presented in Step 2.

Step 2: Checking the frequency of the lemmas and classifying the lemmas produced by all the participants into five groups of different frequency levels

After I removed all the function words from the list of the lemmas produced by all the participants and removed the mistakes generated by the lemmatising software, I the frequency of the lemmas and classified these lemmas into five groups of different frequency levels (the 2,000-, the 3,000-, the 5,000-, the 10,000-, and the academic word levels).

However, the target words in the receptive test and the controlled productive test used in this study were sampled from the target words used in Schmitt et al.'s (2001) VLT. Therefore, to make the analysis of the free productive test consistent with that of the first two tests, the frequency lists used in the processes of word sampling in the VLT (shown in Table 6.3) were then also used in this part of the study. These frequency lists were Thorndike and Lorge's (1944) list, West's (1953) GSL and Coxhead's (2000) AWL. Also, I made a different interpretation of the GSL and the AWL. That is, I treated the word families in the GSL and in the AWL as lemmas. The rationale for this is presented in Section 3.3.

Frequency levels in the VLT	Thorndike and Lorge's (1944) counts	West' (1953) GSL	Coxhead's (2000) AWL
2,000 words	With 'A' occurrence	The first and the second 1,000 words of the GSL	-
3,000 words	With '30-49' occurrences per million running words	-	-

5,000 words	With '14-18' occurrences per million running words	-	-
10,000 words	With '3' occurrences per million running words	-	-
academic words	-	-	AWL

Note: 'A' occurrence refers to 'at least 50 occurrences and not so many as 100 occurrences per million running words'

Table 6.3: Frequency lists used in word sampling process in the VLT

Step 2.1: Checking the frequency in Thorndike and Lorge's (1944) list

I started with checking the occurrence of each of the lemmas in Thorndike and Lorge's (1944) list and typed in its occurrence per million running words in Thorndike and Lorge's (1944) counts in the spreadsheet as shown in Table 6.4.

All the participants' lemmas (excluding function words and mistakes)	Lemmas as shown in Thorndike and Lorge's (1944) List	Occurrence per million running words in Thorndike and Lorge's (1944) counts
abandoned	abandoned (adj.)	3
ability	ability	49
able	able	AA*
abroad	abroad	48
absolutely	absolutely	39
abstract	abstract	6
accept	accept	AA*
acceptance	acceptance	11
accepted	-	-

Note: 'AA occurrence' refers to '100 occurrences or over per million running words

Table 6.4: Occurrence per million in Thorndike and Lorge's (1944) counts for each of the participants' lemmas

There are two limitations of Thorndike and Lorge's (1944) List. First, it is not consistent in including lemmas in the list. According to Oxford Advanced Learner's Dictionary Online, both 'abandoned' and 'accepted' are lemma headwords as adjectives. Thorndike and Lorge's (1944) List includes 'abandoned' and also identifies its word class, yet it does not include 'accepted'. I then treated 'accepted' as a member of the lemma headword 'accept'.

Second, Thorndike and Lorge's (1944) List does not identify word classes for its lemmas. For example, the lemma 'abstract' can be an adjective, a noun, and a verb, but the list showed only the lemma and its occurrence pre million running words in Thorndike and Lorge's (1944) counts. Hence, in this case, I assumed that the occurrence of such lemma (shown in Thorndike and Lorge's (1944) counts) applies to all of its parts of speech.

Step 2.2: Creating five lemma lists

Step 2.2.1: Creating 'All Participants' A&1K2KGSL Lemma List'

When I finished checking the frequency levels of the lemmas in Thorndike and Lorge's (1944) List, I sorted the lemmas and saved only those with 'A' occurrence in a text file so that it could be uploaded and be compared on Lextutor website (Cobb, n.d.). Then I compared the lemmas with 'A' occurrence in Thorndike and Lorge's (1944) list with the GSL by using Text Lex Compare software on Lextutor website (Cobb, n.d.).

This software only offers choices of units for counting words: 'types', 'word families' or 'phrases' in this software. I chose the option 'types' because the list of the 'A' occurrence was already a lemma list and the option 'type' could provide the output in the same form as the input. Therefore, the option 'type' was suitable for this process.

I noticed that, using the option 'type', the hyphenated word and the prepositional phrase in the GSL ('in-law' and 'out of') would be separated as 'in', 'law', 'of' and 'out'. However, this did not affect the data analysis in my study.

Some examples of the lemmas compared between the GSL and the participants' lemmas with 'A' occurrence in Thorndike and Lorge's (1944) List are shown in Table 6.5. The three dots in the table means there are more lemmas than the ones shown here.

Lemmas unique to the GSL	Shared lemmas	Lemmas unique to the list of the participants' lemmas with 'A' occurrence in Thorndike and Lorge's (1944) List
001. in 2	001. accident 2	001. accomplish 1
002. law 2	002. actual 2	002. active 1
003. of 2	003. advantage 2	003. activity 1
004. out 2	004. afford 2	004. actually 1
005. a 1	005. aim 2	005. addition 1
006. able 1	006. alive 2	006. affect 1
007. about 1	007. ancient 2	007. alarm 1
008. above 1	008. apply 2	008. angry 1
009. abroad 1	009. aside 2	009. anyway 1
010. absence 1	010. attend 2	010. appeal 1
.	.	.
.	.	.
.	.	.

Table 6.5: Comparing the GSL to the List of the participants' lemmas with 'A' occurrence in Thorndike and Lorge's (1944) List

I extracted the lemma list shared between the two lists – the middle column in Table 6.5 – and saved it as a text file named 'All Participants' A&1K2KGSL Lemma List'.

Step 2.2.2: Creating 'All Participants' AWL Lemma List'

The next lemma list I created was the list of the lemmas produced by all the participants and also appeared in Coxhead's (Coxhead, 2000) AWL. According to Table 6.3, the lemmas in the AWL themselves did not appear in the GSL, so I did not need to exclude any of the lemmas in the file 'All Participants' A&1K2KGSL Lemma List' from the lemmas produced by all the participants before checking them against the AWL.

To do this, I used the Text Lex Compare software again. Some examples of the lemmas compared between the two lists are shown in Table 6.6.

Lemmas unique to the AWL	Shared lemmas	Lemmas unique to 'All the participants' lemmas'
001. abandon 1	001. abstract 2	001. abandoned 1
002. academy 1	002. access 2	002. ability 1
003. accommodate 1	003. achieve 2	003. able 1
004. accompany 1	004. adapt 2	004. aboard 1
005. accumulate 1	005. adequate 2	005. abroad 1
006. accurate 1	006. adult 2	006. absolutely 1
007. acknowledge 1	007. affect 2	007. accept 1
008. acquire 1	008. apparent 2	008. acceptance 1
009. adjacent 1	009. appreciate 2	009. accessory 1
010. adjust 1	010. appropriate 2	010. accident 1
.	.	.
.	.	.
.	.	.

Table 6.6: Comparing the AWL to 'All the participants' lemmas'

As I treated all the words as lemmas, the word 'abandon' was treated as a different lemma from the word 'abandoned' because 'abandoned' can be an adjective. But if the word 'abandoned' was used by a participant as a past tense verb, I treated it the same way as the lemma 'abandon' (verb). Hence, when I compared them using the option 'type', they both did not appear in the shared column.

I extracted the lemmas shared between the two lists and saved them in a text file named 'All Participants' AWL Lemma List'.

Step 2.2.3: Creating 'All Participants' 3K Lemma List'

The third lemma list I created was named as 'All Participants' 3K Lemma List'. As I used the test lemmas from the VLT in the receptive test and the controlled productive test in my study, I then needed to analyse only the lemmas in the free productive test that also had the same frequency as those in the VLT.

To ensure consistency with the 3,000-word level in the VLT (see Table 6.3), I needed to sort the lemmas produced by all the participants and choose only the lemmas with 30-49 occurrences per million running words in Thorndike and Lorge's (1944) counts. I saved them in a text file named 'All Participants' 30-49 Lemma List'. The lemmas in this list themselves did not appear in the GSL as they had different occurrences (30-49 occurrences per million) from those in the GSL ('A' occurrence – 50-99 occurrences per million in Thorndike and Lorge's (1944) counts). I, therefore,

did not need to exclude any lemmas appearing in the GSL from 'All Participants' 30-49 Lemma List'.

However, some lemmas in this list appeared in the AWL. Accordingly, I needed to exclude those appearing in the AWL. I compared 'All Participants' 30-49 Lemma List' with the list named 'All Participants' AWL Lemma List' (shown in Table 6.7).

Lemmas unique to 'All Participants' AWL Lemma List'	Shared lemmas	Lemmas unique to 'All Participants' 30-49 Lemma List'
001. abandon 1	001. attain 2	001. ability 1
002. abstract 1	002. capable 2	002. abroad 1
003. access 1	003. commit 2	003. absolutely 1
004. achieve 1	004. conflict 2	004. advise 1
005. adapt 1	005. construct 2	005. affection 1
006. adequate 1	006. contact 2	006. agreement 1
007. adult 1	007. contrary 2	007. agriculture 1
008. affect 1	008. contrast 2	008. anxious 1
009. apparent 1	009. convince 2	009. application 1
010. appreciate 1	010. definite 2	010. argue 1
.	.	.
.	.	.
.	.	.

Table 6.7: Comparing 'All Participants' AWL Lemma List' to 'All Participants' 30-49 Lemma List'

I extracted only the lemmas unique to 'All Participants' 30-49 Lemma List' on the right column and saved them in a text file named 'All Participants' 3K Lemma List'.

Step 2.2.4: Creating 'All Participants' 5K Lemma List'

As shown in Table 6.3, the lemmas for the 5,000-word level have 14-18 occurrences per million running words in Thorndike and Lorge's (1944) counts and do not appear in the GSL or the AWL.

The process of creating this list was similar to that of creating 'All Participants' 3K Lemma List. I started by sorting the lemmas produced by the participants (function words and mistakes removed) and saved only the lemmas with 14-18 occurrences in a text file as 'All Participants' 14-18 Lemma List'.

Again, these lemmas did not appear in the GSL, but some appeared in the AWL. Hence, I needed to exclude those appearing in the AWL. I compared 'All Participants' 14-18 Lemma List' to 'All Participants' AWL Lemma List' (shown in Table 6.8).

Lemmas unique to 'All Participants' AWL Lemma List'	Shared lemmas	Lemmas unique to 'All Participants' 14-18 Lemma List'
001. abandon 1	001. adequate 2	001. airplane 1
002. abstract 1	002. adult 2	002. alcohol 1
003. access 1	003. appropriate 2	003. attraction 1
004. achieve 1	004. collapse 2	004. barely 1
005. adapt 1	005. contemporary 2	005. cherish 1
006. affect 1	006. furthermore 2	006. concrete 1
007. apparent 1	007. guarantee 2	007. confident 1
008. appreciate 1	008. index 2	008. confront 1
009. area 1	009. internal 2	009. consequently 1
010. aspect 1	010. interpret 2	010. convenience 1
.	.	.
.	.	.
.	.	.

Table 6.8: Comparing 'All Participants' AWL Lemma List' to 'All Participants' 14-18 Lemma List'

I extracted only the lemmas unique to 'All Participants' 14-18 Lemma List' and saved it as 'All Participants' 5K Lemma List'.

Step 2.2.5: Creating 'All Participants' 10K Lemma List'

Creating 'All Participants' 10K Lemma List' was similar to creating the two previous lemma lists. According to Table 6.3, the lemmas at the 10,000-word level should have 3 occurrences per million running words in Thorndike and Lorge's (1944) counts and should not appear in the GSL or the AWL.

I sorted the lemmas written by the participants (function words and mistakes removed) and saved only the lemmas with 3 occurrences in a text file named 'All Participants' 3 Lemma List'.

I then compared 'All Participants' 3 Lemma List' to 'All Participants' AWL Lemma List' as shown in Table 6.9.

Lemmas unique to 'All Participants' AWL Lemma List'	Shared lemmas	Lemmas unique to 'All Participants' 3 Lemma List'
001. abandon 1	001. concept 2	001. abandoned 1
002. abstract 1		002. animation 1
003. access 1		003. bonfire 1
004. achieve 1		004. capitalism 1
005. adapt 1		005. carefree 1
006. adequate 1		006. coldness 1
007. adult 1		007. confidentially 1
008. affect 1		008. deliriously 1
009. apparent 1		009. disagreement 1
010. appreciate 1		010. drawback 1
.		.
.		.
.		.

Table 6.9: Comparing 'All Participants' AWL Lemma List' to 'All Participants' 3 Lemma List'

I extracted only the lemmas unique to 'All Participants' 3 Lemma List' and saved them in a text file named 'All Participants' 10K Lemma List.

Up to this step, I had created five lists of the lemmas produced by the 111 participants:

1. All Participants' A&1K2KGSL Lemma List
2. All Participants' AWL Lemma List
3. All Participants' 3K Lemma List
4. All Participants' 5K Lemma List
5. All Participants' 10K Lemma List

All these lists would be used again in Step 4 after I lemmatised the individual participants' free writings in Step 3.

Step 3: Lemmatising each of the participants' free writing

In this section, I lemmatised the individual participants' free writings and saved them as separated text files so that I could classify the lemmas into five groups of different frequency levels and check the accuracy of using them in a sentence level later.

I used the free writing of the first participant of the first-year cohort as an example. It is shown in Figure 6.4.

Topic (หัวข้อ): *"A person cannot be poor and happy, because money is always needed to gain something that is important to that person."*

Instruction (คำสั่ง): Write an essay of 300 words in English on the topic provided. Argue for or against this idea. (เขียนเรียงความเป็นภาษาอังกฤษ **300** คำ เพื่อสนับสนุนหรือคัดค้านกับหัวข้อที่กำหนดให้)

Happiness, as everyone may know, is subjective. It cannot be touched or held because it is something you feel from your heart. Happiness can come to you in many different forms which can be from the feeling caused by loving someone or being loved by someone, or else it can be from being successful or getting what you want. However, what is important is that everyone has right to be happy.

All humans are born free with liberty and dignity which means we deserve to do anything or be anything which makes us happy. However, nowadays money has become one of the most important factors that everyone is required to have in order to survive in this such a world and society. For example, those who live in such an agitated metropolis as Bangkok has to pay some money in order to get fresh air into their lungs as well as those who live in cities well far away from the sea also needs to make a payment in order to see beaches and clear water. On the contrary, the revenue of spending some money maybe something you want and you may consider this as your happiness, but as time goes by, this kind of happiness will fade away one day.

Au contraire, happiness should come from your mind of understanding the others and how you look at the world through your own perspective. Happiness is not always what you receive from the others but it can be something produced in your own mind by yourself. I strongly believe that if you can look at the world through positive lens of your eyes and mind, without ego, ignorance or prejudice, happiness will be lit and it will be like an eternal flame shining in your mind no matter how many decades pass by. By this way, richness or poverty would not be an obstacle to be happy.

Figure 6.4: An example of the first participant's free writing

I lemmatised his free writing by using the Lemmatiser software on Lextutor website (Cobb, n.d.). I also chose the choices of 'Lemmas' and 'Headwords'. This process was similar to when I lemmatised all the 111 participants' free writings explained earlier in Step 1.

I extracted the list of the lemmas produced by the first participant and saved them in a text file named 'Year1Student1 Lemma List' (Table 6.10).

a agitated air all also always and anything as at au away bangkok be beach because become being believe born but by can cause city clear come consider contraire contrary day decade deserve different dignity do ego else eternal everyone example eye factor fade far feel feeling flame for form free fresh from get getting go ha happiness happy have heart held how however human i if ignorance important in into is it kind know lens liberty like lit live look loved loving lung make many matter may maybe means metropolis mind money most need no not nowadays obstacle of on one or order other own pass pay payment perspective positive poverty prejudice produced receive required revenue richness right sea see shining should society some someone something spending strongly subjective successful such survive that their this those through time to touched understanding us want water way we well what which who will with without world would you your yourself

Table 6.10: All the lemmas in 'Year1Student1 Lemma List'

In this step, it was not necessary to remove the function words or the mistakes because this list – 'Year 1Student1 Lemma List' – would be compared to the five lemma lists that I created in Step 2 and only shared lemmas would be selected and used.

Then I went through this process with each of the participant' free writings one by one. There were 111 participants in my study (34 participants in Year 1, 24 participants in Year 2, 29 participants in Year 3, and 24 participants in Year 4), so there were 111 lemma lists at the end of this process.

Step 4: Classifying the lemmas produced by each of the participants into five groups of different frequency levels

I classified the lemmas written by each of the participants by comparing his/her lemma list to the five lemma lists I created in Step 2. I also used the free writing of the first student in Year 1 as an example.

Step 4.1: Classifying the first participant's lemmas into the 2,000-word level

I classified the lemmas produced by the first student in Year 1 ('Year1Student1 Lemma List' as an example) into the 2,000-word level by comparing his lemma list to

'All Participants' A&1K2KGSL Lemma List'. I also used the Text Lex Compare software on Lextutor website. The lemmas that were shared between the two lists are in the middle column of Table 6.11.

Lemmas unique to 'All Participants' A&1K2KGSL Lemma List'	Shared lemmas	Lemmas unique to 'Year1Student1 Lemma List'
001. accident 1	001. flame 2	001. a 1
002. actual 1	002. liberty 2	002. agitated 1
003. advantage 1		003. air 1
004. afford 1		004. all 1
005. aim 1		005. also 1
006. alive 1		006. always 1
007. ancient 1		007. and 1
008. apply 1		008. anything 1
009. aside 1		009. as 1
010. attend 1		010. at 1
.		.
.		.
.		.

Table 6.11: Comparing 'Year1Student1 Lemma List' to 'All Participants' A&1K2KGSL Lemma List'

Two lemmas 'flame' and 'liberty' were shared between the two lists. I then put these two lemmas in a spreadsheet before I graded the accuracy of using them in a sentence level (Table 6.12).

Year 1 Student 1 Lemmas									
2K	score	3K	score	5K	score	10K	score	Academic	score
flame liberty									

Table 6.12: Spreadsheet showing lemmas at the 2,000 frequency level produced by the first student of the first-year cohort

Step 4.2: Classifying the lemmas into the 3,000-word level

I classified the lemmas in 'Year1Student1 Lemma List' into the 3,000-word level by comparing 'Year1Student1 Lemma List' against 'All Participants' 3K Lemma List' as shown in Table 6.13.

Lemmas unique to 'All Participants' 3K Lemma List'	Shared lemmas	Lemmas unique to 'Year1Student1 Lemma List'
001. ability 1	001. deserve 2	001. a 1
002. abroad 1	002. dignity 2	002. agitated 1
003. absolutely 1	003. eternal 2	003. air 1
004. advise 1	004. fade 2	004. all 1
005. affection 1	005. payment 2	005. also 1
006. agreement 1	006. strongly 2	006. always 1
007. agriculture 1		007. and 1
008. anxious 1		008. anything 1
009. application 1		009. as 1
010. argue 1		010. at 1
.		.
.		.
.		.

Table 6.13: Comparing 'Year1Student1 Lemma List' to 'All Participants' 3K Lemma List'

I saved only those six lemmas shared between the two lists as the lemmas in the 3,000-word level produced by the first student of the first-year cohort, and I put them in the spreadsheet for the process of grading the usage (Table 6.14).

Year 1 Student 1 Lemmas									
2K	score	3K	score	5K	score	10K	score	Academic	score
flame liberty		deserve dignity eternal fade payment strongly							

Table 6.14: Spreadsheet showing lemmas at the 3,000frequency level produced by the first student of the first-year cohort

Step 4.3: Classifying the lemmas into the 5,000-word level

For the 5,000-word level, I went through the same process with the 2,000- and the 3,000-word levels. I compared 'Year1Student1 Lemma List' against 'All Participants' 5K Lemma List' by using Text Lex Compare software.

There were five lemmas shared between the two lists as shown in the middle column of Table 6.15.

Lemmas unique to 'All Participants' 5K Lemma List'	Shared lemmas	Lemmas unique to 'Year1Student1 Lemma List'
001. airplane 1	001. lit 2	001. a 1
002. alcohol 1	002. lung 2	002. agitated 1
003. attraction 1	003. nowadays 2	003. air 1
004. barely 1	004. obstacle 2	004. all 1
005. cherish 1	005. prejudice 2	005. also 1
006. concrete 1		006. always 1
007. confident 1		007. and 1
008. confront 1		008. anything 1
009. consequently 1		009. as 1
010. convenience 1		010. at 1
.		.
.		.
.		.

Table 6.15: Comparing 'Year1Student1 Lemma List' to 'All Participants' 5K Lemma List'

I then put these shared lemmas in the 5K column of the grading spreadsheet (Table 6.16).

Year 1 Student 1 Lemmas									
2K	score	3K	score	5K	score	10K	score	Academic	score
flame		deserve		lit					
liberty		dignity		lung					
		eternal		nowadays					
		fade		obstacle					
		payment		prejudice					
		strongly							

Table 6.16: Spreadsheet showing lemmas at the 5,000frequency level produced by the first student of the first-year cohort

Step 4.5: Classifying the lemmas into the 10,000-word level

Also, I compared 'Year1Student1 Lemma List' against 'All Participants' 10K Lemma List' to classify the lemmas written by this student into the 10,000-word level.

Lemmas unique to 'All Participants' 10K Lemma List'	Shared lemmas	Lemmas unique to 'Year1Student1 Lemma List'
001. abandoned 1		001. a 1
002. animation 1		002. agitated 1
003. bonfire 1		003. air 1
004. capitalism 1		004. all 1
005. carefree 1		005. also 1
006. coldness 1		006. always 1
007. confidentially 1		007. and 1
008. deliriously 1		008. anything 1
009. disagreement 1		009. as 1
010. drawback 1		010. at 1
.		.
.		.
.		.

Table 6.17: Comparing 'Year1Student1 Lemma List' to 'All Participants' 10K Lemma List'

However, no lemmas were shared between the two lists. Consequently, I left the column of the 10,000-word level in the grading spreadsheet blank (Table 6.18).

Year 1 Student 1 Lemmas									
2K	score	3K	score	5K	score	10K	score	Academic	score
flame		deserve		lit					
liberty		dignity		lung					
		eternal		nowadays					
		fade		obstacle					
		payment		prejudice					
		strongly							

Table 6.18: Spreadsheet showing lemmas at the 10,000-word level produced by the first student of the first-year cohort

Step 4.6: Classifying the lemmas into the academic word level

I classified the lemmas produced by this student by comparing 'Year1Student1 Lemma List' against 'All Participants' AWL Lemma List' (Table 6.19).

Lemmas unique to 'All Participants' AWL Lemma List'	Shared lemmas	Lemmas unique to 'Year1Student1 Lemma List'
001. abandon 1	001. contrary 2	001. a 1
002. abstract 1	002. decade 2	002. agitated 1
003. access 1	003. factor 2	003. air 1
004. achieve 1	004. ignorance 2	004. all 1
005. adapt 1	005. perspective 2	005. also 1
006. adequate 1	006. positive 2	006. always 1
007. adult 1	007. revenue 2	007. and 1
008. affect 1	008. survive 2	008. anything 1
009. apparent 1		009. as 1
010. appreciate 1		010. at 1
.		.
.		.
.		.

Table 6.19: Comparing 'Year1Student1 Lemma List' to 'All Participants' AWL Lemma List'

There were eight lemmas shared between 'Year1Student1 Lemma List' and 'All Participants' AWL Lemma List' as shown in the middle column of Table 6.19. I then put these lemmas in the grading spreadsheet shown in Table 6.20.

Year 1 Student 1 Lemmas										
2K	score	3K	score	5K	score	10K	score	Academic	score
flame		deserve		lit				contrary	
liberty		dignity		lung				decade	
		eternal		nowadays				factor	
		fade		obstacle				ignorance	
		payment		prejudice				perspective	
		strongly						positive	
								revenue	
								survive	

Table 6.20: Spreadsheet showing lemmas at the academic level produced by the first student of the first-year cohort

After I finished classifying the lemmas produced by the first participant into the five groups of different frequency levels, I went through the same processes for the rest of the participants. That is, I compared their lemma lists against those five lemma lists from Step 2. The shared lemmas were put into the grading spreadsheet for a further analysis.

I created a diagram of how to analyse the participants' free writings for a bigger picture of all these processes.

Process 1 (Steps 1 and 2): Lemmatising all the participants' free writings and classifying the lemmas into five frequency levels

Lemmatise all the participants' free writings at once using Lemmatiser on Lextutor website.

Remove function words and mistakes from the lemma list (by using Oxford Advanced Learner's Dictionary and checking word classes in the participants' free writings).

Check the frequency of each lemma in Thorndike and Lorge's (1944) list.

Classify the lemmas into five frequency levels.

Sort the lemmas falling in 'A' categories.	Compare with the GSL, Extract the shared lemmas.	Create 'All Participants' A&1K2KGSL Lemma List'.
Use all the lemmas.	Compare with the AWL, Extract the shared lemmas.	Create 'All Participants' AWL Lemma List'.
Sort the lemmas with 30-49 occurrences*.	Compare with the AWL, Exclude lemmas of AWL.	Create 'All Participants' 3K Lemma List'.
Sort the lemmas with 14-18 occurrences*.	Compare with the AWL, Exclude lemmas of AWL.	Create 'All Participants' 5K Lemma List'.
Sort the lemmas with 3 occurrences*.	Compare with the AWL, Exclude lemmas of AWL.	Create 'All Participants' 10K Lemma List'.

Note: 'Occurrence' refers to occurrences per million running words in Thorndike and Lorge's (1944) counts.

Figure 6.5: Flow chart of how to lemmatise all the participants' free writings and classify the lemmas into five frequency levels

Process 1 (Steps 3 and 4): Lemmatising the individual participants' free writings and classifying the lemmas into five frequency levels

Lemmatise one participant's free writing at a time using Lemmatiser on Lextutor website.

Classify the lemmas written by each participant by comparing them against the five lemma lists from Figure 6.5.

Compare each participant's lemma list with 'All Participants' A&1K2KGSL Lemma List'.

Classify only the shared lemmas as the lemmas in the 2,000-word level.

Compare each participant's lemma list with 'All Participants' 3K Lemma List'.

Classify only the shared lemmas as the lemmas in the 3,000-word level.

Compare each participant's lemma list with 'All Participants' 5K Lemma List'.

Classify only the shared lemmas as the lemmas in the 5,000-word level.

Compare each participant's lemma list with 'All Participants' 10K Lemma List'.

Classify only the shared lemmas as the lemmas in the 10,000-word level.

Compare each participant's lemma list with 'All Participants' AWL Lemma List'.

Classify only the shared lemmas as the lemmas in the academic word

Figure 6.6: Flow chart of how to lemmatise each of the participant' s free writing and classify the lemmas into five frequency levels

These are the flow charts showing the processes of lemmatising the participants' free writings and classifying their lemmas into five frequency levels.

Limitations of analysing tools

- **Limitations of Lemmatiser on Lextutor Website**

In the step of lemmatising the participants' free writings, I found some limitations of the lemmatising software on Lextutor website (Cobb, n.d.).

1) The lemmatiser did not treat a hyphenated word as a single lexical unit. It separated the word 'so-called' in the AWL and provided two lemmas 'so' and 'call'. I found this when I compared the GSL lemma list and the AWL lemma list in order to ensure the accuracy of the two lemma lists. Then I found two lemmas 'so' and 'call' in the shared column between the two lists. This is similar to when I lemmatised the words in the GSL. Therefore, I needed to remove the lemma 'call' from 'All Participants' AWL List', and remove 'so' (conjunction) for both lists.

2) The lemmatiser did not treat a prepositional phrase 'out of' in the GSL as one lemma. However, this did not affect the data analysis in my study because I focused on only content words, not function words.

3) The lemmatiser did not recognise the meaning of the lemma that was used in a wrong tense form.
   For example, in a sentence produced by one of the first-year participants, "A person who love to share and help, they will not only found the happiness inside their heart but they will also gain it from the others smile," the lemmatiser provided the lemma 'found' (in the academic word level) instead of 'find' (not in any of the five frequency levels). Four of the participants wrote 'found' in their essays, so I needed to check whether they meant to use the lemma headword 'found' or 'find'. If they meant to use the lemma 'find' as in the example sentence, I would remove the lemma 'found' from their academic word level.

4) The lemmatiser cannot lemmatise misspelled lemmas. However, these lemmas also showed partial knowledge the participants. Accordingly, I needed to check the misspelled lemmas in each of the participants' free writings manually whether they should be categorised into any of the five groups of the frequency levels. If they were supposed to be in any of the frequency levels, I would include those misspelled lemmas and check for their accuracy when being used in a sentence level.
   For instance, one of the participants produced the misspelled lemma 'now a day' in his free writing. I then checked the frequency of the lemma 'nowadays' and found that 'nowadays' fell into the 5,000-word level. Therefore, I included 'now a day' in the grading spreadsheet as one misspelled lemma in this level for this participant.

- **Limitations of Thorndike and Lorge's (1944) List**

When I checked the participants' lemmas against Thorndike and Lorge's (1944) List in order to fill in the frequency of the lemmas, I found some limitations of this list.

1) I found that Thorndike and Lorge's (1944) List is relatively old. It was published in 1944, yet it was used in this research because of the consistency. Consequently, it does not include many of the contemporary lemmas. Some examples of these lemmas are 'computer', 'exam', 'robot', 'software', 'upgrade', 'delete', and 'video'.

2) The second limitation is that Thorndike and Lorge's (1944) List does not include many types of lemmas. It does not cover British English spellings such as 'mum'. Hence, I checked the frequency of the lemmas in American English spellings instead.
   Moreover, the list does not cover the short forms of some words. These short forms are also lemmas such as 'math', and 'exam', or some abbreviations such as 'VIP'. Thorndike and Lorge's (1944) List has the lemmas 'mathematics', and 'examination', though. However, when checking for their frequency, I did not treat these short forms as their full forms.

3) The third limitation is Thorndike and Lorge's (1944) List does not identify the word classes (parts of speech) of each lemma. This would be problematic particularly for the lemmas that can belong to many word classes such as 'due', 'sound', 'drink', 'order', 'even', 'evidence', 'coin', 'patient', and 'influence'. I then assumed that the frequency as indicated in Thorndike and Lorge's (1944) List for each lemma would apply to all of their word classes.
   In addition, even though Thorndike and Lorge's (1944) List identifies the word classes of the lemmas that look like lemma members such as 'abandoned (adj.)', hurried (adj.)', retired (adj.)', 'retiring (adj.)', and 'limited (adj.)', some of these lemmas such as 'accepted (adj.)' are not included in the list. It has only the base form 'accept'.

4) Thorndike and Lorge's (1944) List does not identify the meanings of a polysemous word. For example, the lemma 'fan' has three meanings, according to Oxford Advanced Learner's Dictionary. It is identified as having 38 occurrence pre million running words in Thorndike and Lorge's (1944) counts, which fell in the 3,000-word level. However, the list does not show which of the three meanings has this occurrence. I then assumed that all the meanings of the lemma 'fan' have the same occurrence.

After the process of lemmatisation, I graded the accuracy and appropriateness of the lemma use in terms of meaning and grammar. It is presented in Process 2.

Process 2: Checking the accuracy (in meaning and grammar) of all the classified lemmas

This section shows the process of checking the accuracy of the lemmas produced and used by the participants in order to explore the participants' free productive vocabulary knowledge. The analysing process of the participants' free productive vocabulary knowledge was designed to be consistent to that of the participants' controlled productive vocabulary knowledge. That is, each lemma was checked by employing a three-scale scoring system:

- two points for the lemma which was used correctly in meaning and in grammar
- one point for the lemma which was used correctly in meaning only
- zero point for the lemma which was used incorrectly in meaning (no need to check its grammatical usage)

The two references used to check its accuracy were Oxford Advanced Learner's Dictionary Online and the text analysing software – the Sketch Engine (Kilgarriff et al., 2014). The dictionary was used for word classes, meanings, and some grammatical functions. The function of Word Sketch and the concordance in the Sketch Engine were used to check grammatical structures and collocational behaviours. The English Web 2013 (enTenTen13) corpus was used for this purpose because this is a large English corpus of almost 20 billion running words as justified in Section 4.6.3. It, therefore, covers sufficient grammatical structures produced by the participants in this study.

When these two references were unable to provide evidence that a particular lemma was used correctly, the sentence with that lemma was then graded by the two raters (English native speakers). Similar to when the two raters graded the data of the controlled productive vocabulary knowledge, they were instructed how to grade the data, they graded the first 20 sentences together, and the scores given by them both were then analysed using statistics for the reliability of their grading. Then the sentences receiving different scores would be revised by the raters for the agreement.

Here is the flow chart showing how to score the participants' lemma use in the free productive test in this study (Figure 6.7).

```
┌─────────────────────────────────────────────────────────────────┐
│  Process 2: Scoring the participants' lemma use in the free       │
│                   productive vocabulary test                       │
└─────────────────────────────────────────────────────────────────┘
```

| Check the **meaning** of the lemma by using Oxford Advanced Learner's Dictionary Online. | Check the **grammar** of the lemma by using Oxford Advanced Learner's Dictionary Online and the software called Sketch Engine. |

| The lemma is used with its correct meaning. | The lemma is used with vague meaning. | The lemma is used with incorrect meaning. | Its grammatical functions match those of the references. | Its grammatical functions did not match those of the references. |

| Give 1 point. | Consult the two raters. | Give 0 point. | Give 1 more point. | Consult the two raters. |

| The meaning is acceptable. | The meaning is not acceptable. | | The lemma is grammatically correct. | The lemma is not grammatically correct. |

| Give 1 point. | Give 0 point. | | Give 1 more point. | Give 0 point. |

| Check its grammatical functions. | Do not check its grammatical functions. |

Note: Meaning score + Grammar score = Free productive knowledge score

Figure 6.7: Flow chart of how to score the participants' answers in the free productive test

The process of analysing the free productive vocabulary is similar to that of analysing the controlled productive vocabulary. However, one main difference of the data on the controlled productive vocabulary knowledge and that on the free productive vocabulary knowledge in this study was the estimation of the number of unknown lemmas in each frequency level.

For example, if a participant were able to use five out of ten lemmas in the 2,000-word level correctly both in meaning and in grammar in the controlled productive test, this implied this participant had the controlled productive vocabulary of approximately 1,000 lemmas and did not know how to use around 1,000 lemmas at a sentence level. Differently, the participants' free productive vocabulary could not be used to estimate the number of lemmas they did not know the meaning of or that of lemmas they did not know how to use at a paragraph or essay level. The number of the lemmas that the participants knew the meaning and knew how to use but did not write in the free writings could not be investigated either.

These are the limitations of the data on the free productive vocabulary knowledge in the present study. The study, therefore, focuses on only the lemmas the participants knew and/or provided in the free productive test.

Examples of how I scored the lemma uses from a participant's free writing

I explain how to grade the use of each lemma. The free writing of the first student in Year 1 was again used as an example. Only six of the lemmas the student produced were shown here as they belonged to every word class of content words (nouns, verbs, adjectives, and adverbs). They were also from all the frequency levels, except the 10,000-word level – the lemmas from his composition did not fall into this level.

Example 1: Checking the accuracy of the usage of the lemma '**liberty**' (a noun in the 2,000-word level)

The excerpt: "All humans are born free with **liberty** and dignity which means we deserve to do anything or be anything which makes us happy."

I first checked the meaning of this lemma by using Oxford Advanced Learner's Dictionary Online. This lemma means "freedom to live as you choose without too many limits from government or authority". From the surrounding text, he used the words 'free' and the phrase 'to do anything'. These matched the meaning provided by the dictionary. I then gave one point to the sentence for the correct meaning.

Next, I checked its grammatical functions. According to the dictionary, this lemma is an uncountable noun. The participant used this lemma as an uncountable noun coming after a preposition. However, whether this lemma can be used with the preposition 'with' is not provided by this dictionary. I then used the Sketch Engine (Kilgarriff et al., 2014) to search the English Web 2013 (enTenTen13) corpus in order to identify typical patterns. Word Sketch showed that the lemma 'liberty' could come after the preposition 'with'.

I also examined the concordances with the prepositional phrase 'with liberty', and I found the following sentences as shown in Table 6.21. The sentences used a similar structure to the sentence produced by the participant.

1. I was **born** an American woman **with liberties** and rights that are the envy of the world.
2. and unlike many in much of the world today, we're **born with liberty** and face no great struggle in keeping most of it.

Table 6.21: Examples of concordances with the prepositional phrase 'with liberty' found in English Web 2013 (enTenTen13) corpus

Based on what the references provided, the lemma 'liberty' was used in a sentence with both semantic accuracy and grammatical appropriateness. Therefore, two points were given to this lemma.

Example 2: Checking the accuracy of the usage of the lemma '**fade**' (a verb in the 3,000-word level)

The excerpt: "On the contrary, the revenue of spending some money maybe something you want and you may consider this as your happiness, but as time goes by, this kind of happiness will **fade** away one day."

I started by looking up this lemma in Oxford Advanced Learner's Dictionary Online.

According to the dictionary, one of the meanings of the lemma 'fade' is "to disappear gradually", and this meaning fitted in the sentence produced by the participant. The surrounding context such as 'times goes by' and 'away' supported this meaning. So, one point was given to the correct meaning in the sentence.

In terms of grammatical functions, the dictionary shows that the lemma can be used as an intransitive verb or a transitive verb. It also emphasises the frequently used phrase 'fade away' and provides some examples such as "The laughter faded away."

The participant used it as an intransitive verb as in 'will fade away', which matched the frequently used phrase provided by the dictionary. For certainty, I also used the Sketch Engine to identify whether 'fade away' was a typical pattern for the lemma 'fade'. I also looked at the concordances to find whether the lemma could be used with the subject 'happiness', and I found two sentences shown in Table 6.22.

1. This feeling of **happiness faded away** and it was immediately replaced by a severe reality.
2. That **happiness faded away** and now I have new purposes, a new drive.

Table 6.22: Examples of concordance with 'happiness' and 'fade away'

Therefore, this lemma was used with grammatical appropriateness. One more point was then given to this lemma. Together with the meaning score, it made two points for the lemma 'fade' produced by this participant.

Example 3: Checking the accuracy of the usage of the lemma '**eternal**' (an adjective in the 3,000-word level)

The excerpt: "I strongly believe that if you can look at the world through positive lens of your eyes and mind, without ego, ignorance or prejudice, happiness will be lit and it will be like an **eternal** flame shining in your mind no matter how many decades pass by. By this way, richness or poverty would not be an obstacle to be happy."

Like the previous lemmas, I checked the meaning of the lemma 'eternal' in Oxford Advanced Learner's Dictionary Online. There are two meanings of the lemma 'eternal': 1) "without an end; existing or continuing forever" and 2) "happening often and seeming never to stop".

The surrounding words used by the participant such as 'happiness', 'positive', and 'flame' and 'no matter how many decades pass by' seemed to support the first meaning for the lemma 'eternal'. Hence, the meaning "without an end; existing or

continuing forever" then fitted the participant's sentence. Then one point was given to this lemma.

The dictionary shows that the lemma is an adjective. In every one of the examples it provides, the lemma is used before a noun. This is also how the participant used it in his free writing as well.

To check whether the lemma 'eternal' can be used with the noun 'flame', I used the Sketch Engine and Word Sketch showed 'eternal flame' in the column of nouns modified by 'eternal' (Table 6.23).

1. Our love is like an **eternal flame**: No one can touch it, it can't be blown out, it carries many colors, and burns until the end of time.
2. As one of the most popular ceremonies, the third candle often represents love's **eternal flame.**

Table 6.23: Examples of concordances with 'eternal flame'

I checked the concordances, and they also showed that the lemma 'eternal' could be used with the noun 'flame'. I then gave one point to the lemma 'eternal' produced in his free writing for correct grammar. Hence, two points were given to the usage of this lemma.

Example 4: Checking the accuracy of the usage of the lemma '**nowadays**' (an adverb in the 5,000-word level)

The excerpt: "However, **nowadays** money has become one of the most important factors that everyone is required to have in order to survive in this such a world and society."

In Oxford Advanced Learner's Dictionary Online, the lemma 'nowadays' means "at the present time, in contrast with the past". This meaning fitted in the context of the excerpt of Example 4.

In the excerpt, the lemma modified the sentence and used with the verb 'become' in present perfect tense. I used the Sketch Engine and Word Sketch showed that the lemma and the verb could be used together (Table 6.24).

1. Many people **nowadays become** stuck employed in a normal workplace.
2. Information **nowadays becomes** dated very quickly, which is why it's important to create new videos as often as possible.

Table 6.24: Example of concordances with 'nowadays' and 'become'

All the evidence mentioned above suggested that the lemma 'nowadays' was used with semantic and grammatical accuracy by this participant and two points were given to this lemma.

Example 5: Checking the accuracy of the usage of the lemma '**revenue**' (a noun in the academic word level)

The excerpt: "On the contrary, the **revenue** of spending some money maybe something you want and you may consider this as your happiness, but as time goes by, this kind of happiness will fade away one day."

I looked up the lemma 'revenue' in Oxford Advanced Learner's Dictionary Online firstly and found that the meaning provided in the dictionary did not seem to fit in the context of the sentence produced by the participant. According to the dictionary, the lemma 'revenue' means "the money that a government receives from taxes or that an organization, etc. receives from its business". However, surrounding words written by the participant such as 'spending' seemed to contradict the lemma itself.

This showed that the lemma 'revenue' was not used correctly in meaning, so I did not give any point and then did not check its grammatical functions.

Example 6: Checking the accuracy of the usage of the lemma '**factor**' (a noun in the academic word level)

The excerpt: "However, nowadays money has become one of the most important **factors** that everyone is required to have in order to survive in this such a world and society."

According to Oxford Advanced Learner's Dictionary Online, the lemma 'factor' is a polysemous word. It has five meanings: 1) "one of the several things that cause or influence something", 2) "a number that divides into another number exactly", 3) "the amount by which something increases or decreases", 4) "a particular level on a scale of measurement", and 5) "a substance in the blood that helps the clotting process. There are several types of this substance".

In the main clause of the sentence produced by the participant, the first meaning of the lemma 'factor' seemed to fit in. However, the surrounding words in the relative clause modifying the lemma 'factor' did not seem to support this meaning. The verbs 'require' and 'have' as in 'require factors' or 'have factors to survive' did not suggest the key words 'cause' or 'influence' in the meaning provided by the dictionary at all. Instead, the lemma 'factor' should be replaced by 'basic needs'.

In terms of grammatical functions, I used the Sketch Engine (Kilgarriff et al., 2014) to identify typical patterns of the lemma 'factor'. Word Sketch did not show whether the lemma could be used as an object of the verb 'require'.

Because the sentence with the lemma 'factor' produced by this participant was not clear enough and even though the two references were used, the two raters needed to grade this sentence.

All the six examples above show how I graded the accuracy of the lemmas produced in the participants' free writings. Four of the examples show the correct usage of the lemmas in both meaning and grammar. Example 5 shows that the lemma was used with incorrect meaning; while the lemma in Example 6 seemed vague in meaning, so the two raters were asked to grade it.

After I graded all the lemmas used by a participant, I converted the scores into percentages so that I could compare these percentages to the percentages from the receptive and the controlled productive tests in order to examine the relationship among them. The percentage scores of the lemmas used correctly by a participant were calculated based on the possible total scores of all the lemmas produced by such participant. In this way, the percentages would not be biased because of the length of that participant's essay.

6.3.5 Analysing the relationship among receptive, controlled productive and free productive vocabulary knowledge

I analysed the relationship among receptive, controlled productive and free productive vocabulary knowledge in three ways. I firstly compared the vocabulary scores or percentages from the three tests in terms of differences and ratios among them. This is the quantitative analysis within the dichotomous approach. The three scores of all the four year cohorts were also compared.

The second one is the quantitative analysis within the developmental approach. I averaged the developmental scores of the test lemmas at each frequency levels and of each year cohorts before comparing them. I also analysed the data qualitatively by examining some participants' answers to the three vocabulary tests. This analysis is not the main analysis in the present study, but it was conducted for more insights about the relationship between receptive and productive vocabulary knowledge.

Chapter 7 Findings 1

Participants' background and receptive vocabulary knowledge

7.1   Introduction

Chapter 7 first presents the participants' background information received from the questionnaire. I analysed the data by using descriptive statistics. The main focus of the chapter is addressing the first research question about the participants' receptive vocabulary knowledge in relation to word frequency levels and at different years of study.

Receptive vocabulary knowledge in this study refers to the knowledge of the forms and the meanings of the lemmas at each frequency level. The unit for counting words in the study was the lemma. The five frequency levels focused in this study were the 2,000-word level, the 3,000-word level, the 5,000-word level, the 10,000-word level, and the academic word level. Each of the five levels included different number of lemmas. The 2,000-word level included 2,000 lemmas. The 3,000-, the 5,000-, and the 10,000-word levels included 1,000 lemmas each. The academic word level included 570 lemmas. The participants were the 111 English-major students in Thailand. The first- and the second-year students were grouped as the low-proficiency group, while the third- and the fourth-year students were grouped together as the higher year group. The receptive scores of the four year cohorts were compared to examine the receptive vocabulary knowledge at different years of study.

7.2   Participants' background

In this section, I present information about the participants' general background information: the number of the participants of each year group, their age, gender, types of secondary schools, school programmes, English learning experience, exposure to English in English-speaking countries, and number of hours they usually spent doing different activities to practice English outside class.

Year of study	Number of participants	%	Proficiency group	Number of participants	%
1	34	30.63	Low	58	52.25
2	24	21.62			
3	29	26.13	High	53	47.75
4	24	21.62			
Total	111	100.00	Total	111	100.00

Table 7.1: Number of the participants of each year group

Among the four year cohorts, the biggest group was the first-year students (34 students, or 30.63%). There were more participants in a low-proficiency group (58 participants, or 52.25%) than in the higher year group (53 participants, or 47.75%).

Year of study	Age (years old)	Number of participants	%
1	19	14	41.18
	20	7	20.59
	21	1	2.94
	Total	34	100.00
2	20	12	50.00
	21	2	8.33
	22	1	4.17
	Total	24	100.00
3	21	14	48.28
	22	5	17.24
	Total	29	100.00
4	22	14	58.33
	23	5	20.83
	Total	24	100.00

Table 7.2: Participants' age

Most of the first-year participants were nineteen years old (41.18%). The second-year participants on average were 20 years old. The third-year participants were around 21, and the fourth-year participants were 22 years old.

Year of study	Gender				
	Male	%	Female	%	Total%
1	6	17.65	28	82.35	100.00
2	2	8.33	22	91.67	100.00
3	7	24.14	22	75.86	100.00
4	2	8.33	22	91.67	100.00
Total	17	15.32	94	84.68	100.00

Table 7.3: Number of male and female participants

Around 85% of the participants from all the four year cohorts were females, and 15% were males.

Year of study	Secondary school type							
	Thai	%	English	%	International	%	Total%	
1	29	85.29	5	14.71	0	0.00	100.00	
2	22	91.67	2	8.33	0	0.00	100.00	
3	25	86.21	4	13.79	0	0.00	100.00	
4	23	95.83	1	4.17	0	0.00	100.00	
Total	99	89.19	12	10.81	0	0.00	100.00	

Table 7.4: Number of participants in different secondary school types

None of the participants had studied in an international programme. Approximately 90% of them finished their secondary school with a Thai programme, while around 10% of them finished their school with an English programme.

Year of study	Programme in a secondary school								
	S&M*	%	E&M*	%	E&L*	%	Others	%	Total%
1	24	70.59	1	2.94	7	20.59	2	5.88	100.00
2	19	79.17	0	0.00	5	20.83	0	0.00	100.00
3	27	93.10	0	0.00	2	6.90	0	0.00	100.00
4	16	66.67	0	0.00	8	33.33	0	0.00	100.00
Total	86	77.48	1	0.90	22	19.82	2	1.80	100.00

*Note: S&E stands for science and math, E&M stands for English and math, and E&L stands for English and another language.

Table 7.5: Number of participants from different programmes in secondary schools

Around 77% had studied in the science and math programme in secondary schools, while 20% of them studied in English and another foreign language programme. A few of them (3%) studied in English and math programme and other programmes.

Year of study	Number of years having been studying English	Number of months studying or living in English-speaking countries
1	13.53	3.91
2	13.87	2.04
3	15.24	3.36
4	16.67	2.58
Average	**14.73**	**3.08**

Table 7.6: Number of years studying English and months living/studying in an English-speaking countries

The average number of years that the participants had been studying English ranged from around 14 to 17 years. The older they were, the longer time they had been studying English.

The average number of months that the participants from each year cohort spent in English-speaking countries was 3 months, with the first-year participants having spent the longest time (3.91 months), followed by the third-year participants, the fourth-year participants, and the second-year participants.

Listening	Watching	Speaking
songs	movies/series with and without subtitles	to foreign friends
podcasts	Videos on YouTube	to Thai friends
BBC radio	TV shows	to himself/herself
	TED talks	
Reading	**Writing**	**Others**
textbooks	posts on social media such as Facebook and Instagram	thinking in English
grammar books		singing and learning from lyrics
fictions		playing board games such as crossword
non-fictions		using flash cards
online news		doing English tests such as TOEFL tests
online articles		tutoring English
comics		joining a language exchange club
		taking an extra English conversation course
		practicing English pronunciation

Table 7.7: Activities that the participants did to practice English

Year of study	Number of hours per month participants spent practicing English through these skills							
	Listening	Watching	Speaking	Reading	Writing	Others	Total per month	Total since admission
1	8.94	21.21	2.24	7.68	0.00	3.79	**43.86**	526
2	25.42	17.71	0.33	6.46	0.00	1.04	**50.96**	1,716
3	15.90	27.86	0.76	5.66	0.00	5.93	**56.11**	2,020
4	9.54	21.71	1.71	4.38	0.67	3.04	**41.05**	1,970
Average per month	**14.45**	**22.30**	**1.32**	**6.17**	**0.14**	**3.59**	N/A	N/A

Table 7.8: Number of hours per month the participants spent practicing English outside classroom

The participants practiced English outside class in different ways as shown in Table 7.7. The participants on average spent the longest hours (22.30 hours per month) watching movies, YouTube videos, English TV shows and TED talks, followed by listening activities (14 hours per month) such as listening English songs, podcasts,

and BBC radio. Activities that the participants on average did not spend much time doing were reading, speaking, writing, and other activities (e.g., singing, playing board games, tutoring English, etc.).

I assumed that the participants had been practicing English outside classroom in these ways since their admission. The total number of hours that each year cohort spent per month was then multiplied by 12 months and by the number of years they had been in the English major. The third years students had spent most time practicing English outside classroom (2020 hours within three years), compared to the other three year cohorts.

7.3   Research question 1: What is the participants' receptive vocabulary knowledge?

In this section, I present the findings about the participants' receptive vocabulary knowledge in the order of the two sub-questions as shown below.

Research question 1: What is the participants' receptive vocabulary knowledge?

> **Sub-question 1:** What is the participants' receptive vocabulary knowledge at different word frequency levels?

> **Sub-question 2:** What is the participants' receptive vocabulary knowledge at different years of study?

7.3.1   Sub-question 1: What is the participants' receptive vocabulary knowledge at different word frequency levels?

This section presents the findings about the participants' receptive vocabulary knowledge at different frequency levels: the 2,000-, the 3,000-, the 5,000-, the 10,000-, and the academic word levels.

The participants' knowledge was evaluated based on their translations or the synonyms supplied by the participants at each of the five frequency levels (two points for correct translations/synonyms, one point for partially correct translations/synonyms, zero point for incorrect translations/ synonyms, and 20 points as a maximum score), I analysed the scores and the percentage by using descriptive statistics in SPSS. The findings are shown in Table 7.9

and Figure 7.1. The percentage scores of the receptive vocabulary refer to the lemmas that the participants on average knew their meanings out of 100 lemmas in each of the five frequency levels.

Frequency level	Mean score	%	SD	%
2,000 words	18.40	91.98	1.69	8.43
3,000 words	16.09	80.45	2.23	11.15
5,000 words	8.83	44.14	3.44	17.20
10,000 words	2.15	10.77	2.10	10.51
academic words	16.09	80.45	3.15	15.73

Table 7.9: Participants' receptive vocabulary knowledge scores at the five different word frequency levels
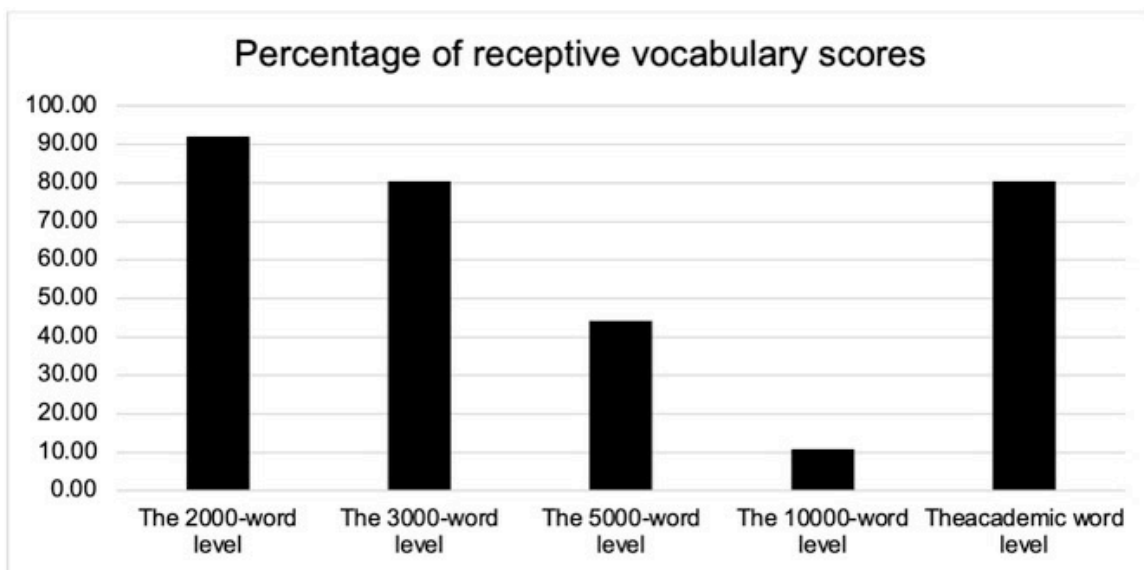
Figure 7.1: Percentage of receptive vocabulary scores at the five different word frequency levels

The participants' receptive vocabulary scores were lower when the word frequency levels became lower. The score at the 2,000-word level was the highest (91.98%), followed by that of the 3,000-, the 5,000-, and the 10,000-word levels (80.45%, 44.14%, 10.77%, respectively). The score at the academic word level was equal to that at the 3,000-word level (80.45%).

In this study, the mastery level of the receptive vocabulary knowledge at a frequency level means being able to recognise the written form and knowing the meaning of most of the words at such frequency level with correct meaning and appropriate grammar. If 80 per cent of the score at each frequency level is the threshold of mastery level, the participants mastered receptive vocabulary knowledge at the three frequency levels: the 2,000-, the 3,000-, and the academic word levels.

Hence, word frequency levels affected the participants' receptive vocabulary knowledge. The participants tended to know the meanings of more lemmas at the higher frequency levels and of fewer lemmas at the lower frequency levels. In addition, the participants reached the mastery levels of vocabulary at the high-frequency levels and the academic word level.

I ran a one-way repeated measures ANOVA to test statistical significance of the findings. The p-value of less than 0.001 from the tests of within-subjects effects and the tests of within-subjects contrasts showed that there was very low probability that the significant differences of the receptive scores at the five frequency levels were through random error. In terms of pairwise comparisons, the receptive scores at all the frequency levels were significantly different, except those between the 3,000-word level and the academic word level (the p-value of 1.000).

This can be interpreted that the receptive scores among the frequency levels were statistically significant, or the frequency levels affected the receptive scores. The differences in scores at the 3,000-word level and the academic word level were not statistically significant, though.

7.3.2   Sub-question 2: What is the participants' receptive vocabulary knowledge at different years of study?

This section attempts to answer the second sub-question of the first research question. It is about the participants' receptive vocabulary knowledge at different years of study. Only the mean scores of the known receptive vocabulary were used and referred to receptive vocabulary knowledge scores.

The present study is a cross-sectional study, so the receptive vocabulary knowledge scores of the participants from different year cohorts were compared. This involves two comparisons. The first comparison is between the higher year group (Year 3 students and Year 4 students) and the lower year group (Year 1 students and Year 2 students). The second comparison is among the four year cohorts.

Comparison 1: Receptive vocabulary knowledge between the higher year group and the lower year group

A normality test was conducted. The p-values of the Shapiro-Wilk test were above 0.05 (0.081 for the lower year group and 0.704 for the higher year group). We can assume that the data were approximately normally distributed. To check the assumption of homogeneity of variances, the Levene's test was conducted. The p-value (0.067) was greater 0.05, so the variances between groups were not significantly different and the assumption of equality of variance was met.

I compared the average scores of the receptive vocabulary knowledge of the higher year group to those of the lower year group by using independent samples t-test and Cohen's (1988) effect size. There were 50 target lemmas for all the five frequency levels, or ten lemmas per frequency level. Two points were given to a correct translation/synonym, so the possible score was 100. The statistics (with equal variances assumed) are shown in Table 7.10.

Year group	N	Mean score (Out of 100)	SD	MD	Sig. (2-tailed)	Effect size
Higher	53	66.62	7.52	9.69**	0.000	1.12
Lower	58	56.93	9.62			

Note: ** = statistically significant at the 0.001 level

Table 7.10: Receptive vocabulary scores of the higher and the lower year groups

As I expected, the finding showed that the higher year group highly significantly had higher receptive vocabulary knowledge score than the lower year group (approximately 10%) with the p value of less than 0.001. The effect size of this t-test (1.12) was found to exceed Cohen's (1988) convention of a large effect size (0.80).

I then compared their receptive vocabulary knowledge scores at the five frequency levels to examine at which frequency levels the receptive scores between the two groups were statistically significant. The percentages, the significance, and the effect size are shown in Table 7.11.

Frequency level	Year group	N	Mean	%	SD	%	MD	%	Sig. (2-tailed)	Effect size
2,000 words	Higher	53	19.21	96.04	1.22	6.07	1.55**	7.76**	0.000	1.04
	Lower	58	17.66	88.28	1.72	8.61				
3,000 words	Higher	53	16.94	84.72	1.88	9.38	1.63**	8.17**	0.000	0.79
	Lower	58	15.31	76.55	2.26	11.29				
5,000 words	Higher	53	9.94	49.72	3.14	15.70	2.13*	10.67*	0.001	0.65
	Lower	58	7.81	39.05	3.41	17.05				
10,000 words	Higher	53	2.77	13.87	1.80	9.02	1.19*	5.94*	0.003	0.59
	Lower	58	1.59	7.93	2.21	11.04				
academic words	Higher	53	17.75	88.77	2.24	11.18	3.19**	15.93**	0.000	1.18
	Lower	58	14.57	72.84	3.10	15.51				

Note: * = statistically significant at the 0.05 level, ** = statistically significant at the 0.001 level

Table 7.11: Receptive vocabulary knowledge scores between the higher year group and the lower year group at the five frequency levels
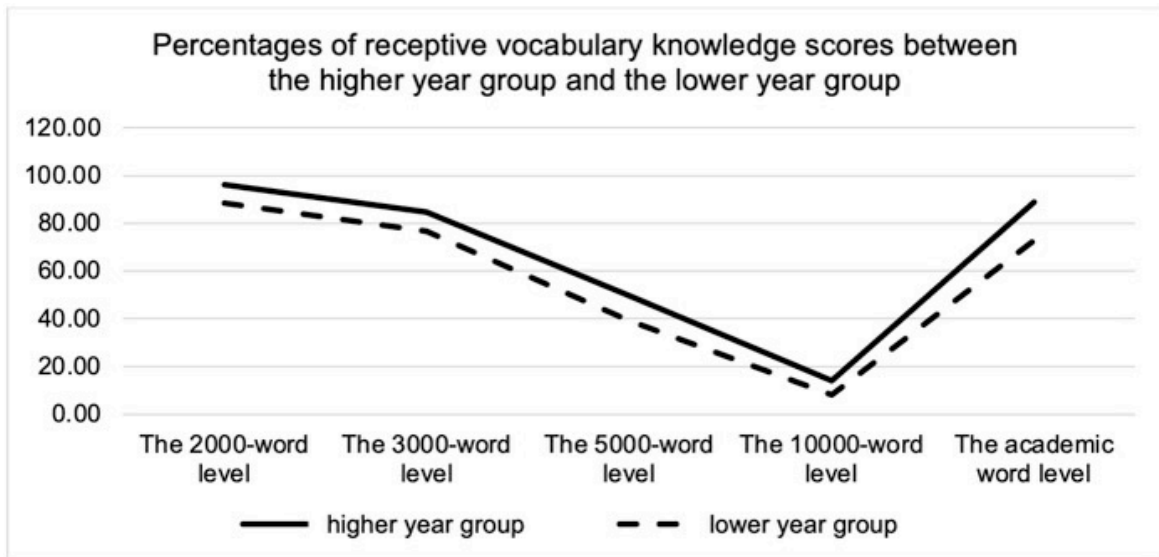
Figure 7.2: Percentages of the receptive vocabulary knowledge scores between the higher year group and the lower year group at the five frequency levels

The findings showed that the higher year group highly significantly had higher receptive vocabulary knowledge scores than the lower year group at all the five frequency levels with the p values of less than 0.001 at the 2,000-, the 3,000- and the academic word levels and the p values of less than 0.005 at the 5,000- and the 10,000-word levels. According to Cohen's (1988) convention, the effect sizes were found to be large at the 2,000-, the 3,000- and the academic word levels, and were found to be medium at the 5,000- and the 10,000-word levels.

The biggest difference in the scores between these two year groups was 15.93% at the academic word level, while the smallest difference was 5.94% at the 10,000-word level. However, the differences at all the five levels were broadly similar. According to Figure 7.2, the lines seemed to run parallel to each other.

In terms of vocabulary mastery, the higher year group mastered the vocabulary at the three frequency levels: the 2,000-, the 3,000-, and the academic word levels, while the lower year group mastered the vocabulary at the 2,000-word level only and nearly reached the mastery level at the other two levels. The higher year group mastered the vocabulary at more levels than the lower year group.

Comparison 2: Receptive vocabulary knowledge at different years of study

When I found that there was high significance of the differences in receptive vocabulary knowledge scores at all the five frequency levels between the higher year group and the lower year group, I then decided to analyse further to examine the receptive vocabulary knowledge at different years of study (from Year 1 to Year 4) and whether there would be significant differences among the four year cohorts.

The number of the participants per year cohort did not reach the optimal number for the comparison, though. According to Cohen et al. (2017, p.203), "Typically an anticipated minimum of thirty cases per variable should be used as a 'rule of thumb', i.e., one must be assured of having a minimum of thirty cases for each variable" for reliable statistics.

A normality test was conducted. The p-values of the Shapiro-Wilk test were above 0.05 (0.058, 0.524, 0.357, and 0.47 for the first-year group to the fourth-year group). We can assume that the data were approximately normally distributed. To check the assumption of homogeneity of variances, the Levene's test was conducted. The p-value (0.138) was greater 0.05, so the assumption of equality of variances was met.

After checking these assumptions, I analysed their receptive vocabulary knowledge scores (out of 100 scores – five frequency levels, ten lemmas each, and two points

for each lemma) by using descriptive statistics in One-Way ANOVA. The statistics are shown in Table 7.12 and Figure 7.3.

Year of study	N	Mean score (Out of 100)	SD
1	34	55.88	10.66
2	24	58.42	7.91
3	29	64.10	7.98
4	24	69.67	5.72
Average	N/A	61.56	9.92

Table 7.12: Receptive vocabulary knowledge scores of the four year cohorts
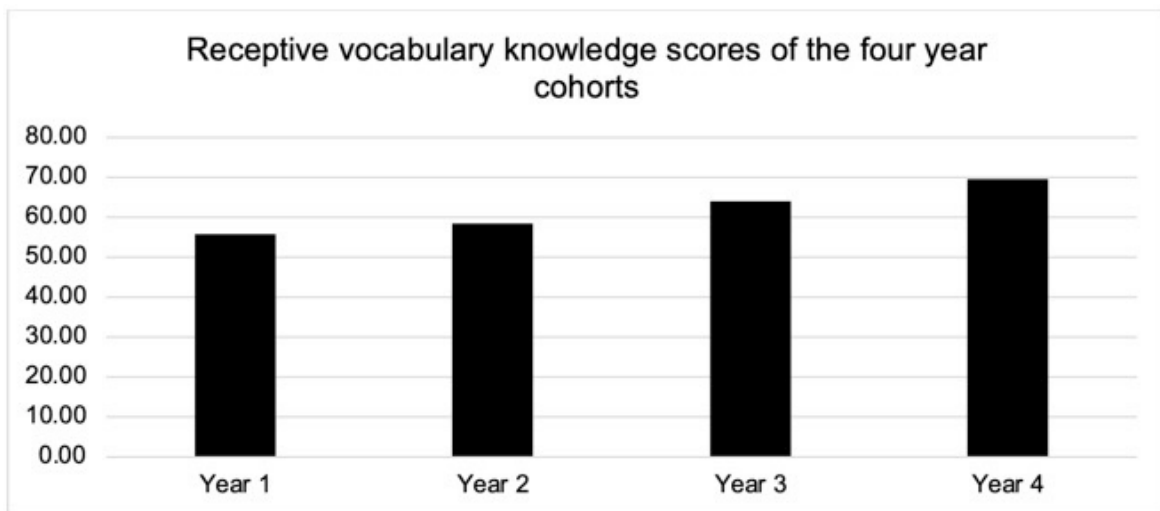
Figure 7.3: Receptive vocabulary knowledge scores of the four year cohorts

The findings showed that the participants in later years of study had higher average scores of receptive vocabulary knowledge than those in earlier years of study with less variation for the higher year group.

I also examined the participants' receptive vocabulary knowledge scores of the four year cohorts at each of the five frequency levels by using descriptive statistics and a line graph as shown in Table 7.13 and Figure 7.4.

Frequency level	Year of study	N	Mean	%	SD	%
2,000 words	1	34	17.32	86.62	1.45	7.25
	2	24	18.13	90.63	1.99	9.93
	3	29	18.93	94.66	1.36	6.81
	4	24	19.54	97.71	0.93	4.66
	Average	**N/A**	**18.40**	**91.98**	**1.69**	**8.43**
3,000 words	1	34	14.97	74.85	2.42	12.09
	2	24	15.79	78.96	1.96	9.78
	3	29	16.45	82.24	1.79	8.92
	4	24	17.54	87.71	1.84	9.21
	Average	**N/A**	**16.09**	**80.45**	**2.23**	**11.15**
5,000 words	1	34	8.09	40.44	3.27	16.35
	2	24	7.42	37.08	3.64	18.17
	3	29	8.93	44.66	3.43	17.16
	4	24	11.17	55.83	2.26	11.29
	Average	**N/A**	**8.83**	**44.14**	**3.44**	**17.20**
10,000 words	1	34	1.74	8.68	2.72	13.61
	2	24	1.38	6.88	1.17	5.86
	3	29	2.62	13.10	1.86	9.30
	4	24	2.96	14.79	1.76	8.78
	Average	**N/A**	**2.15**	**10.77**	**2.10**	**10.51**
academic words	1	34	13.76	68.82	3.45	17.24
	2	24	15.71	78.54	2.12	10.58
	3	29	17.17	85.86	2.62	13.10
	4	24	18.46	92.29	1.41	7.07
	Average	**N/A**	**16.09**	**80.45**	**3.15**	**15.73**

Table 7.13: Receptive vocabulary knowledge scores of the four year cohorts at the five different frequency levels
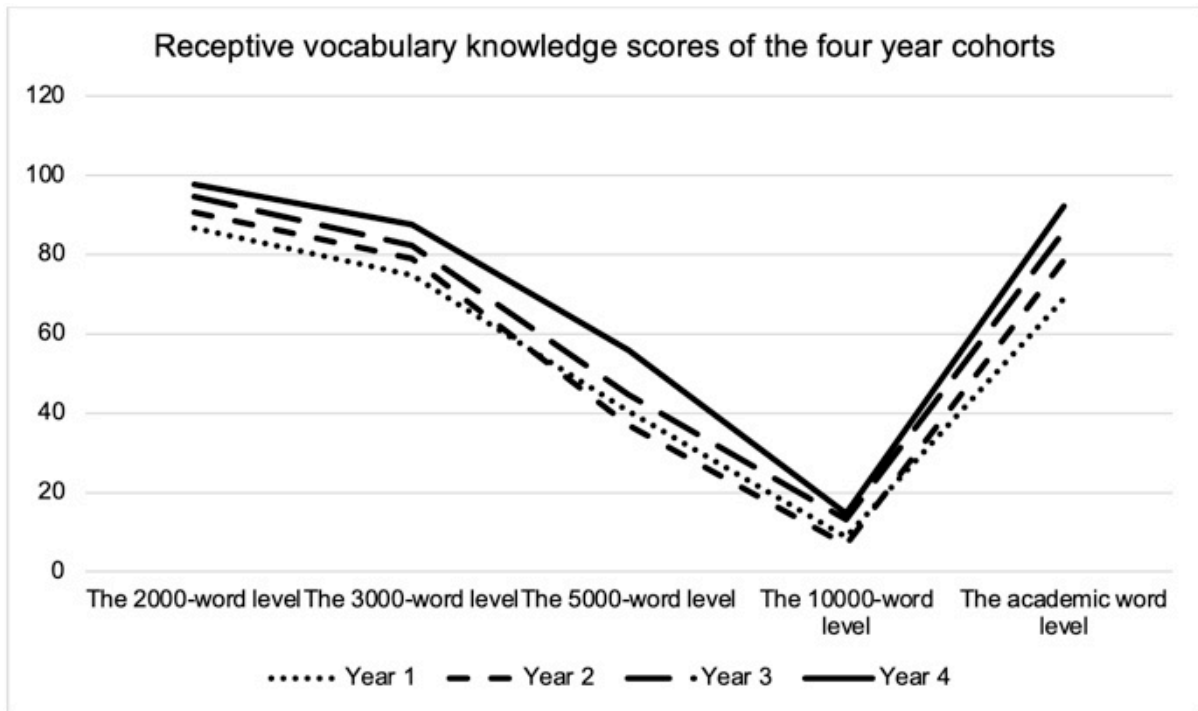
Figure 7.4: Receptive vocabulary knowledge scores of the four year cohorts at the five different frequency levels

When I compared the participants' receptive vocabulary knowledge scores at each of the five frequency levels, I found that the participants in the later years had larger receptive vocabulary knowledge at all the five frequency levels. However, at some frequency levels (the 5,000- and the 10,000-word levels), the first-year students had slightly higher scores than the second-year students.

To check whether these differences were statistically significant, I analysed the data by using MANOVA and One-Way ANOVA with Tukey HSD post hoc test. The findings are shown in Table 7.14 and I summarised the significant differences in Table 7.14. The effect sizes (partial eta squared) are the values of partial eta squared of the tests of between-subjects effects.

Frequency level	(A) Year of study	(B) Year of study	MD (A-B)	%	Sig.	Partial Eta Squared
		Multiple Comparisons				
		Post hoc test: Tukey HSD				
2,000 words	1	2	-0.80	-4.01	0.179	0.26
		3	-1.61**	-8.04**	0.000	
		4	-2.22**	-11.09**	0.000	
	2	1	0.80	4.01	0.179	
		3	-0.81	-4.03	0.200	
		4	-1.42*	-7.08*	0.006	
	3	1	1.61**	8.04**	0.000	
		2	0.81	4.03	0.200	
		4	-0.61	-3.05	0.439	
	4	1	2.22**	11.09**	0.000	
		2	1.42*	7.08*	0.006	
		3	0.61	3.05	0.439	
3,000 words	1	2	-0.82	-4.11	0.438	0.18
		3	-1.48*	-7.39*	0.026	
		4	-2.57**	-12.86**	0.000	
	2	1	0.82	4.11	0.438	
		3	-0.66	-3.28	0.652	
		4	-1.75*	-8.75*	0.019	
	3	1	1.48*	7.39*	0.026	
		2	0.66	3.28	0.652	
		4	-1.09	-5.47	0.219	
	4	1	2.57**	12.86**	0.000	
		2	1.75*	8.75*	0.019	
		3	1.09	5.47	0.219	
5,000 words	1	2	0.67	3.36	0.861	0.15
		3	-0.84	-4.21	0.727	
		4	-3.08*	-15.39*	0.003	
	2	1	-0.67	-3.36	0.861	
		3	-1.51	-7.57	0.324	
		4	-3.75*	-18.75*	0.001	
	3	1	0.84	4.21	0.727	
		2	1.51	7.57	0.324	
		4	-2.24	-11.18	0.062	
	4	1	3.08*	15.39*	0.003	

		2	3.75*	18.75*	0.001	
		3	2.24	11.18	0.062	
10,000 words	1	2	0.36	1.80	0.911	0.09
		3	-0.89	-4.43	0.319	
		4	-1.22	-6.12	0.116	
	2	1	-0.36	-1.80	0.911	
		3	-1.25	-6.23	0.125	
		4	-1.58*	-7.92*	0.040	
	3	1	0.89	4.43	0.319	
		2	1.25	6.23	0.125	
		4	-0.34	-1.69	0.932	
	4	1	1.22	6.12	0.116	
		2	1.58*	7.92*	0.040	
		3	0.34	1.69	0.932	
academic words	1	2	-1.94*	-9.72*	0.032	0.33
		3	-3.41**	-17.04**	0.000	
		4	-4.69*	-23.47*	0.000	
	2	1	1.94*	9.72*	0.032	
		3	-1.46	-7.32	0.185	
		4	-2.75*	-13.75*	0.002	
	3	1	3.41**	17.04**	0.000	
		2	1.46	7.32	0.185	
		4	-1.29	-6.43	0.289	
	4	1	4.69**	23.47**	0.000	
		2	2.75*	13.75*	0.002	
		3	1.29	6.43	0.289	

Note: * = statistically significant at the 0.05 level, ** = statistically significant at the 0.001 level

Table 7.14: ANOVA analysis of receptive vocabulary knowledge scores among the four year cohorts

The 2,000- word level	MD (%)	The 3,000- word level	MD (%)	The 5,000- word level	MD (%)	The 10,000- word level	MD (%)	The academic word level	MD (%)
-		-		-		-		Y2 > Y1	9.72
Y3 > Y1	8.04	Y3 > Y1	7.39	-		-		Y3 > Y1	17.04
Y4 > Y1	11.09	Y4 > Y1	12.86	Y4 > Y1	15.39	-		Y4 > Y1	23.47
Y4 > Y2	7.08	Y4 > Y2	8.75	Y4 > Y2	18.75	Y4 > Y2	7.29	Y4 > Y2	13.75

Table 7.15: Years with significant differences in receptive vocabulary knowledge scores at the five frequency levels

According to the data, there were statistically significant differences in receptive vocabulary knowledge scores at all the five frequency levels but no significant difference between some year groups. This is summarised in Table 7.15. Most of the significant differences were between the first-year participants and the third or the fourth-year participants or between the second-year participants and the fourth-year participants. The effect sizes of the two frequency levels (the 2,000- and the academic word levels) were found to exceed Cohen's (1988) convention of a small effect.

The widest gap of their receptive knowledge was between the fourth-year and the first-year students at the academic words. Approximately, the fourth-year students knew 23.47% more academic words (134 lemmas from the AWL) than the first year cohort did. The smallest gap was between the first year group and the second year group at the 10,000-word level, with the first-year students knowing around 1.80% or only 18 more lemmas than the second-year students.

There were no significant differences in receptive vocabulary knowledge scores between the two continuous years at the five frequency levels, except that of the scores between Year 1 and Year 2 participants at the academic word level. At the 5,000- and the 10,000-word levels, the second-year students significantly had higher scores than the first-year students

In conclusion, the participants' receptive vocabulary is affected by different frequency levels; they knew more lemmas at high frequency levels and fewer lemmas at lower frequency levels. Their knowledge of academic words was similar to that at the 3,000-word level. The higher year group had a wider receptive knowledge of

lemmas. The biggest gain of the receptive vocabulary was at the academic word level, while the smallest gain was at the 10,000-word level.

Chapter 8 Findings 2

Controlled productive vocabulary knowledge

8.1   Introduction

Chapter 8 addresses the second research question about the participants' controlled productive vocabulary knowledge at the five frequency levels. The controlled productive vocabulary in this study refers to the lemmas that the participants knew how to use in sentences when the lemmas were given. The number of the test lemmas in the controlled productive vocabulary test was the same as the number of those in the receptive vocabulary test. I also investigated the changes of this knowledge over the years of study by comparing the controlled productive knowledge scores of the higher to that of the low-proficiency groups and comparing the scores of the four year cohorts. The findings are presented in the order of the two sub-questions.

8.2   Research question 2: What is the participants' controlled productive vocabulary knowledge?

Research question 2: What is the participants' controlled productive vocabulary knowledge?

>   **Sub-question 1:** What is the participants' controlled productive vocabulary knowledge at different word frequency levels?

>   **Sub-question 2:** What is the participants' controlled productive vocabulary knowledge at different years of study?

8.2.1   Sub-question 1: What is the participants' controlled productive vocabulary knowledge at different word frequency levels?

This section presents the findings about the participants' controlled productive vocabulary knowledge at the five different frequency levels: the 2,000-word level, the

3,000-word level, the 5,000-word level, the 10,000-word level, and the academic word level.

In the controlled productive test, the test lemmas were the same as those in the receptive test. When grading, I gave two points to the target lemmas that were used correctly in meaning and grammar (precisely known controlled productive vocabulary), one point to the lemmas used with correct meaning only (partially known controlled productive vocabulary), and zero points to the lemmas used incorrectly in meaning in sentences. Therefore, the maximum score for all the 50 target lemmas was 100 (20 points for each of the frequency level).

To answer the second research question, I analysed the scores by using descriptive statistics. The findings are shown in Table 8.1.

Frequency level	Mean score	%	SD	%
2,000 words	14.20	70.99	3.44	17.19
3,000 words	10.43	52.16	4.08	20.40
5,000 words	5.14	25.68	3.08	15.40
10,000 words	1.12	5.59	1.74	8.68
academic words	9.60	48.02	4.80	23.97

Table 8.1: Participants' controlled productive vocabulary knowledge scores at the five different word frequency levels
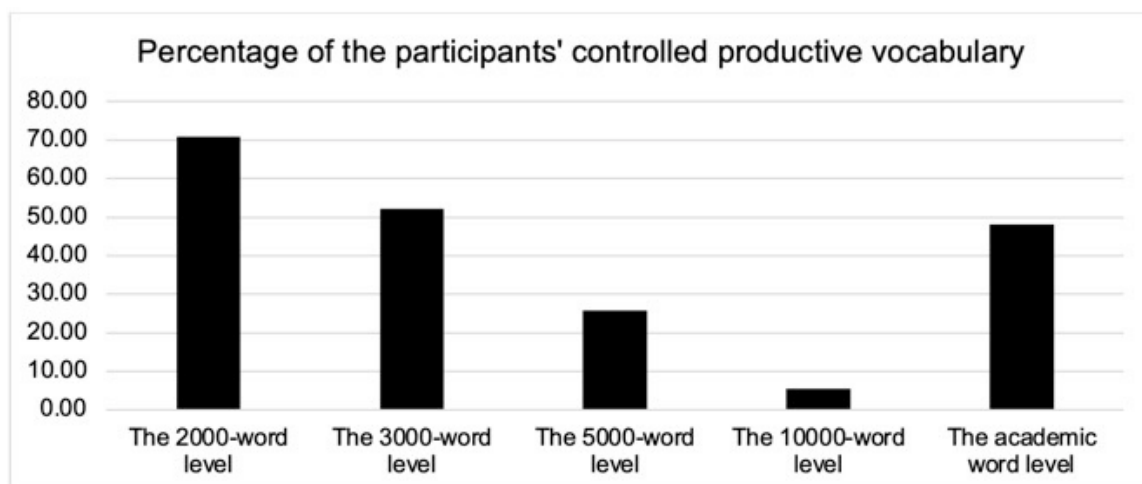
Figure 8.1: Percentage of the participants' controlled productive vocabulary knowledge scores at the five different word frequency levels

The participants' controlled productive vocabulary scores decreased when the word frequency levels became lower. The mean score at the 2,000-word level was the highest (70.99%), and that at the 10,000-word level was the lowest (5.59%). The score at the academic word level (48.02%) was between the 3,000- and the 5,000-word levels (52.16% and 25.68%). This implied that the participants had larger controlled productive vocabulary knowledge at higher frequency levels and smaller at lower frequency levels.

The mastery level of the controlled productive vocabulary knowledge at a frequency level means being able to use most of the words at such frequency level with correct meaning and appropriate grammar. The participants did not master controlled productive vocabulary at any of the five frequency levels as their controlled productive scores did not reach 80% of the total scores. However, they nearly reached the mastery level at the 2,000-word level. Therefore, word frequency levels also influence controlled productive vocabulary, and the participants did not master controlled productive level at any frequency level.

I ran a one-way repeated measures ANOVA to test statistical significance of the findings. The p-value of less than 0.001 from the tests of within-subjects effects and the tests of within-subjects contrasts showed that there was very low probability that the significant differences of the controlled productive vocabulary scores at the five frequency levels were through random error. In terms of pairwise comparisons, the controlled productive scores at all the frequency levels were significantly different, except those between the 3,000-word level and the academic word level (the p-value of 0.091).

This can be interpreted that the controlled productive scores among the frequency levels were statistically significant, or the frequency levels affected the controlled productive scores. The differences in scores at the 3,000-word level and the academic word level were not statistically significant, though.

8.2.2 Sub-question 2: What is the participants' controlled productive vocabulary knowledge at different years of study?

Comparison 1: Controlled productive vocabulary knowledge between the higher and the lower year groups

A normality test was conducted. The p-values of the Shapiro-Wilk test were above 0.05 (0.231 for the lower year group and 0.181 for the higher year group). We can assume that the data were approximately normally distributed. To check the assumption of homogeneity of variances, the Levene's test was conducted. The p-value (0.503) was greater 0.05, so the assumption of equality of variances was met.

I compared the mean score of the higher year group with that of the low-proficiency group.

Year group	N	Mean score (Out of 100)	SD	MD	Sig. (2-tailed)	Effect size
Higher	53	46.70	11.48	11.89**	0.000	0.91
Lower	58	34.81	14.56			
Note: ** = statistically significant at the 0.001 level						

Table 8.2: Controlled productive vocabulary knowledge scores between the higher and the lower year groups regardless of word frequency levels

The t-test results showed that the higher year group highly significantly had higher controlled productive vocabulary score than the lower year group (MD = 11.89%) with the p value of less than 0.001, regardless of word frequency levels. The effect size of this t-test (0.91) was found to exceed Cohen's (1988) convention of a large effect size (0.80).

In addition, I also investigated their controlled productive vocabulary knowledge at each of the five frequency levels. I conducted an independent samples t-test and calculated the effect sizes. The statistics are shown in Table 8.3.

Frequency levels	Year group	N	Mean	%	SD	%	MD	%	Sig. (2-tailed)	Effect size
2,000 words	Higher	53	15.64	78.21	3.09	15.45	2.76**	13.81**	0.000	0.87
	Lower	58	12.88	64.40	3.22	16.12				
3,000 words	Higher	53	11.89	59.43	3.47	17.37	2.78**	13.92**	0.000	0.73
	Lower	58	9.10	45.52	4.17	20.83				
5,000 words	Higher	53	5.85	29.25	2.79	13.95	1.37*	6.83*	0.019	0.46
	Lower	58	4.48	22.41	3.21	16.04				
10,000 words	Higher	53	1.32	6.60	1.40	6.99	0.39	1.95	0.239	0.23
	Lower	58	0.93	4.66	1.99	9.95				
academic words	Higher	53	12.00	60.00	3.85	19.27	4.59**	22.93**	0.000	1.09
	Lower	58	7.41	37.07	4.54	22.69				

Note: * = statistically significant at the 0.05 level, ** = statistically significant at the 0.001 level

Table 8.3: Controlled productive vocabulary knowledge scores between the higher and the lower year groups at the five frequency levels
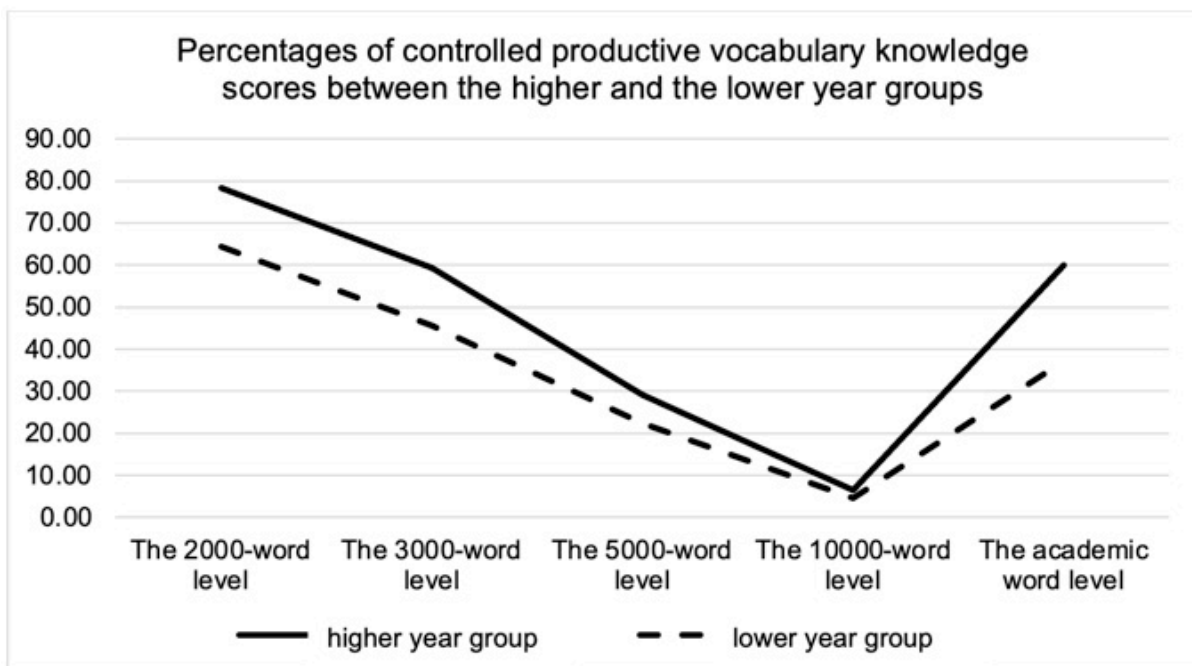
Figure 8.2: Percentages of the controlled productive vocabulary knowledge scores between the higher and the lower year groups at the five frequency levels

The findings showed that the participants from both groups scored less when the frequency became lower, with the higher year group scoring higher than the low-proficiency group.

The higher year group significantly had more controlled productive vocabulary knowledge scores than the lower year group at all the five frequency levels, except at the 10,000-word level. The p values of less than 0.001 and with the large effect sizes at the 2,000-, the 3,000- and the academic word levels, but the p value at the 5,000-word level was less than 0.005 with a small effect size (Cohen, 1988).

In Figure 8.2, the biggest gap between their scores was at the academic word level (MD = 22.93%), and the smallest gap at the 10,000-word level (MD = 1.95%).

Regarding mastery of controlled productive vocabulary knowledge, none of them reached the mastery level. However, the higher year group nearly reached the mastery level at the 2,000-word level.

Comparison 2: Controlled productive vocabulary knowledge at different years of study

Similar to the participants' receptive vocabulary knowledge, I also examined the changes of their controlled productive vocabulary knowledge at each of the five frequency levels.

A normality test was conducted. The p-values of the Shapiro-Wilk test were above 0.05 (0.077, 0.052, 0.987, and 0.762 for the first-year group to the fourth-year group). We can assume that the data were approximately normally distributed. To check the assumption of homogeneity of variances, the Levene's test was conducted. The p-value (0.284) was greater 0.05, so the assumption of equality of variances was met.

After that, I analysed their controlled productive vocabulary knowledge scores conducting descriptive statistics in One-Way ANOVA and present the statistics in Table 8.4 and Figure 8.3.

Year group	N	Mean score (Out of 100)	SD
1	34	35.32	15.48
2	24	34.08	13.45
3	29	43.31	12.10
4	24	50.79	9.37
Average	N/A	40.49	14.42

Table 8.4: Controlled productive vocabulary knowledge scores of the four year cohorts
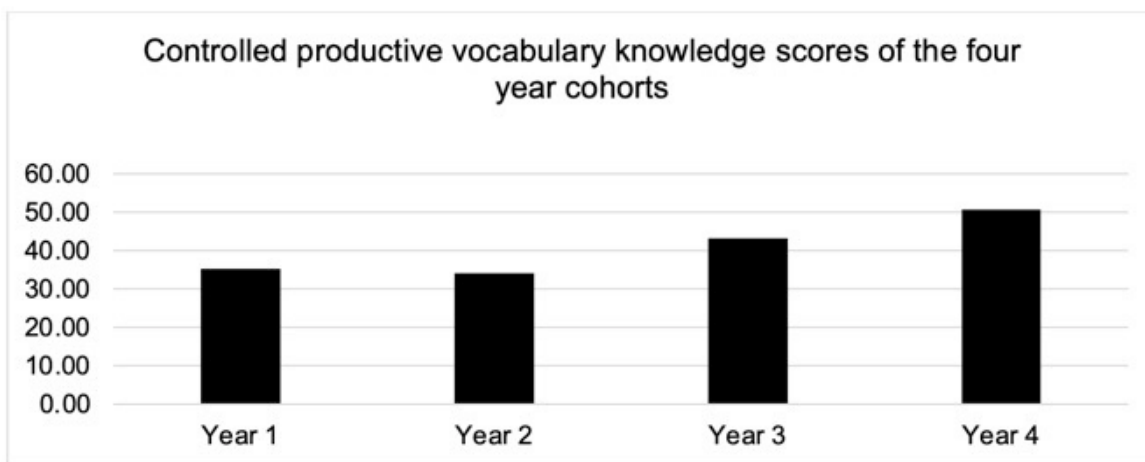
Figure 8.3: Controlled productive vocabulary knowledge scores of the four year cohorts

The findings showed that the participants in higher years of study had more controlled productive scores than those in lower years of study. The participants in Year 4 had the highest scores. However, the first-year participants had slightly higher scores than the second-year students. The standard deviation of the first-year students was slightly higher than that of the second-year students, showing that the data had more variability. Overall, there seemed to be the growth of the participants' controlled productive vocabulary knowledge over the years of study.

I also present the participants' controlled productive vocabulary knowledge scores at each of the five frequency levels as shown in Table 8.5 and Figure 8.4.

Frequency level	Year of study	N	Mean	%	SD	%
2,000 words	1	34	12.41	62.06	3.47	17.33
	2	24	13.54	67.71	2.78	13.91
	3	29	14.72	73.62	3.44	17.21
	4	24	16.75	83.75	2.19	10.96
	Average	**N/A**	**14.20**	**70.99**	**3.44**	**17.19**
3,000 words	1	34	9.41	47.06	4.16	20.82
	2	24	8.67	43.33	4.22	21.09
	3	29	11.14	55.69	3.79	18.93
	4	24	12.79	63.96	2.87	14.37
	Average	**N/A**	**10.43**	**52.16**	**4.08**	**20.40**
5,000 words	1	34	5.15	25.74	3.43	17.15
	2	24	3.54	17.71	2.65	13.27
	3	29	5.45	27.24	2.65	13.27
	4	24	6.33	31.67	2.93	14.65
	Average	**N/A**	**5.14**	**25.68**	**3.08**	**15.40**
10,000 words	1	34	1.09	5.44	2.47	12.33
	2	24	0.71	3.54	1.00	5.00
	3	29	1.03	5.17	1.35	6.75
	4	24	1.67	8.33	1.40	7.02
	Average	**N/A**	**1.12**	**5.59**	**1.74**	**8.68**
academic words	1	34	7.26	36.32	4.74	23.69
	2	24	7.63	38.13	4.33	21.66
	3	29	10.97	54.83	4.02	20.11
	4	24	13.25	66.25	3.30	16.50
	Average	**N/A**	**9.60**	**48.02**	**4.79**	**23.97**

Table 8.5: Controlled productive vocabulary knowledge scores of the four year cohorts at the five different frequency levels
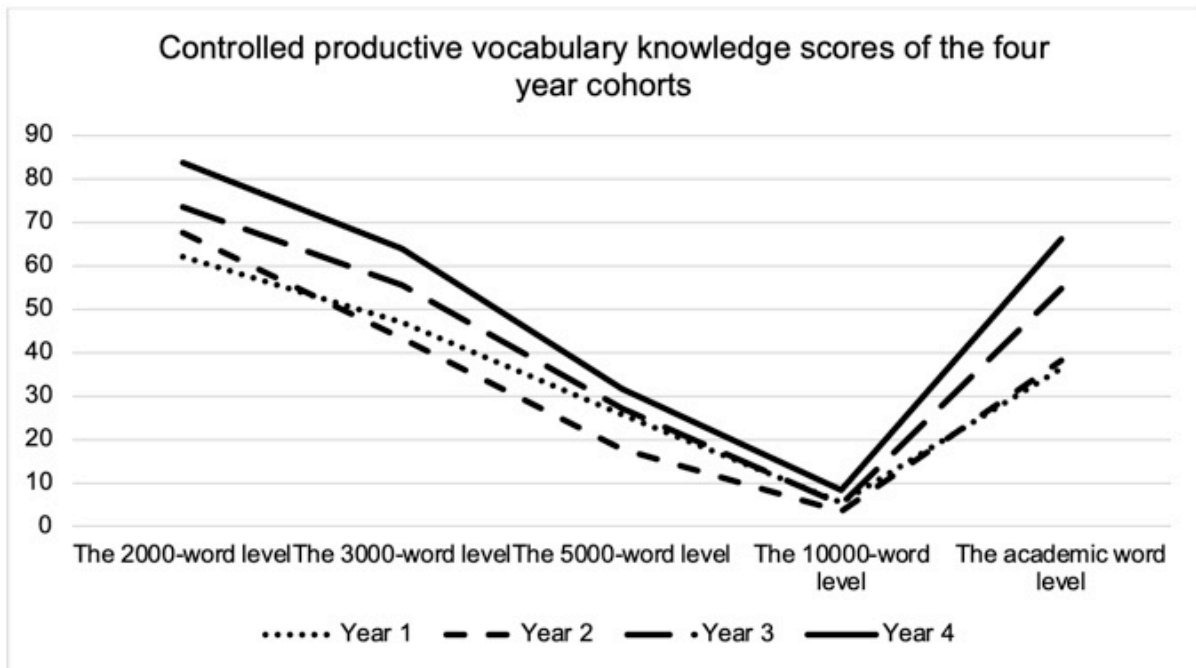
Figure 8.4: Controlled productive vocabulary knowledge scores of the four year cohorts at the five different frequency levels

According to the data, the participants in the later years tended to have larger controlled productive vocabulary knowledge at all the five frequency levels, with the largest gain at the academic word level and the smallest gain at the 10,000-word level. Only the fourth-year participants master the controlled productive vocabulary at the 2,000-word level (83.75%). However, the first-year students had higher scores than the second-year students at the 3,000-, the 5,000- and the 10,000-word levels. I would explore and discuss this in Chapter 11.

I conducted MANOVA and One-Way ANOVA with Tukey post hoc test to check whether these differences were statistically significant. The findings are shown in Table 8.6 and the summary table of years with significant differences are shown in Table 8.7. The effect sizes are the values of partial eta squared of the tests of between-subjects effects.

235

Multiple Comparisons						
Post hoc test: Tukey HSD						
Frequency level	(A) Year of study	(B) Year of study	Mean Difference (A-B)	%	Sig.	Partial Eta Squared
2,000 words	1	2	-1.13	-5.65	0.518	0.22
		3	-2.31*	-11.56*	0.019	
		4	-4.34**	-21.69**	0.000	
	2	1	1.13	5.65	0.518	
		3	-1.182	-5.91	0.508	
		4	-3.21*	-16.04*	0.003	
	3	1	2.31*	11.56*	0.019	
		2	1.18	5.91	0.508	
		4	-2.03	-10.13	0.087	
	4	1	4.34*	21.69*	0.000	
		2	3.21*	16.04*	0.003	
		3	2.03	10.13	0.087	
3,000 words	1	2	0.75	3.73	0.885	0.14
		3	-1.73	-8.63	0.288	
		4	-3.38*	-16.90*	0.007	
	2	1	-0.75	-3.73	0.885	
		3	-2.47	-12.36	0.096	
		4	-4.13*	-20.63*	0.002	
	3	1	1.73	8.63	0.288	
		2	2.47	12.36	0.096	
		4	-1.65	-8.27	0.404	
	4	1	3.38*	16.90*	0.007	
		2	4.13*	20.63*	0.002	
		3	1.65	8.27	0.404	
5,000 words	1	2	1.61	8.03	0.185	0.09
		3	-0.30	-1.51	0.978	
		4	-1.19	-5.93	0.443	
	2	1	-1.61	-8.03	0.185	
		3	-1.91	-9.53	0.099	
		4	-2.79*	-13.96*	0.008	
	3	1	0.30	1.51	0.978	
		2	1.91	9.53	0.099	

		4	-0.89	-4.43	0.703	
	4	1	1.19	5.93	0.443	
		2	2.79*	13.96*	0.008	
		3	0.89	4.43	0.703	
10,000 words	1	2	0.38	1.90	0.843	0.04
		3	0.05	0.27	0.999	
		4	-0.58	-2.89	0.594	
	2	1	-0.38	-1.90	0.843	
		3	-0.33	-1.63	0.903	
		4	-0.96	-4.79	0.226	
	3	1	-0.05	-0.27	0.999	
		2	0.33	1.63	0.903	
		4	-0.63	-3.16	0.549	
	4	1	0.58	2.89	0.594	
		2	0.96	4.79	0.226	
		3	0.63	3.16	0.549	
academic words	1	2	-0.36	-1.80	0.988	0.26
		3	-3.70*	-18.50*	0.004	
		4	-5.99**	-29.93**	0.000	
	2	1	0.36	1.80	0.988	
		3	-3.34*	-16.70*	0.024	
		4	-5.63**	-28.13**	0.000	
	3	1	3.70*	18.50*	0.004	
		2	3.34*	16.70*	0.024	
		4	-2.28	-11.42	0.203	
	4	1	5.99**	29.93**	0.000	
		2	5.63**	28.13**	0.000	
		3	2.28	11.42	0.203	

Note: * = statistically significant at the 0.05 level, ** = statistically significant at the 0.001 level

Table 8.6: ANOVA analysis of controlled productive vocabulary knowledge scores among the four year cohorts

The 2,000-word level	MD (%)	The 3,000-word level	MD (%)	The 5,000-word level	MD (%)	The 10,000-word level	MD (%)	The academic word level	MD (%)
Y2 > Y1	5.65	-	-	-	-	-	-	-	-
Y3 > Y1	11.56	-	-	-	-	-	-	Y3 > Y1	18.50
-	-	-	-	-	-	-	-	Y3 > Y2	16.70
Y4 > Y1	21.69	Y4 > Y1	16.90	-	-	-	-	Y4 > Y1	29.93
Y4 > Y2	16.04	Y4 > Y2	20.63	Y4 > Y2	13.96	-	-	Y4 > Y2	28.13

Note: Y = years of study

Table 8.7: Years with significant differences in controlled productive vocabulary knowledge scores at the five frequency levels

The ANOVA analysis showed that there were statistically significant differences in controlled productive vocabulary knowledge scores at all the five frequency levels, except at the 10,000-word level. Most of the differences were between the fourth-year participants and the other year groups.

There were significant differences in controlled productive vocabulary knowledge scores between the two continuous years: Year 2 and Year 1 at the 2,000-word level and Year 3 and Year 2 at the academic word level. The biggest differences were between the fourth-year students and the first-year and the second-year students at the academic word level (29.93% and 28.13%, respectively).

The effect sizes of these differences at the 2,000- and the academic word levels exceeded Cohen's (1988) convention of a small effect.

In conclusion, the participants' controlled productive vocabulary decreased when the frequency levels became lower. The participants at higher years of study tended to have larger controlled productive vocabulary, and they seemed to have learned a lot of academic words throughout their years of study. However, the first-year students had higher scores than the second-year students at some frequency levels (at the 3,000-, 5,000-, and 10,000-word levels), but these were not statistically significant.

Chapter 9 Findings 3

Free productive vocabulary knowledge

9.1   Introduction

Chapter 9 presents the findings regarding the participants' free productive knowledge in relation to different word frequency levels and different years of study. Free productive vocabulary knowledge in this study refers to the knowledge of how to recall and use lemmas at each of the five frequency levels. The third research question and its sub-questions are shown below.

9.2   Research question 3: What is the participants' free productive vocabulary knowledge?

Research question 3: What is the participants' free productive vocabulary knowledge?

>   **Sub-question 1:** What is the participants' free productive vocabulary knowledge at different word frequency levels?

>   **Sub-question 2:** What is the participants' free productive vocabulary knowledge at different years of study?

9.2.1   Sub-question 1: What is the participants' free productive vocabulary knowledge at different word frequency levels?

This section presents the findings regarding the participants' free productive vocabulary knowledge at the five different frequency levels: the 2,000-word level, the 3,000-word level, the 5,000-word level, the 10,000-word level, and the academic word level.

The findings in this section are different from the findings of the participants' receptive and controlled productive vocabulary knowledge in two ways. First, there

were not the target lemmas in the free productive vocabulary test – argumentative essays. Second, there was not maximum scores (possible scores) at each frequency level. That is, all the lemmas in the test were supplied by the participants, and the findings did not cover the lemmas that were not supplied in the participants' free writings. There might be other lemmas known well by the participants but were not included in their essays. Hence, I present only what the participants provided in the test.

The number of the lemmas produced was multiplied by two points and used as the total scores of the free productive vocabulary knowledge in this section.

Group	N	Tokens		Lemmas	
		Mean	SD	Mean	SD
Year 1	34	312.41	56.21	128.68	20.80
Year 2	24	349.00	57.08	136.29	20.39
Year 3	29	373.03	85.83	152.55	23.28
Year 4	24	356.08	77.21	151.83	20.97
Lower year	58	327.55	58.94	131.83	20.80
Higher year	53	365.36	81.71	152.23	22.06
All the participants	111	345.60	72.91	141.57	23.64

Table 9.1: Number of running words and lemmas produced per essay by the participants

The word limit per essay for the free writing task was 300 running words and the participants on average produced 345.60 running words or 141.57 lemmas, with the students in the later years producing more running words or lemmas than those in earlier years of study. However, the students from Year 3 produced slightly more running words or lemmas than the fourth-year students.

To answer the second research question, I investigated the participants' free productive vocabulary knowledge at the five different frequency levels. As mentioned in data analysing process, for the consistency with the receptive and the controlled productive tests in this study, I classified the lemmas into five frequency-based groups by comparing the lemmas against the word lists used in the word sampling process of Schmitt et al.'s (2001) VLT, which are the GSL (West, 1953) for the most 2,000 frequent words, the AWL (Coxhead, 2000) for the academic words, and the

Teacher's word book of 30,000 words (Thorndike and Lorge, 1944) for the 2,000-, the 3,000-, the 5,000-, and the 10,000-word levels. I present the changes in number of the different lemmas when I classified them with different word lists in Table 9.2 and Table 9.3. I used the number of the different lemmas produced by the first participant as an example.

Group of lemmas produced by the first participant	When comparing against two word lists	Number of different lemmas	%
The 2,000-word level	In the GSL, with both 'AA' and 'A' occurrences in Thorndike and Lorge's (1944) counts, not in the AWL	106	66.67
The 3,000-, 5,000-, 10,000-, and other word levels	Not in the GSL or the AWL	44	27.67
The academic word level	in the AWL	9	5.66
Number of different lemmas produced by the first participant	**including function words and mistakes**	**Total = 159**	**Total = 100.00**

Table 9.2: The number of lemmas (produced by the first participant as an example) classified by using only the GSL and the AWL

Group of lemmas produced by the first participant	When compared against three word lists used for the VLT	Number of different lemmas	%
2,000 words	in the GSL, with only 'A' occurrence in Thorndike and Lorge's (1944) count, not in the AWL	2	1.26
3,000 words	not in the GSL/AWL, 30-49 occurrences in Thorndike and Lorge's (1944) count	6	3.77
5,000 words	not in the GSL/AWL, 14-18 occurrences in Thorndike and Lorge's (1944) count	5	3.14
10,000 words	not in the GSL/AWL, 3 occurrences in Thorndike and Lorge's (1944) count	0	0.00
academic words	in the AWL	9	5.66
The rest of the lemmas (not classified or graded in this study)	including function words in the essay and mistakes produced by the software, not in the five lists above	137	86.16
Number of different lemmas produced by the first participant	**including function words and mistakes**	**Total = 159**	**Total = 100.00**

Table 9.3: The number of lemmas (produced by the first participant as an example) classified by using all the three word lists used in the VLT

When I classified the lemmas produced by this participant by using only the GSL and the AWL, most of the lemmas (66.67%) fell into the 2,000-word level as shown in Table 9.2. However, when I classified them by using the three word lists, for the consistency with the receptive and controlled productive tests, I found that the number of the lemmas produced by the participant and falling into the 2,000-word level had dropped drastically from 66.67% to only 1.26% as shown in Table 9.3 because most of the lemmas in the GSL have 'AA' occurrence in Thorndike and Lorge (1944). 'AA' occurrence refers to "100 [occurrences] or over per million [tokens in the count]" (Thorndike and Lorge, 1944, p. ix). Some of the lemmas in the GSL has 'A' occurrence. Only those with 'A' occurrence were sampled in the VLT. Therefore, I sorted only those with 'A' occurrence, for consistency of the analysis. This made the number of lemmas at the 2,000-word level in the analysis of free

productive vocabulary in my study would then be much lower, from 106 lemmas (Table 9.2) and to 2 lemmas (Table 9.3).

In spite of this, I decided to classify the lemmas produced by the participants by using the three word lists for the consistency of all the three sets of data. The lemmas I checked and reported were all content words that fell into the categories shown in Table 9.3 only.

For the first participant (see Table 9.3), only two lemmas at the 2,000-word level, six lemmas at the 3,000-word level, five lemmas at the 5,000-word level, none at the 10,000-word level and nine lemmas at the academic word level were graded.

The possible total score of each essay was two points multiplied by the number of the produced lemmas per essay (two points for a lemma used with correct meaning and grammar and one point for a lemma used with correct meaning. Each participant's total possible score was different because they supplied different number of lemmas per essay. Here are the formulas I used to calculate the percentage of score of the lemmas at each frequency level (Equation 9.1 – Equation 9.5).

$$\text{Percentage of 2K FP score per essay} = \frac{\text{2K FP score x 100}}{\text{Number of all lemmas in the essay x 2}}$$

Equation 9.1: Percentage score of the produced lemmas per essay at the 2,000-word level

Note: 2K = at the 2,000-word level, FP = (produced) free productive vocabulary

$$\text{Percentage of 3K FP score per essay} = \frac{\text{3K FP score x 100}}{\text{Number of all lemmas in the essay x 2}}$$

Equation 9.2: Percentage score of the produced lemmas per essay at the 3,000-word level

Note: 3K = at the 3,000-word level, FP = (produced) free productive vocabulary

Percentage of 5K FP score per essay =
$$\frac{\text{5K FP score x 100}}{\text{Number of all lemmas in the essay x 2}}$$

Equation 9.3: Percentage score of the produced lemmas per essay at the 5,000-word level

Note: 5K = at the 5,000-word level, FP = (produced) free productive vocabulary

Percentage of 10K FP score per essay =
$$\frac{\text{10K FP score x 100}}{\text{Number of all lemmas in the essay x 2}}$$

Equation 9.4: Percentage score of the produced lemmas per essay at the 10,000-word level

Note: 5K = at the 10,000-word level, FP = (produced) free productive vocabulary

Percentage of AWL FP score per essay =
$$\frac{\text{AWL FP score x 100}}{\text{Number of all lemmas in the essay x 2}}$$

Equation 9.5: Percentage score of the produced lemmas per essay at the academic word level

Note: AWL = at the academic word level (based on Coxhead's (2000) AWL), FP = (produced) free productive vocabulary

To answer the first sub-question of the third research question, I analysed the lemma scores of each of the five frequency levels by using descriptive statistics. The findings are shown in Table 9.4 and Figure 9.1.

As shown in Table 9.4, the participants on average produced 141.57 lemmas per essay, so the possible total scores for the lemma use was approximately 283.40 per essay (the number of all the lemmas per essay x 2 points).

Frequency level	Average number of produced lemmas per essay	%	Average score of the produced lemmas per essay	%
2,000 words	5.96	4.14	10.25	3.55
3,000 words	3.86	2.69	7.17	2.48
5,000 words	1.46	0.98	2.65	0.89
10,000 words	0.35	0.24	0.55	0.18
academic words	5.43	3.70	9.28	3.15
Not classified into these five levels*	124.50	88.24	249.00 (Not graded the correctness of usage)	89.75 (Not graded the correctness of usage)
Total	**141.57**	**100.00**	**283.14**	**100.00**

Note:
1) * = function words and words outside the five bands and the AWL
2) Each participant's total possible score was different because they supplied different number of lemmas per essay.

Table 9.4: Number and scores of the produced lemmas per essay at the five frequency levels
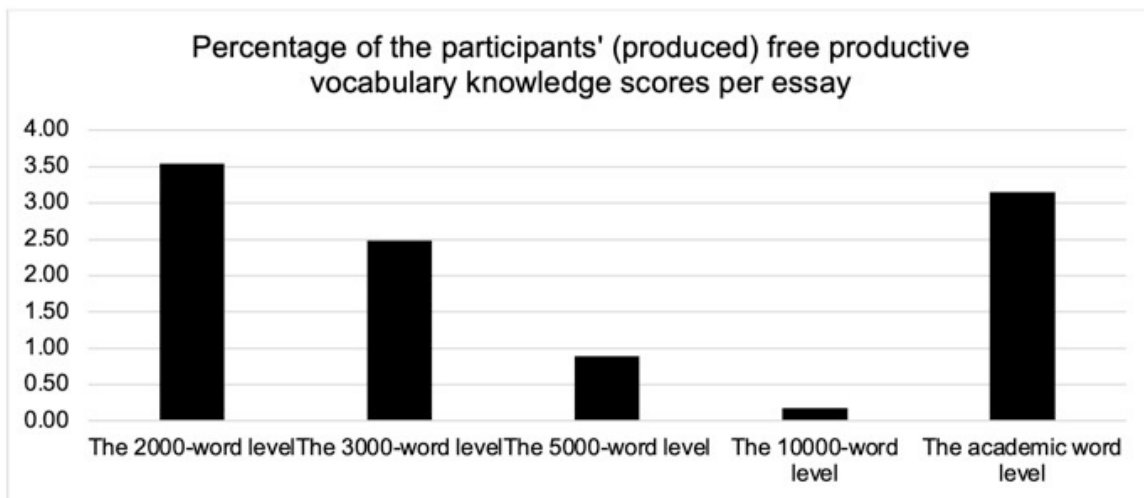
Figure 9.1: The participants' (produced) free productive vocabulary knowledge scores at the five frequency levels per essay

When calculating the number and the scores into percentage, I found that on average the participants produced the most lemmas at the 2,000-word level (around 41 lemmas per 1,000 produced lemmas). The number was roughly equal to that at the academic word level (37 lemmas per 1,000 produced lemmas). The participants

produced fewer lemmas at lower-frequency levels, or approximately 27 lemmas at the 3,000-word level. They produced ten lemmas per 1,000 produced lemmas at the 5,000-word level and only two lemmas per 1,000 produced lemmas at the 10,000-word level.

In terms of scores, the participants on average had the highest scores at the 2,000-word level (3.55%). Similar to that at the 2,000-word level, the average score of the produced lemmas per essay at the academic word level was 3.15%. Their free productive scores at the 3,000-, the 5,000-, and the 10,000-word levels were relatively low. 2.48%, 0.89%, and 0.18%.

When I compared the percentages in Table 9.4, I found that, out of the lemmas they produced, only a few lemmas were used incorrectly, compared to those used correctly. This supports the view that the students had receptive knowledge of the lemmas before they could recall and use them in their free writings, and that the students preferred to use the lemmas they were certain about using them in sentences.

Regarding the mastery level, these figures cannot be used to show whether the students mastered the vocabulary at each frequency level or not because the number of the lemmas that were well-known but did not produce in their free writings was not known.

I ran a one-way repeated measures ANOVA to test statistical significance of the findings, based on the percentage scores of the lemmas written in the free productive test. The p-value of less than 0.001 from the tests of within-subjects effects and the tests of within-subjects contrasts showed that there was very low probability that the significant differences of the free productive vocabulary scores at the five frequency levels were through random error. In terms of pairwise comparisons, the free productive scores at all the frequency levels were significantly different, except those between the 2,000-word level and the academic word level (the p-value of 0.713).

This can be interpreted that the free productive scores among the frequency levels were statistically significant, or the frequency levels affected the controlled productive scores. The differences in scores at the 2,000-word level and the academic word level were not statistically significant, though.

9.2.2 Sub-question 2: What is the participants' free productive vocabulary knowledge at different years of study?

In this section, I investigated the participants' free productive vocabulary knowledge at different years of study by comparing the percentage scores of the produced lemmas between groups of the participants (between the proficiency groups and among the four year cohorts).

Comparison 1: Free productive vocabulary knowledge between the higher and the lower year groups

A normality test was conducted. The p-values of the Shapiro-Wilk test were above 0.05 (0.058 for the lower year group and 0.529 for the higher year group). We can assume that the data were approximately normally distributed. To check the assumption of homogeneity of variances, the Levene's test was conducted. The p-value (0.279) was greater 0.05, so the assumption of equality of variances was met.

I compared the percentage score of the lemmas produced by the higher year group to that of the lemmas produced by the lower year group by conducting an independent samples t-test. The percentages of the scores were calculated out of the total possible lemma scores per essay if they used all the lemmas (I sorted) correctly. The statistics are shown in Table 9.5.

Year group	N	Score of the lemmas per essay	%	SD	%	MD	%	Sig. (2-tailed)	Effect size
Higher	53	36.81	11.92	13.52	3.56	13.23**	3.20**	0.000	1.05
Lower	58	23.59	8.72	11.46	3.56				

Note: ** = statistically significant at the 0.001 level

Table 9.5: Free productive vocabulary knowledge scores between the higher and the lower year groups

The higher year group significantly had higher score of the lemmas produced and used in their free writings than the lower year group (MD = 13.23 scores, 3.20%) with the p value of less than 0.001. The effect size of this t-test (1.05) was found to exceed Cohen's (1988) convention of a large effect size (0.80).

I also compared the scores of the lemmas (at each of the five frequency levels) produced and used by these two groups to examine how their free productive vocabulary knowledge changes at different frequency levels. I conducted an

independent samples t-test and calculated the effect sizes. The percentages of the scores were calculated out of the possible total lemma scores per essay (the number of the lemmas produced per essay multiplied by two points). The statistics are shown in Table 9.6.

Frequency levels	Year group	N	Mean	%	SD	%	MD	%	Sig. (2-tailed)	Effect size
2,000 words	Higher	53	12.08	3.93	6.12	1.90	3.49*	0.73*	0.001	0.64
	Lower	58	8.59	3.20	4.77	1.60				
3,000 words	Higher	53	8.72	2.84	4.02	1.20	2.96**	0.69**	0.000	0.81
	Lower	58	5.76	2.15	3.29	1.14				
5,000 words	Higher	53	3.28	1.05	2.71	0.81	1.21*	0.30*	0.011	0.48
	Lower	58	2.07	0.75	2.26	0.79				
10,000 words	Higher	53	0.70	0.22	1.07	0.33	0.28	0.06	0.120	0.30
	Lower	58	0.41	0.15	0.84	0.31				
academic words	Higher	53	12.04	3.88	6.37	1.83	5.28**	1.41**	0.000	0.94
	Lower	58	6.76	2.47	5.21	1.79				

Note: * = statistically significant at the 0.05 level, ** = statistically significant at the 0.001 level

Table 9.6: Free productive vocabulary knowledge scores between the higher and the lower year groups at the five frequency levels

Percentages of free productive vocabulary knowledge scores between the higher and the lower year groups
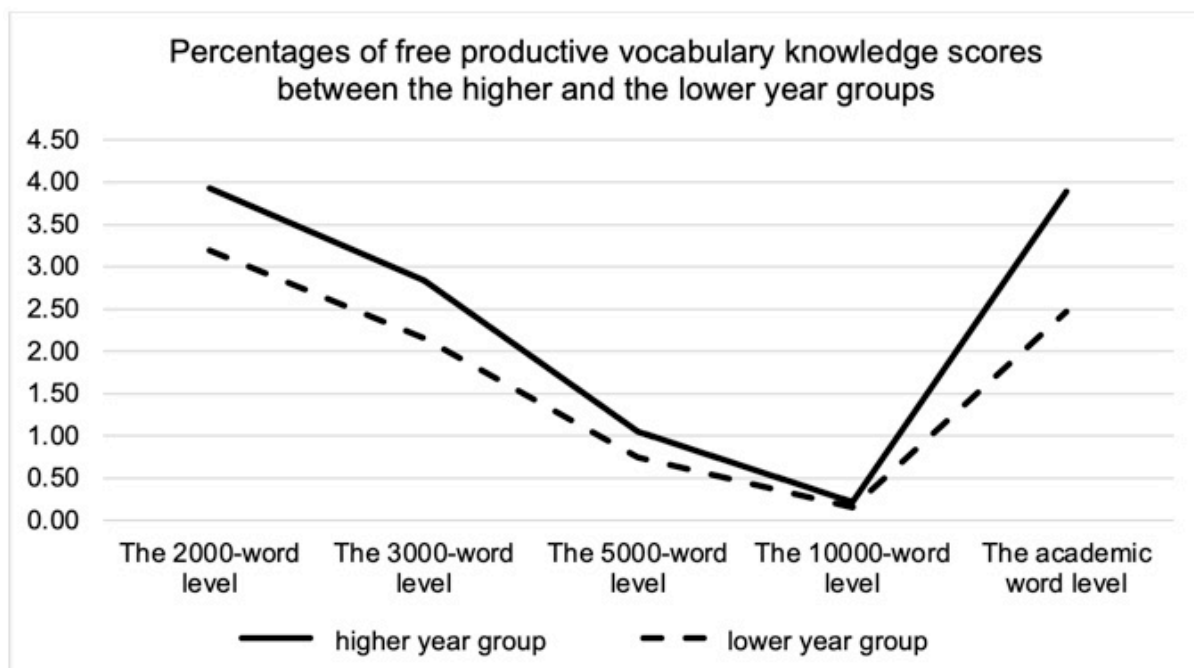
Figure 9.2: Percentages of the free productive vocabulary knowledge scores between the higher and the lower year groups at the five frequency levels

The findings showed that the higher year group had higher percentage scores for free productive vocabulary knowledge than the lower year group at all the five frequency levels. The differences were statistically significant, except at the 10,000-word level.

Their scores at the 2,000-, the 3,000- and the academic word levels were highly significantly different with the p values of less than 0.001 and medium to large effect sizes (0.64, 0.81, and 0.94) (Cohen, 1988). The difference at the 5,000-word level was also significant ($p < 0.05$) but with the effect size of 0.48.

The largest difference was at the academic word level (MD = 1.41%) but their scores at the 10,000-word level were only slightly different (MD = 0.06%).

Comparison 2: Free productive vocabulary knowledge at different years of study

A normality test was conducted. The p-values of the Shapiro-Wilk test were above 0.05 (0.051, 0.723, 0.931, and 0.063 for the first-year group to the fourth-year group). We can assume that the data were approximately normally distributed. To check the assumption of homogeneity of variances, the Levene's test was conducted. The p-value (0.918) was greater 0.05, so the variances among groups were not significantly different and the assumption of equality of variance was met.

When I examined their free productive vocabulary knowledge at different years of study, I again found that their free productive vocabulary knowledge was high for the participants at higher years of study. However, the percentage of free vocabulary use for the fourth-year students were slightly lower than that for the third-year students. This difference was not statistically significant, though.

Year of study	N	Mean	%	SD	%
1	34	20.12	7.61	10.92	3.45
2	24	28.50	10.29	10.56	3.15
3	29	37.97	12.22	14.47	3.61
4	24	35.42	11.56	12.42	3.54
Average	N/A	**29.90**	**10.25**	**14.09**	**3.89**

Table 9.7: Free productive vocabulary knowledge scores per essay of the four year cohorts
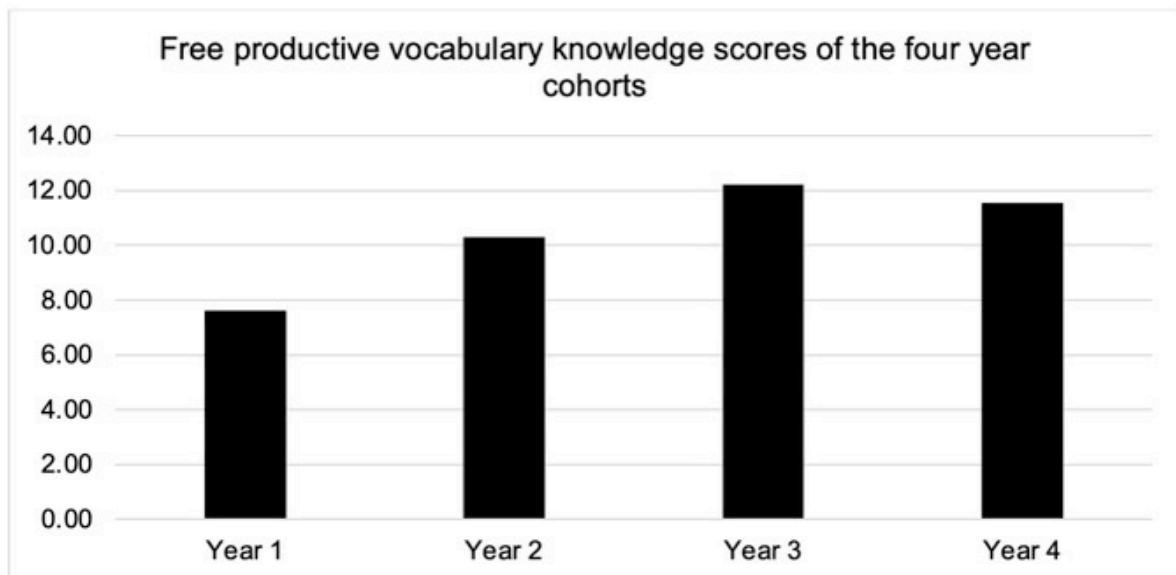
Figure 9.3: Free productive vocabulary knowledge scores per essay of the four year cohorts

I also compared the participants' free productive vocabulary knowledge scores at the five frequency levels and present the findings in Table 9.8 and Figure 9.4.

Frequency level	Year of study	N	Mean	%	SD	%
2,000 words	1	34	7.59	2.92	4.67	1.66
	2	24	10.00	3.59	4.63	1.46
	3	29	11.79	3.83	6.11	1.86

	4	24	12.42	4.05	6.23	1.98
	Average	**N/A**	**10.25**	**3.55**	**5.70**	**1.78**
3,000 words	1	34	4.59	1.75	3.04	1.04
	2	24	7.42	2.72	2.96	1.05
	3	29	9.45	3.06	3.82	1.09
	4	24	7.83	2.57	4.15	1.30
	Average	**N/A**	**7.17**	**2.48**	**3.93**	**1.22**
5,000 words	1	34	1.91	0.71	2.11	0.76
	2	24	2.29	0.80	2.49	0.85
	3	29	3.41	1.07	3.12	0.91
	4	24	3.13	1.02	2.15	0.68
	Average	**N/A**	**2.65**	**0.89**	**2.55**	**0.81**
10,000 words	1	34	0.32	0.13	0.73	0.31
	2	24	0.54	0.18	0.98	0.31
	3	29	0.72	0.22	1.07	0.34
	4	24	0.67	0.21	1.09	0.33
	Average	**N/A**	**0.55**	**0.18**	**0.96**	**0.32**
academic words	1	34	5.71	2.09	5.25	1.80
	2	24	8.25	3.01	4.88	1.66
	3	29	12.59	4.04	6.63	1.89
	4	24	11.38	3.70	6.13	1.79
	Average	**N/A**	**9.28**	**3.15**	**6.35**	**1.94**

Table 9.8: Free productive vocabulary knowledge scores of the four year cohorts at the five different frequency levels
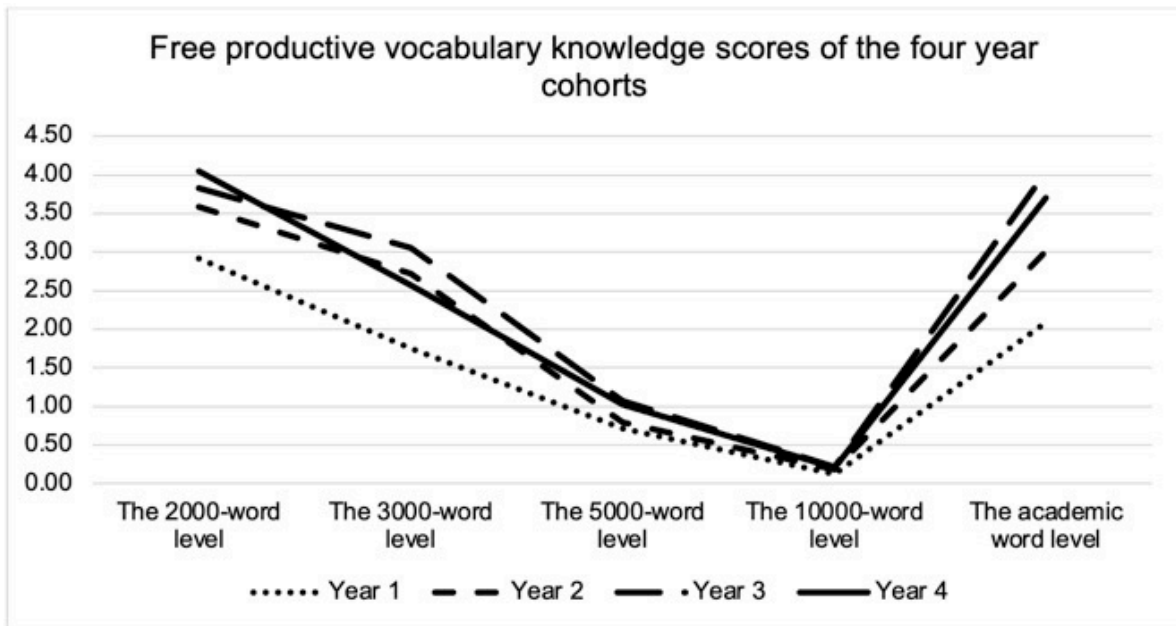
Figure 9.4: Free productive vocabulary knowledge scores of the four year cohorts at the five frequency levels

I found that the third-year participants had higher scores than the fourth-year students at all the frequency levels, except the 2,000-word level, and that the second-year group had a slightly higher score than the fourth-year cohort at the 3,000-word level. Of the four groups, the first-year students had the lowest scores.

However, to check whether these differences in their scores were statistically significant, I conducted MANOVA and One-Way ANOVA with Tukey post hoc test. The findings are shown in Table 9.9 and Table 9.10. The effect sizes are the values of partial eta squared of the tests of between-subjects effects.

Multiple Comparisons						
Post hoc test: Tukey HSD						
Dependent Variable	**(A) Year of Study**	**(B) Year of Study**	**Mean Difference (A-B)**	**%**	**Sig.**	**Partial Eta Squared**
2,000 words	1	2	-2.41	-0.67	0.346	0.10
		3	-4.21*	-0.90*	0.014	
		4	-4.83*	-1.13*	0.006	
	2	1	2.41	0.67	0.346	
		3	-1.79	-0.24	0.630	
		4	-2.42	-0.47	0.416	
	3	1	4.21*	0.90*	0.014	
		2	1.79	0.24	0.630	
		4	-0.62	-0.23	0.976	
	4	1	4.83*	1.13*	0.006	
		2	2.42	0.47	0.416	
		3	0.62	0.23	0.976	
3,000 words	1	2	-2.83*	-0.97*	0.016	0.18
		3	-4.86**	-1.31**	0.000	
		4	-3.25*	-0.82*	0.004	
	2	1	2.83*	0.97*	0.016	
		3	-2.03	-0.34	0.159	
		4	0.42	0.15	0.976	
	3	1	4.86**	1.31**	0.000	
		2	2.03	0.34	0.159	
		4	1.62	0.49	0.343	
	4	1	3.25*	0.82*	0.004	
		2	-0.42	-0.15	0.976	
		3	-1.62	-0.49	0.343	
5,000 words	1	2	-0.38	-0.09	0.941	0.07
		3	-1.50	-0.36	0.088	
		4	-1.21	-0.31	0.270	
	2	1	0.38	0.09	0.941	
		3	-1.12	-0.27	0.369	
		4	-0.83	-0.22	0.657	
	3	1	1.50	0.36	0.088	
		2	1.12	0.27	0.369	

		4	0.29	0.05	0.975	
	4	1	1.21	0.31	0.270	
		2	0.83	0.22	0.657	
		3	-0.29	-0.05	0.975	
10,000 words	1	2	-0.22	-0.04	0.829	0.02
		3	-0.40	-0.09	0.354	
		4	-0.34	-0.07	0.539	
	2	1	0.22	0.04	0.829	
		3	-0.18	-0.05	0.901	
		4	-0.13	-0.03	0.969	
	3	1	0.40	0.09	0.354	
		2	0.18	0.05	0.901	
		4	0.06	0.02	0.996	
	4	1	0.34	0.07	0.539	
		2	0.13	0.03	0.969	
		3	-0.06	-0.02	0.996	
academic words	1	2	-2.54	-0.91	0.352	0.16
		3	-6.88**	-1.94**	0.000	
		4	-5.67*	-1.60*	0.002	
	2	1	2.54	0.91	0.352	
		3	-4.34*	-1.03*	0.037	
		4	-3.13	-0.69	0.243	
	3	1	6.88**	1.94**	0.000	
		2	4.34*	1.03*	0.037	
		4	1.21	0.34	0.871	
	4	1	5.67*	1.60*	0.002	
		2	3.13	0.69	0.243	
		3	-1.21	-0.34	0.871	

Note: * = statistically significant at the 0.05 level, ** = statistically significant at the 0.001 level

Table 9.9: ANOVA analysis of free productive vocabulary knowledge scores among the four year cohorts

The 2,000-word level	MD %	The 3,000-word level	MD %	The 5,000-word level	MD %	The 10,000-word level	MD %	The academic word level	MD %
-	-	Y2 > Y1	0.97	-	-	-	-	-	-
Y3 > Y1	0.90	Y3 > Y1	1.31	-	-	-	-	Y3 > Y1	1.94
-	-	-	-	-	-	-	-	Y3 > Y2	1.03
Y4 > Y1	1.13	Y4 > Y1	0.82	-	-	-	-	Y4 > Y1	1.60
Note: Y = years of study									

Table 9.10: Years with significant differences in free productive vocabulary knowledge scores at the five frequency levels

The findings from the One-Way ANOVA analysis showed that there were statistically significant differences in free productive vocabulary knowledge scores between some year groups at the 2,000-, the 3,000-, and the academic word levels. Most of the differences were between the third or the fourth-year participants and the first-year participants. The biggest difference was between the third-year and the first-year students (1.94%) at the academic word level, but the smallest difference was between the third-year and the fourth-year students at the 10,000-word level (MD = 0.02%). Therefore, even though the differences in free productive vocabulary scores were statistically significant, the effect sizes did not exceed Cohen's (1988) convention of a small effect (0.20). This implied that these differences were trivial.

In conclusion, the participants at the higher years of study seemed to be able to produce and used more lemmas at all the five frequency levels, particularly the academic word level. However, the third-year and the second-year had used slightly more lemmas than the fourth-year students.

Chapter 10 Findings 4

Relationship among receptive, controlled productive, and free productive vocabulary knowledge

10.1 Introduction

This chapter presents the findings to the final research question. I present the findings in the order of the sub-research questions. The findings about the relationship among the participants' receptive, controlled productive, and free productive vocabulary knowledge were analysed in three ways. The first analysis was the quantitative analysis within the dichotomous approach (Section 10.2). I compared the findings of the receptive vocabulary with those of the productive vocabulary (both controlled and free productive vocabularies). The second analysis that I used was the quantitative analysis within the developmental approach (Section 10.3). This approach was explained by presenting the developmental scores as used in Paribakht and Wesche's (1997) study. It can show the shifts of vocabulary knowledge from one stage to another stage on a vocabulary knowledge continuum. The third analysis examines participants' answers qualitatively in terms of degrees of knowledge aspects. This was done to provide more insight into the relationship between receptive, controlled productive, and free productive vocabulary (Section 10.4).

10.2 Research question 4: What is the relationship among the participants' receptive, controlled productive, and free productive vocabulary knowledge? (Using the dichotomous approach)

Research question 4: What is the relationship among the participants' receptive, controlled productive, and free productive vocabulary knowledge?

> **Sub-question 1:** What is the relationship among the participants' receptive, controlled productive, and free productive vocabulary knowledge at different word frequency levels?

Sub-question 2: What is the relationship among the participants' receptive, controlled productive, and free productive vocabulary knowledge at different years of study?

10.2.1 Sub-question 1: What is the relationship among the participants' receptive, controlled productive, and free productive vocabulary knowledge at different word frequency levels?

In this section, I present the relationship among the participants' receptive, controlled productive, and free productive vocabulary knowledge by comparing the percentage scores from the three tests.

However, even though the word sampling for the three tests was consistent and the scoring system was the same (a three-scale scoring system), the limitation of this study is that the total scores of the three tests were not equal or did not refer to the same things. That is, the total possible score for the receptive vocabulary test at the 2,000-word level was 20 points (10 target lemmas and two points each for precise knowledge) of form and meaning. The maximum score at the 2,000-word level represents the most frequently used 2,000 lemmas that a learner can understand when encountering. The maximum score at the 3,000-, the 5,000-, and the 10,000-word levels represents the 1,000 lemmas at such level. That at the academic word level represents 570 lemmas of the AWL. The maximum scores of the controlled productive test represent the same but controlled productive vocabulary refers to lemmas that a learner can use correctly when forced to.

However, the free productive vocabulary scores in this study refer to only the percentage scores of the lemmas produced by the participants in their free writing tasks and falling into each of the five frequency levels that are focused in the present study. Therefore, the lemmas the participants were able to use in sentences but not produced by the participants were not included in the calculation. Also, the mastery level for the free productive vocabulary cannot be calculated.

These limitations meant that it was not possible to compare the participants' mean scores of these three kinds of vocabulary straightforwardly, so I adopted the following strategy. Only the percentage scores of the lemmas produced in the free productive vocabulary test were analysed and compared to the mean percentage scores of the receptive and the controlled productive vocabulary tests, as shown in Table 10.1.

Frequency level	Receptive vocabulary	%	Controlled productive vocabulary	%	(Produced) Free productive vocabulary per essay*	%
2,000 words	18.40	91.98	14.20	70.99	10.25	3.55
3,000 words	16.09	80.45	10.43	52.16	7.17	2.48
5,000 words	8.83	44.14	5.14	25.68	2.65	0.89
10,000 words	2.15	10.77	1.12	5.59	0.55	0.18
academic words	16.09	80.45	9.6	48.02	9.28	3.15

Note: The mean score and the percentage of free productive vocabulary were calculated based on the lemmas produced by each of the participants.

Table 10.1: Participants' receptive, controlled productive, and (produced) free productive vocabulary knowledge
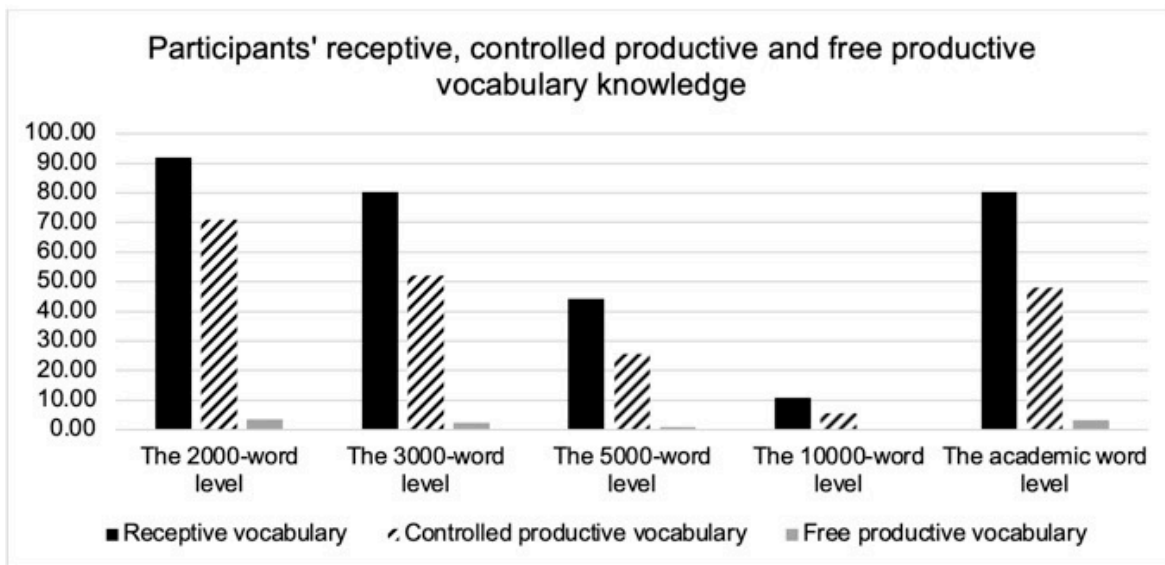
Note: the mean score and the percentage of free productive were calculated based on the lemmas produced by each of the participants.

Figure 10.1: Participants' receptive, controlled productive, and free productive vocabulary knowledge

Overall, the participants' receptive, controlled productive, and free productive vocabularies were consistently affected by frequency levels. The patterns of the three groups of vocabulary were similar. The percentage scores for the three groups were the highest at the 2,000-word level. They became lower when the frequency levels were lower. The percentage scores at the 10,000-word level were the lowest,

and those at the academic word level seemed to be similar to those at the 3,000-word level.

Of all the three vocabularies, the receptive vocabulary was the largest at all the five frequency levels. The controlled productive vocabulary was smaller than the receptive vocabulary, and the free productive vocabulary was the smallest.

The participants only mastered the receptive vocabulary at more frequency levels (the 2,000-, the 3,000-, and the academic word levels) than they did for controlled productive vocabulary (on average nearly mastered the 2,000-word level). The data on free productive vocabulary cannot be compared.

The correlations among the percentage scores of the receptive, controlled productive, and free productive vocabularies

Pearson correlation coefficient was conducted to check whether the percentage scores of the receptive, controlled productive, and free productive vocabularies were correlated. The correlations are presented in Table 10.2.

Frequency levels	Vocabulary	Correlation	CP (%)	FP (%)
The five levels	Re (%)	Pearson Correlation	0.680**	0.447**
		Sig. (2-tailed)	0.000	0.000
	CP (%)	Pearson Correlation	1	0.288*
		Sig. (2-tailed)		0.002
2,000 words	Re (%)	Pearson Correlation	0.535**	0.090
		Sig. (2-tailed)	0.000	0.347
	CP (%)	Pearson Correlation	1	0.083
		Sig. (2-tailed)		0.389
3,000 words	Re (%)	Pearson Correlation	0.496**	0.166
		Sig. (2-tailed)	0.000	0.081
	CP (%)	Pearson Correlation	1	0.105
		Sig. (2-tailed)		0.271
5,000 words	Re (%)	Pearson Correlation	0.628**	0.098
		Sig. (2-tailed)	0.000	0.308
	CP (%)	Pearson Correlation	1	0.090
		Sig. (2-tailed)		0.345
10,000 words	Re (%)	Pearson Correlation	0.526**	0.082
		Sig. (2-tailed)	0.000	0.392

			1	0.159
	CP (%)	Pearson Correlation	1	0.159
		Sig. (2-tailed)		0.095
academic words	Re (%)	Pearson Correlation	0.661**	0.421**
		Sig. (2-tailed)	0.000	0.000
	CP (%)	Pearson Correlation	1	0.350**
		Sig. (2-tailed)		0.000

Note: *** = significance at the 0.001 level, Re = receptive vocabulary, CP = controlled productive vocabulary, FP = free productive vocabulary, % = the percentage score

Table 10.2: Correlations among the percentage scores of the receptive, controlled productive, and free productive vocabularies were correlated

The total percentage scores of all the three types of vocabulary were significantly and positively correlated to one another. The strongest correlation (r = 0.68; p value = 0.000) was between the percentage scores of the receptive vocabulary and those of the controlled productive vocabulary. The weakest correlation (r = 0.288; p value = 0.002) was between the percentage scores of the controlled productive vocabulary.

When examining the correlation at each of the five frequency level, the percentage scores of the receptive vocabulary were moderately correlated with those of the controlled productive vocabulary at all the five levels, with the p-values of less than 0.001. The strongest correlation (r = 0.661, p value = 0.000) between them was at the academic word level.

However, the percentage scores of the free productive vocabulary positively but not significantly correlated with those of the other two at the first four frequency levels. Only at the at the academic word level, these three were significantly correlated. The percentage scores of the free productive vocabulary had a moderate correlation (r = 0.35, p-value = 0.000) with those of the receptive vocabulary but a weak correlation (r = 0.42, p-value = 0.000) with those of the controlled productive vocabulary.

The ratios among receptive, controlled productive, and free productive vocabularies

The proportions among them at the five frequency levels were also calculated as the three following ratios: the controlled productive/receptive ratio, the free productive/receptive ratio, and the free productive/controlled productive ratio.

Frequency level	Receptive vocabulary (%)	Controlled productive vocabulary (%)	Free productive vocabulary (%)	CP/Re ratio (%)	FP/Re ratio (%)	FP/CP ratio (%)
2,000 words	91.98	70.99	3.55	77.18	3.86	5.00
3,000 words	80.45	52.16	2.48	64.84	3.08	4.75
5,000 words	44.14	25.68	0.89	58.18	2.02	3.47
10,000 words	10.77	5.59	0.18	51.90	1.67	3.22
academic words	80.45	48.02	3.15	59.69	3.92	6.56

Note: The mean score and the percentage of free productive vocabulary were calculated based on the lemmas produced by each of the participants.

The abbreviation 'Re' refers to receptive, 'CP' refers to controlled productive, and 'FP' refers to free productive vocabulary.

Table 10.3: Ratios among receptive, controlled productive, and free productive vocabularies

All the ratios varied but tended to be high at high frequency levels and tended to decrease when the frequency became lower, with those at the academic word level similar to those of the high-frequency lemmas. The gaps between them were small at high frequency levels. The gap between the receptive and the controlled productive vocabularies was the smallest, but that between the receptive and the free productive vocabularies was the biggest.

If based on the idea that receptive vocabulary becomes productive vocabulary at some point, there was more chance of receptive lemmas becoming controlled productive vocabulary than becoming free productive vocabulary.

10.2.2 Sub-question 2: What is the relationship among the participants' receptive, controlled productive, and free productive vocabulary knowledge at different years of study?

Comparison 1: Receptive, controlled productive, and free productive vocabularies between the higher and the lower year groups

In this section, I compared the percentage scores of all the three vocabularies of the higher year group to those of the low-proficiency group.

The data shows that the higher year group had higher percentage scores for all the three vocabularies and their mean differences were all highly significant (p < 0.001) with the large effect size (exceeding 0.80), according to Cohen's (1988) convention.

The mean scores for receptive vocabulary of these two groups were higher than those for the controlled productive and free productive vocabulary. However, the mean difference for the controlled productive vocabulary between the two groups was the largest (11.89%), followed by that for the receptive vocabulary (9.69%) and then that for the free productive vocabulary (3.20%).

Vocabulary	Year group	N	Mean score (Out of 100)	SD	MD	Sig. (2-tailed)	Effect size
Receptive	Higher	53	66.62	7.52	9.69**	0.000	1.12
	Lower	58	56.93	9.62			
Controlled productive	Higher	53	46.70	11.48	11.89**	0.000	0.91
	Lower	58	34.81	14.56			
Free productive	Higher	53	11.92	3.56	3.20**	0.000	1.05
	Lower	58	8.72	3.56			
Note: The mean score and the percentage of free productive were calculated based on the lemmas produced by each of the participants., * = statistically significant at the 0.05 level, ** = statistically significant at the 0.001 level							

Table 10.4: Receptive, controlled productive, and free productive vocabularies between the higher and the lower year groups

The higher year group had higher scores for all the three types of vocabulary, especially the controlled productive vocabulary.

I also compared the mean scores of these two proficiency groups to examine how their receptive, controlled productive, and free productive vocabulary knowledge change at different frequency levels (Table 10.5).

Frequency level	Vocabulary	Year group	N	Mean (%)	SD (%)	MD (%)	Sig. (2-tailed)	Effect size
2,000 words	Receptive	Higher	53	96.04	6.07	7.76**	0.000	1.04
		Lower	58	88.28	8.61			
		Higher	53	78.21	15.45	13.81**	0.000	0.87

262

	Controlled productive	Lower	58	64.40	16.12			
	Free productive	Higher	53	3.93	1.90	0.73*	0.001	0.64
		Lower	58	3.20	1.60			
3,000 words	Receptive	Higher	53	84.72	9.38	8.17**	0.000	0.79
		Lower	58	76.55	11.29			
	Controlled productive	Higher	53	59.43	17.37	13.92**	0.000	0.73
		Lower	58	45.52	20.83			
	Free productive	Higher	53	2.84	1.20	0.69**	0.000	0.81
		Lower	58	2.15	1.14			
5,000 words	Receptive	Higher	53	49.72	15.70	10.67*	0.001	0.65
		Lower	58	39.05	17.05			
	Controlled productive	Higher	53	29.25	13.95	6.83*	0.019	0.46
		Lower	58	22.41	16.04			
	Free productive	Higher	53	1.05	0.81	0.30*	0.011	0.48
		Lower	58	0.75	0.79			
10,000 words	Receptive	Higher	53	13.87	9.02	5.94*	0.003	0.59
		Lower	58	7.93	11.04			
	Controlled productive	Higher	53	6.60	6.99	1.95	0.239	0.23
		Lower	58	4.66	9.95			
	Free productive	Higher	53	0.22	0.33	0.06	0.120	0.30
		Lower	58	0.15	0.31			
academic words	Receptive	Higher	53	88.77	11.18	15.93**	0.000	1.18
		Lower	58	72.84	15.51			
	Controlled productive	Higher	53	60.00	19.27	22.93**	0.000	1.09
		Lower	58	37.07	22.69			
	Free productive	Higher	53	3.88	1.83	1.41**	0.000	0.94
		Lower	58	2.47	1.79			

Note: * = statistically significant at the 0.05 level, ** = statistically significant at the 0.001 level

Table 10.5: Receptive, controlled productive, free productive vocabulary knowledge scores between the higher and the lower year groups at the five frequency levels

When I examined the differences in percentage scores of the receptive, controlled productive, and free productive vocabulary knowledge between the higher and the lower year groups, I found that all the percentage scores for the three vocabularies

of the higher year group were significantly higher than the low-proficiency groups at all every frequency level.

The biggest difference of the receptive vocabulary between the higher and the lower year groups was at the academic word level (MD = 15.93%). The largest gap of the controlled productive vocabulary between these two groups of participants was at the academic word level (MD = 22.93%), similar to the free productive vocabulary (MD = 1.41%).

The biggest differences for the controlled productive vocabulary were at the 2,000-, the 3,000-word levels and the academic words, and for the receptive vocabulary at the 5,000- and the 10,000-word levels.

In terms of mastery, only the higher year group could reach the mastery level for their receptive vocabulary at the 2,000-, the 3,000-, and the academic word levels. They did not reach the mastery of the productive vocabularies at all.

The ratios among receptive, controlled productive, and free productive vocabularies between the higher and the lower year groups

I analysed the rations among receptive, controlled productive, and free productive vocabularies between the higher and the lower year groups by using the percentages from Table 10.5. The ratios are presented in Table 10.6.

Frequency level	Proficiency group	CP/Re ratio (%)	FP/Re ratio (%)	FP/CP ratio (%)
2,000 words	High	81.43	4.09	5.02
	Low	72.95	3.62	4.97
3,000 words	High	70.15	3.35	4.78
	Low	59.46	2.81	4.72
5,000 words	High	58.83	2.11	3.59
	Low	57.39	1.92	3.35
10,000 words	High	47.58	1.59	3.33
	Low	58.76	1.89	3.22
academic words	High	67.59	4.37	6.47
	Low	50.89	3.39	6.66

Note:

1) The abbreviation 'Re' refers to receptive, 'CP' refers to controlled productive, and 'FP' refers to free productive vocabulary.

2) The percentage score of free productive vocabulary were calculated based on the lemmas produced by each of the participants.

Table 10.6: Ratios among receptive, controlled productive, and free productive vocabularies between the higher and the lower year groups

The controlled productive/receptive ratios of the higher year group tended to be higher than those of the lower year group at all the five frequency levels, except the 10,000-word level. This implied that there was more chance of learners' receptive vocabulary becoming controlled productive vocabulary for the higher year group; the gap between their receptive and controlled productive became small. However, even though the higher year group had a larger receptive vocabulary than the low-proficiency group, not much of their receptive vocabulary at low-frequency level became controlled productive vocabulary. I discuss this in more detail in the discussion chapter (Chapter 11).

Both the free productive/receptive ratios and the free productive/controlled productive ratios were relatively low at all the five levels. The ratios of the higher year group were only slightly higher than those of the low-proficiency group. The differences did not seem to be noticeable.

Comparison 2: Receptive, controlled productive, and free productive vocabularies at different years of study

Vocabulary	Year group	N	Mean	%	SD	%
Receptive	1	34	55.88	55.88	10.66	10.66
	2	24	58.42	58.42	7.91	7.91
	3	29	64.10	64.10	7.98	7.98
	4	24	69.67	69.67	5.72	5.72
Controlled productive	1	34	35.32	35.32	15.48	15.48
	2	24	34.08	34.08	13.45	13.45
	3	29	43.31	43.31	12.10	12.10
	4	24	50.79	50.79	9.37	9.37
Free productive	1	34	20.12	7.61	10.92	3.45
	2	24	28.50	10.29	10.56	3.15
	3	29	37.97	12.22	14.47	3.61
	4	24	35.42	11.56	12.42	3.54

Note: the mean score and the percentage of free productive were calculated based on the lemmas produced by each of the participants.

Table 10.7: Receptive, controlled productive, free productive vocabulary knowledge scores of the four year cohorts

The data showed that the participants in later years of study had higher scores for all the three groups of vocabulary, with the receptive vocabulary scores always the highest. However, the first-year students had slightly higher controlled productive scores than the second-year students. This is discussed in the discussion chapter.

Frequency level	Year of study	N	Receptive vocabulary	%	Controlled productive vocabulary	%	Free productive vocabulary	%
2,000 words	1	34	17.32	86.62	12.41	62.06	7.59	2.92
	2	24	18.13	90.63	13.54	67.71	10.00	3.59
	3	29	18.93	94.66	14.72	73.62	11.79	3.83
	4	24	19.54	97.71	16.75	83.75	12.42	4.05
	Average	N/A	**18.40**	**91.98**	**14.20**	**70.99**	**10.25**	**3.55**
3,000 words	1	34	14.97	74.85	9.41	47.06	4.59	1.75
	2	24	15.79	78.96	8.67	43.33	7.42	2.72
	3	29	16.45	82.24	11.14	55.69	9.45	3.06

	4	24	17.54	87.71	12.79	63.96	7.83	2.57
	Average	**N/A**	**16.09**	**80.45**	**10.43**	**52.16**	**7.17**	**2.48**
5,000 words	1	34	8.09	40.44	5.15	25.74	1.91	0.71
	2	24	7.42	37.08	3.54	17.71	2.29	0.80
	3	29	8.93	44.66	5.45	27.24	3.41	1.07
	4	24	11.17	55.83	6.33	31.67	3.13	1.02
	Average	**N/A**	**8.83**	**44.14**	**5.14**	**25.68**	**2.65**	**0.89**
10,000 words	1	34	1.74	8.68	1.09	5.44	0.32	0.13
	2	24	1.38	6.88	0.71	3.54	0.54	0.18
	3	29	2.62	13.10	1.03	5.17	0.72	0.22
	4	24	2.96	14.79	1.67	8.33	0.67	0.21
	Average	**N/A**	**2.15**	**10.77**	**1.12**	**5.59**	**0.55**	**0.18**
academic words	1	34	13.76	68.82	7.26	36.32	5.71	2.09
	2	24	15.71	78.54	7.63	38.13	8.25	3.01
	3	29	17.17	85.86	10.97	54.83	12.59	4.04
	4	24	18.46	92.29	13.25	66.25	11.38	3.70
	Average	**N/A**	**16.09**	**80.45**	**9.60**	**48.02**	**9.28**	**3.15**

Note: The mean score and the percentage of free productive were calculated based on the lemmas produced by each of the participants.

Table 10.8: Receptive, controlled productive, and free productive vocabulary knowledge scores of the four year cohorts at the five different frequency levels

At every frequency level, the participants' receptive, controlled productive, and free productive vocabularies tended to be larger when the participants study in later years of study. However, they could be low for some years of study as well. For example, the receptive and controlled productive scores of the first-year students were slightly higher than those of the second-year group, and the free productive score of the third-year participants at the 3,000-word level was slightly higher than that of the fourth-year students.

Their receptive vocabulary was larger than the controlled productive vocabulary and the controlled productive vocabulary was larger than the free productive vocabulary at all the five levels and for all the year groups. I calculated the ratios and presented them in Table 10.9.

The ratios among receptive, controlled productive, and free productive vocabularies among the four years cohorts

The ratios among receptive, controlled productive, and free productive vocabularies among the four years cohorts are presented in Table 10.9. They were calculated based on the percentages of the mean scores for the three groups of vocabulary shown in Table 10.8.

Frequency level	Year of study	CP/Re (%)	FP/Re ratios (%)	FP/CP ratios (%)
2,000 words	1	71.65	3.37	4.71
	2	74.71	3.96	5.30
	3	77.77	4.05	5.20
	4	85.71	4.14	4.84
3,000 words	1	62.87	2.34	3.72
	2	54.88	3.44	6.28
	3	67.72	3.72	5.49
	4	72.92	2.93	4.02
5,000 words	1	63.65	1.76	2.76
	2	47.76	2.16	4.52
	3	60.99	2.40	3.93
	4	56.73	1.83	3.22
10,000 words	1	62.67	1.50	2.39
	2	51.45	2.62	5.08
	3	39.47	1.68	4.26
	4	56.32	1.42	2.52
academic words	1	52.78	3.04	5.75
	2	48.55	3.83	7.89
	3	63.86	4.71	7.37
	4	71.78	4.01	5.58

Note: the mean score and the percentage of free productive were calculated based on the lemmas produced by each of the participants.
The abbreviation 'Re' refers to receptive, 'CP' refers to controlled productive, and 'FP' refers to free productive vocabulary.

Table 10.9: The ratios among receptive, controlled productive, and free productive vocabularies among the four years cohorts

All the three types of ratios seemed to be high at higher frequency levels but low at lower frequency levels. For example, at the 2,000-, the 3,000-word levels and the academic word level, the fourth-year had the highest controlled productive/receptive ratios. At the 5,000- and the 10,000-word levels the first-year students had the highest controlled productive/receptive ratios. This implied that there was more

chance of receptive vocabulary becoming controlled productive vocabulary at higher frequency levels and at higher years of study.

However, the patterns for all the three ratios were not clear. For the controlled productive/receptive ratios, the first-year students had the highest ratios at the 5,000- and the 10,000-word levels, but the second-year or the third-year students had the lowest ratios at these two levels. For the free productive/receptive ratios, the fourth-year students or the third-year participants seemed to have the highest free productive/receptive ratios for all the frequency levels, except the 10,000-word level, where the second-year students had the highest.

For the free productive/controlled productive ratios, the second-year students had the highest free productive/controlled productive ratios at all the five frequency levels, and the first-year participants had the lowest ratios at all the levels, except the academic word level, where the fourth-year had the highest ratio. Therefore, when I examined in detail, the ratios of vocabularies for each year group did not seem to always relate to the vocabulary size, frequency levels, or years of study.

10.3 Findings about vocabulary knowledge within the developmental approach

I analysed the findings by summarising the number of the test lemmas with the developmental scores used in the receptive and the controlled productive tests in the present study. These two tests can be combined and used as one test (the self-reported format with verification). I used the five developmental score, following Paribakht and Wesche (1997).

- Score of 1 referred to no knowledge of the lemmas in such frequency level.
- Score of 2 referred recognising the written form of the lemma.
- Score of 3 referred to recognising the written form of the lemma and knowing its meaning.
- Score of 4 referred to recognising the written form, knowing its meaning, and being able to use it with semantic appropriateness in a sentence.
- Score of 5 referred to recognising the written form, knowing its meaning, and being able to use it with semantic appropriateness and grammatical accuracy in a sentence.

The highest numbers are written in bold in Table 10.10.

Frequency level	Year of study	Percentage of the test lemmas with these developmental scores				
		1 (%)	2 (%)	3 (%)	4 (%)	5 (%)
2,000 words	1	2.06	7.35	20.59	20.00	**50.00**
	2	0.83	5.42	20.00	12.92	**60.83**
	3	0.00	2.41	19.31	11.72	**66.55**
	4	0.00	0.83	9.58	15.00	**74.58**
3,000 words	1	8.53	15.88	19.41	15.88	**40.29**
	2	2.92	18.33	31.67	7.92	**39.17**
	3	1.38	15.17	21.03	9.31	**53.10**
	4	2.92	8.33	18.33	8.33	**62.08**
5,000 words	1	21.18	**37.94**	10.29	9.12	21.47
	2	14.58	**47.92**	15.83	7.92	13.75
	3	13.79	**39.31**	15.17	8.28	23.45
	4	11.25	**32.50**	20.42	9.17	26.67
10,000 words	1	**47.94**	43.24	2.65	2.35	3.82
	2	40.00	**52.50**	3.33	1.67	2.50
	3	**46.55**	38.97	7.93	1.72	4.83
	4	32.50	**51.67**	6.25	2.08	7.50
academic words	1	3.53	25.29	27.06	12.94	**31.18**
	2	0.42	19.17	**37.92**	10.00	32.50
	3	0.69	10.00	26.55	16.21	**46.55**
	4	0.00	4.58	18.75	18.33	**58.33**

Note:
Score of 1 = no knowledge of the lemmas in such frequency level
Score of 2 = recognising the written form of the lemma
Score of 3 = recognising the written form of the lemma and know its meaning
Score of 4 = recognising the written form, know its meaning, and being able to use it with semantic appropriateness in a sentence
Score of 5 = recognising the written form, know its meaning, and being able to use it with semantic appropriateness and grammatical accuracy in a sentence

Table 10.10: Percentage of the test lemmas with the five-scale developmental scores

The higher developmental scores were assumed to include the knowledge at the lower developmental scores. For example, 50% of the lemmas at the 2,000-word level for the first-year group were at the highest score, a score of 5. Around 20% of the lemmas were at a score of 4, and another 20% were at a score of 3. This implied that the first-year students knew the meaning of around 90% of the lemmas at the 2,000-word level. They were able to use around 70% of the lemmas with correct meanings and around 50% with correct meaning and grammar.

The majority of the lemmas at the 2,000- and the 3,000-word levels were also at the score of 5, with the higher percentages of the lemmas at the 2,000-word level. This implied that the participants knew the meanings of the lemmas at these two frequency levels and to use them in a sentence level.

The percentages of the lemmas at the 5,000-word level at the score of 5 dropped around 2-3 times, compared to the first two levels. The highest percentages (around 30-50%) were at the score of 2. This implied that the participants seemed to be able to use and knew the meanings of some of the lemmas and the 5,000-word level, but mostly only recognise their written forms.

The participants knew the meanings of only around 10% of the lemmas at the 10,000-word level and were able to use only around 2-8% of them. Half of the lemmas at the 10,000-word level were only at the scores of 1 and 2 of the developmental scoring system. It implied that the participants never encountered or recognised half of the lemmas at the 10,000-word level and they did not know the meanings of around 32-50% of the lemmas at this level. This showed that frequency level affected the participants' word knowledge and the ability to use the lemmas.

The knowledge of the academic lemmas of the participants was rated as 5 (around 30-60%) and 3 (20-40%). This implied that they knew the meanings most of the academic lemmas. Once they knew the lemma meanings, they used them with both semantic appropriateness and grammatical accuracy.

When I compared the data among the four year cohorts, I found that, the participants in higher years of study tended to know and were able to use more lemmas at the 2,000-word level and the academic word level. The percentages of the lemmas at the academic word level for the higher year group were higher.

Similarly, the percentages at the 3,000-word level for the higher year group was also higher. However, that for the second-year participants was 1% lower than that for the first-year participants.

At the 5,000-word level, the percentages at the scores of 3 to 5 for the participants at higher years of study seemed to be higher than those for the lower year group. This implied that the fourth-year participants still knew and were able to use most of the lemmas at this level. If the percentages of scores of 4 and 5 were combined, the fourth-year students were able to use most of the lemmas at this level, at least with correct meaning.

At the 10,000-word level, the percentages among the four year cohorts were roughly the same throughout the five scores.

At the academic word level, the percentages of the scores of 1 and 2 were lower but those at the scores of 4 and 5 were higher at higher years of study. The participants knew more academic words when they were at higher years of study.

The trends of the developmental scores showed that the participants tended to know more lemmas at high frequency levels. The percentages at the scores of 5 seemed

high but those at the scores of 1 and 2 seemed low at higher frequency levels. Moreover, the participants in higher years of study tended to know the meanings of more lemmas and were able to use them. At some frequency levels, participants from lower years of study could use more lemmas than those from higher years of study.

In conclusion, the frequency level and years of study had an impact on the participants' word knowledge and their ability to use lemmas in sentences. Even though the current study is a cross-sectional study, using the developmental score can somehow imply the gradual changes from one stage to another stage along a vocabulary knowledge continuum. One end of a continuum is having no knowledge or having never encountered the lemma, through recognising and knowing its meaning (receptive vocabulary knowledge), and to the other end – being able to use the lemma in a sentence with semantic appropriateness and grammatical accuracy (productive vocabulary knowledge). This is discussed in more detail in the discussion chapter.

I also analysed some of the participants' answers qualitatively for more insights of how receptive vocabulary knowledge is related to productive vocabulary knowledge. This is presented in Section 10.4.

10.4 Findings from qualitative analysis

10.4.1 Participants' answers in all the three tests

I analysed the relationship among receptive, controlled productive, and free productive vocabularies at different frequency levels qualitatively. I created a table for each of the test lemmas and put in the average scores that each year cohorts received. This was done to examine whether the participants would produce the test lemmas in the free productive test when they knew them receptively and were able to use them at a sentence level when forced to.

In the three tests, two points were given to correct answers and one point was given to partial correct answers. In Table 10.11, the number '2' indicates precise knowledge, the number '1' indicates partial knowledge, and the number '0' indicates to no knowledge. The numbers represented the points that the majority of the participants got for a particular test lemma. The symbols for free productive vocabulary are different. In Table 10.11, the symbol '/' means that such lemma did

not appear at least once in the free writings of that year group, and the symbol '-' means that such lemma never appeared in the participants' free writings.

The 2,000-word level

		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
sport	Y4	2	2	-	victory	Y4	2	2	-	birth	Y4	2	2	/	temperature	Y4	2	2	-	flesh	Y4	2	1	-
	Y3	2	2	/		Y3	2	1	-		Y3	2	1	/		Y3	2	2	-		Y3	2	1	-
	Y2	2	2	/		Y2	2	1	/		Y2	2	1	/		Y2	2	2	-		Y2	1	0	-
	Y1	2	2	/		Y1	2	1	-		Y1	2	1	/		Y1	2	2	-		Y1	0	0	-

		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
salary	Y4	2	2	/	education	Y4	2	2	/	scale	Y4	2	1	/	journey	Y4	2	1	-	treasure	Y4	2	2	/
	Y3	2	2	/		Y3	2	2	/		Y3	2	1	-		Y3	2	2	-		Y3	2	2	/
	Y2	2	2	/		Y2	2	2	/		Y2	2	0	-		Y2	2	2	-		Y2	2	2	/
	Y1	2	2	/		Y1	2	2	/		Y1	2	0	/		Y1	2	1	-		Y1	2	1	/

The 3,000-word level

		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
notion	Y4	1	0	/	palm	Y4	2	1	-	belt	Y4	2	2	-	chill	Y4	2	1	-	ox	Y4	2	2	-
	Y3	0	0	-		Y3	2	1	-		Y3	2	2	-		Y3	2	1	-		Y3	2	1	-
	Y2	1	0	-		Y2	2	1	-		Y2	2	1	-		Y2	2	1	-		Y2	2	1	-
	Y1	1	0	-		Y1	2	1	-		Y1	2	1	-		Y1	2	1	-		Y1	2	1	-

		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
structure	Y4	2	2	/	bench	Y4	2	1	-	charity	Y4	2	1	/	province	Y4	2	2	-	lieutenant	Y4	1	1	-
	Y3	2	2	-		Y3	2	1	-		Y3	2	1	/		Y3	2	2	-		Y3	1	0	-
	Y2	2	1	-		Y2	1	1	-		Y2	2	1	-		Y2	2	1	-		Y2	0	0	-
	Y1	2	2	-		Y1	1	1	-		Y1	2	1	/		Y1	2	2	-		Y1	0	0	-

The 5,000-word level

		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
pail	Y4	0	0	-	novelty	Y4	0	0	-	balloon	Y4	2	2	-	phase	Y4	1	1	-	mess	Y4	2	1	0
	Y3	0	0	-		Y3	0	0	-		Y3	2	2	-		Y3	1	0	-		Y3	2	1	0
	Y2	0	0	-		Y2	0	0	-		Y2	2	1	-		Y2	1	0	-		Y2	2	1	0
	Y1	0	0	-		Y1	0	0	-		Y1	2	2	-		Y1	1	0	-		Y1	2	1	/

		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
apron	Y4	2	1	-	compliment	Y4	2	1	-	apparatus	Y4	0	0	-	revenue	Y4	1	0	-	mare	Y4	1	0	-
	Y3	1	1	-		Y3	1	1	-		Y3	0	0	-		Y3	0	0	-		Y3	0	0	-
	Y2	1	0	-		Y2	1	1	-		Y2	0	0	-		Y2	0	0	-		Y2	1	0	-
	Y1	0	0	-		Y1	2	1	-		Y1	0	0	-		Y1	1	0	1		Y1	0	0	-

The 10,000-word level

		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
antics	Y4	0	0	-	batch	Y4	0	0	-	connoisseur	Y4	0	0	-	jumble	Y4	0	0	-	saliva	Y4	2	1	-
	Y3	0	0	-		Y3	0	0	-		Y3	0	0	-		Y3	0	0	-		Y3	2	1	-
	Y2	0	0	-		Y2	0	0	-		Y2	0	0	-		Y2	0	0	-		Y2	1	1	-
	Y1	0	0	-		Y1	0	0	-		Y1	0	0	-		Y1	0	0	-		Y1	1	1	-

		R	CP	FP			R	CP	FP			R	CP	FP			R	CP	FP			R	CP	FP
dregs	Y4	0	0	-	casualty	Y4	1	0	-	seclusion	Y4	0	0	-	revelry	Y4	0	0	-	apparition	Y4	0	0	-
	Y3	1	0	-		Y3	1	0	-		Y3	0	0	-		Y3	0	0	-		Y3	0	0	-
	Y2	0	0	-		Y2	0	0	-		Y2	0	0	-		Y2	0	0	-		Y2	0	0	-

	Y1	0	0	-		Y1	0	0	-		Y1	0	0	-		Y1	0	0	-		Y1	0	0	-
The academic word level																								
		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
labour	Y4	2	1	0	percent	Y4	2	2	-	principle	Y4	2	1	-	fund	Y4	2	1	0	technique	Y4	2	2	-
	Y3	2	1	/		Y3	2	1	-		Y3	1	1	-		Y3	2	1	/		Y3	2	2	-
	Y2	2	1	/		Y2	2	1	-		Y2	1	1	-		Y2	2	0	0		Y2	2	1	-
	Y1	1	1	0		Y1	2	1	-		Y1	1	1	-		Y1	1	1	/		Y1	2	1	-
		Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP			Re	CP	FP
philosophy	Y4	2	2	-	sum	Y4	2	1	0	consent	Y4	0	0	0	investigation	Y4	2	1	-	decade	Y4	2	2	
	Y3	2	2	-		Y3	2	1	/		Y3	1	0	0		Y3	2	1	-		Y3	2	1	/
	Y2	2	1	-		Y2	1	0	/		Y2	0	0	0		Y2	2	1	-		Y2	1	1	0
	Y1	2	1	-		Y1	2	1	/		Y1	1	0	/		Y1	1	1	-		Y1	1	1	/

Note: 'Y1' = Year 1 students, 'Y2' = Year 2 students, 'Y3' = Year 3 students, 'Y4' = Year 4 students, 'Re' = receptive vocabulary, 'CP' = controlled productive vocabulary, 'FP' = free productive vocabulary, 'Y' = years of study, '2' = two points for precise knowledge, '1' = one point for partial knowledge, '/' = appearing in the free writings

Table 10.11: Qualitative analysis of relationship among receptive, controlled productive, and free productive vocabularies

The data showed that the participants had wider precise receptive vocabulary knowledge. Fewer of them had precise controlled productive vocabulary knowledge, and even fewer test lemmas were retrieved and produced in the participants' compositions. Some of receptive vocabulary, especially words at high-frequency levels and academic words, could also be used productively both in a sentence level and an essay level.

10.4.2 Examples of participants' answers in all the three tests

Some of the participants supplied answers for the same test words in all the three tests: the receptive test, the controlled productive test, and the free productive test. I also examined how they supplied their answers. These are some examples:

Example 1: different degrees of receptive and productive knowledge of some aspects of a lemma

Tested lemma: birth

Answer in receptive test: 'การเกิด' (kan-koet, meaning 'birth/ being born')

Answer in controlled productive test:

"An offspring given **birth** by humans is not in a form of egg because humans are mammals."

Answer in free productive test:

"These kind of people do not have a chance to choose their **birth** they are what they are."

In this case, the participant had precise receptive knowledge, knowing the meaning of the lemma 'birth'. He was able to use the lemma in a sentence with correct meaning and grammar even though the raters stated that the whole sentence did not sound completely right. The sentence should have been written with an active voice rather than a passive voice. The lemma 'birth' was used as an uncountable noun and used with the verb 'give', so the raters gave two points for his sentence in the controlled productive test. However, when the participant retrieved this lemma again in the essay, he did not know that the lemma 'birth' was not normally used with the verb 'choose'. According to the raters, it should be 'choose where they were born', 'choose how they are born', or 'choose their birthplace'. Yet, as the word 'birth' was used with the meaning – being born, one point was given. Hence, the form and the meaning were correct, but the use was not completely right.

If based on Nation's (2013) taxonomy of word knowledge aspects for testing, this case showed that the participant had precise knowledge of some aspects and partial knowledge of some aspects of the lemma 'birth'. The knowledge of some aspects cannot be examined from the participants' answers, though. I present these in the following table – Table 10.12.

Form		
Spoken	Re	N/A: The three tests are tests for written vocabulary.
	Pro	N/A: The three tests are tests for written vocabulary.
Written	Re	Precise: He recognised it in the receptive test and in the controlled productive test
	Pro	Precise: He spelled the lemma correctly in the free productive test.
Word parts	Re	Precise: This knowledge was precise only for the form and meaning sense he used. He knew the base form of this lemma was a noun.
	Pro	N/A: He did not produce inflected or derived forms.
Meaning		

Form and meaning	Re	Precise for a particular sense: He gave one correct meaning of the lemma.
	Pro	Precise for a particular sense: He used this meaning sense in the sentences.
Concept and referents	Re	N/A: He knew its central concept but did not show whether he knew a range of uses of the lemma.
	Pro	N/A: He did not use the lemma to refer to a range of items.
Associations	Re	Partial or Precise: He produced associations such as 'offspring', 'give', humans', 'egg', 'mammal', but these words only might not be sufficient to judge whether this knowledge is precise.
	Pro	Precise: He could recall this lemma together with related ideas.
Use		
Grammatical functions	Re	N/A: The tests did not show context for this lemma.
	Pro	Partial: He did not always use this lemma in correct grammatical patterns.
collocations	Re	N/A: The tests did not initially show collocations for this lemma.
	Pro	Partial: He did not always produce its appropriate collocations.
Constraints on use	Re	N/A: The tests did not test this knowledge.
	Pro	Partial to Precise: He used the lemma 'birth' to talk about general issue such as money and life and the lemma 'birth' is a high-frequency lemma, which can be used to express general ideas.
Note: 'Re' = receptive knowledge, 'Pro' = productive knowledge, 'N/A' = not being able to judge or tested from the three tests used in the present study		

Table 10.12: Qualitative analysis of a participant's answers from the three tests (Example 1) using Nation's (2013) taxonomy

The qualitative analysis showed that a learner can have receptive knowledge of some aspects of a particular lemma and productive knowledge of some aspects of such lemma. The receptive and the productive vocabulary knowledge can also be classified into degrees: partial or precise knowledge. In Example 1, the knowledge of form and meaning tended to be precise, but the knowledge of use was partial. However, some aspects such as the receptive knowledge and the productive knowledge of spoken form as well as the receptive knowledge of the category 'use' could not be analysed by using these three tests. This finding is discussed in Section 11.2.3.

Example 2: Incorrect answer in the receptive test but correct meaning and grammar in the productive tests

Another example I present in this section is an example of answers where receptive knowledge is considered as no knowledge but actually found to be at precise degree when grading sentences.

Tested lemma: decade

Answer in receptive test: 'ศตวรรษ' (sa-ta-wat, meaning '100 years')

Answer in controlled productive test:

"Although **decades** have gone by, the Genocide done by Red Khmer is still in Cambodian people's head."

Answer in free productive test:

"it will be like an eternal flame shining in your mind no matter how many **decades** pass by."

From Example 2, even though the participant did not provide the correct translation for the tested lemma, he did use the lemma with its correct meaning and grammatical functions in both of the productive tests. The correct translation should be 'ทศวรรษ' (tot-sa-wat, meaning '10 years').

In terms of meaning used in the sentences, the surrounding context showed that the lemma did not mean '100 years' but around 10-90 years. The event of Cambodian genocide happened during 1975 and 1979, which is around 39-43 years before the year 2018 where the participant took the test.

Also, his second sentence relates to ideas in someone's mind and that few people live for a hundred years. Therefore, he seemed to know the meaning of the lemma, which is about time of around ten years.

Despite the wrong translation, he actually knew its meaning and he used the lemma 'decade' correctly as a countable noun and as a subject to the verb 'go' and 'pass'. I used Nation's (2013) taxonomy to examine the participant's receptive and productive knowledge level of the test word 'decade' (see Table 10.13).

Form		
Spoken	Re	N/A: The three tests are tests for written vocabulary.
	Pro	N/A: The three tests are tests for written vocabulary.
Written	Re	Precise: He recognised it in the receptive test and in the controlled productive test
	Pro	Precise: He spelled the lemma correctly in the free productive test.
Word parts	Re	Precise: This knowledge was precise only for the form and meaning sense he used. He knew the base form of this lemma was a noun.
	Pro	Precise: He produced an inflected form 'decades'.
Meaning		
Form and meaning	Re	Partial: He knew it was a period of time, about ten years, but he supplied a wrong Thai translation 'ศตวรรษ' (sa-ta-wat, meaning 'a hundred years'), the pronunciation of which is similar to the correct translation 'ทศวรรษ' (tot-sa-wat, meaning 'ten years').
	Pro	Precise: He used the lemma with correct meaning 'ten years' in the sentences.
Concept and referents	Re	Precise: He knew the concept of the lemma.
	Pro	N/A: He did not use the lemma to refer to a range of items.
Associations	Re	Partial or Precise: He produced associations such as 'many', 'go by', 'pass by', but these words only might not be sufficient to judge whether this knowledge is precise.
	Pro	Precise: He could recall this lemma together with related ideas in both productive tests.
Use		
Grammatical functions	Re	N/A: The tests did not show context for this lemma.
	Pro	Precise: He used this lemma with appropriate grammar.
collocations	Re	N/A: The tests did not initially show collocations for this lemma.
	Pro	Precise: He produced its appropriate collocations.
	Re	N/A: The tests did not test this knowledge.

Constraints on use	Pro	Precise: He used the lemma at appropriate times, writing the lemma to refer to time both in the past and the future.

Note: 'Re' = receptive knowledge, 'Pro' = productive knowledge, 'N/A' = not being able to judge or tested from the three tests used in the present study

Table 10.13: Qualitative analysis of a participant's answers from the three tests (Example 2) using Nation's (2013) taxonomy

These two examples of participants' answers to the three vocabulary tests in the present study indicate the complexity of the relationship between receptive and productive vocabulary knowledge. The participants had precise knowledge of receptive vocabulary and partial knowledge of productive vocabulary, or vice versa. Their knowledge of some aspects was precise and that of some aspects was partial. These also showed some limitations of the scoring methods. All these points are discussed in Section 11.2.3.

10.5 Chapter conclusion

In this chapter, I quantitatively examined the relationship among receptive, controlled productive and free productive vocabulary knowledge in relation to frequency levels and years of study. I examined this within the dichotomous approach by comparing the average scores and by analysing the ratios among them.

I found that frequency levels affected the participants' knowledge of the three vocabularies. With the idea that these three were separate groups of vocabulary and they competed one another, their receptive vocabulary was always the largest, followed by controlled productive vocabulary and free productive vocabulary. They were large at high-frequency levels and at academic word level. They became smaller when the frequency levels became lower.

At different years of study, the three groups of vocabulary were larger in size, with the receptive vocabulary always the largest. The scores for the controlled productive vocabulary showed the biggest gap at the high-frequency levels (the 2,000- and the 3,000-word levels) and the academic word level. This might be because the four year cohorts had relatively high scores of the receptive vocabulary at these three frequency levels.

With the idea that receptive vocabulary could become productive vocabulary in the dichotomous approach, the ratios showed that there were more chance for the receptive vocabulary to become the controlled productive vocabulary and the free productive vocabulary at higher frequency levels and the academic word level.

I also examined the relationship between receptive and productive vocabulary knowledge within the developmental approach by observing the changes of the developmental scores. I found that the participants' knowledge of the lemmas was at the score of 5 at high-frequency levels and the academic word level. More lemmas at the score of 5 were known by the learners at high years of study.

If the differences in scores can imply the vocabulary development as in a longitudinal study, the vocabulary knowledge over the years of study was incremental. However, at mid- and low-frequency levels, the participants' knowledge seemed to be only at the score of 1 or 2. This means the majority of the participants did not recognise or did not know the meaning of the lemmas at these frequency levels.

Finally, I qualitatively examined the relationship among these three groups of vocabulary. I found that the participants had receptive vocabulary knowledge before they were able to use the test lemmas in the two productive vocabulary tests. When I examined some of the participants' answers by using Nation's taxonomy, I found that the participants' receptive knowledge of some aspects was precise and that of some was partial. This also applied to the productive knowledge of some aspects. Using all these analyses, I found that the relationship between receptive and productive vocabulary knowledge is quite complex. I discussed and concluded this in the discussion chapter, Chapter 11.

Chapter 11 Discussion and conclusion

11.1 Introduction

In this chapter, I discuss two main issues, which are the methodological insights and the research findings. The methodological insights include research designs, word resources, the test formats, and the limitations of the research tools. The findings include the participants' receptive, controlled productive, and free productive vocabularies and the relationship among the three from both the quantitative analysis (the dichotomous and the developmental approaches) and the qualitative analysis based on Nation's (2013) taxonomy.

11.2 Methodological insights

11.2.1 Research designs (longitudinal vs. cross-sectional study and number of participants)

The current study aimed to explore the vocabulary knowledge of the English major students from all the four year cohorts, but time for conducting this study did not allow four years of data collection. Because of time limitations, this research study adopted a cross-sectional research design that compared the data of the participants from different year cohorts in order to examine the vocabulary knowledge at different years of study. I also bear in mind that the scores obtained from my cross-sectional study cannot completely refer to the development of the participants' vocabulary knowledge. Unless tracking the participants' vocabulary scores along the four years of study can be done, a longitudinal study is recommended for more solid results regarding vocabulary changes at different years of study.

With the cross-sectional design, the number of participants per year group should be at least 30 participants as suggested by (2017). Because the number of the participants in each year group did not reach the optimal number of participants per variable, I then firstly divided the participants into two groups. The first- and the second-year groups were also considered as the lower year group, and the third- and the fourth-year groups were considered as the higher year group. Hence, it is recommended that future research recruit at least 30 participants per year group for statistical reliability.

11.2.2 Word lists as resources of test words and the number of test lemmas included

This study aimed to examine the participants' knowledge of vocabulary in general. As the participants were at a university level, academic words were included because university students are expected to use a lot of them throughout their years of study.

For comparability of the findings, I decided to use the test words from a widely-used receptive vocabulary – the VLT by Schmitt et al. (2001). The reliability and validity of the test have also been tested. It tests academic words and words from four frequency bands (the 2,000-, the 3,000-, the 5,000-, and the 10,000-word levels), which cover high-, mid-, and low-frequency levels. However, the word lists used in the word sampling process of the VLT were relatively old. The general word lists are Thorndike and Lorge's (1944) list, West's (1953) GSL, and the academic word list is Coxhead's (2000) AWL.

This limitation affects the participants' knowledge scores for some lemmas. For example, the lemma 'lieutenant' was in the 3,000-word level of the VLT, but in other more recent corpus such as the BNC (Nation, 2012) this lemma was in a lower-frequency level – the 6000-word level. Therefore, the majority of the participants from each year cohort was not familiar with the lemma because it was not frequently used any longer. Most of the participants seemed to be able to use the rest of the lemmas at this frequency level, though.

Since the aim of my study was not to examine the knowledge of particular words, this did not affect the whole picture of vocabulary knowledge in general. Also, this was not directly related to the research question. I then did not present this in the findings chapter. However, it is recommended for the research in the future use more recent word lists for word sampling process such as the BNC/COCA lists (Nation, 2017). It might reflect participants' vocabulary knowledge more clearly.

Another issue is to do with how words are counted. I used the lemma in my study, while the other studies used word families. I decided to use a lemma because it is more straightforward than word families in terms of the members they include. It is also suitable for analysing productive vocabulary as suggested by Schmitt (2010). Therefore, to make the three tests consistent and comparable, I used the lemma to count word units in my study. Hence, only the patterns of how vocabulary knowledge changes were compared with those of other studies.

In terms of the number of test items, the controlled productive test task (a sentence-writing task) in the present study was relatively demanding. Including a larger

number of test lemmas might have led to participants' fatigue effect. Hence, only ten lemmas could be included both in the receptive test and the controlled productive test. If any research study in the future uses a less demanding task, more test lemmas would be ideal for a better understanding of vocabulary knowledge.

11.2.3 Test formats and scoring methods

There were three vocabulary tests used in this study: the receptive test, the controlled productive test, and the free productive test.

The receptive test was in the self-reported scale with verification – translation task. It tested the ability to recognise the written form and to know the meaning of the test lemmas. It did not test the participants' reading ability or whether the participants understood the test items in reading texts, though. The participants needed to recognise the written forms of the test lemmas before they supplied the meanings of such lemmas. Therefore, it focused on only form recognition and meaning recall. Even though I acknowledged that many aspects are involved when knowing a word, the receptive test did not test the knowledge of other aspects. Here are the reasons.

> 1) A word is meaning-driven unit (Nation, 2013), so the aspect of form and meaning is the most fundamental aspect (Laufer and Goldstein, 2004).

> 2) The aspect of form and meaning is most related to productive vocabulary and can predict around one-third of productive vocabulary knowledge (Zhong, 2016).

Based on these ideas, I then decided to include only one aspect – form and meaning – in the receptive test.

The receptive test was combined with the controlled productive test into a vocabulary knowledge scale test. The controlled productive test was in a sentence-writing format. Although the controlled productive test was considered by Schmitt and McCarthy (1997) as a big jump from receptive knowledge, I considered that it represented a productive skill in real life – using words in context to convey what we think.

Another task – an essay-writing task – was used as a free productive test. When writing in these two productive tests, it is unavoidable for the participants to use other knowledge aspects such as collocation, derived forms, associations, grammatical knowledge, and register. The two productive tests in this study involves other knowledge aspects. They did not only involve the aspect of form and meaning but

also the aspects of use. Hence, these three tests in my study involves the three main categories in Nation's (2013) taxonomy, i.e., form, meaning, and use.

The two main things that had been done in the present study differently from other research studies in the field are 1) the attempt to score the three tests in the same way for the consistency of the analysis and 2) scoring the receptive and the controlled productive tests by adopting both the dichotomous and the developmental approaches.

The first thing that has been done differently from other research studies in the field is using the same word lists and the same scoring system for the three tests. Future research can adopt this method for the consistency. In other studies exploring the relationship among receptive, controlled productive, and free productive vocabularies such as the studies by Laufer (1998) and Laufer and Paribakht (1998), the participants' compositions were analysed into lexical profiles by using LFP (Laufer and Nation, 1995). The LFP checked the frequency of words produced against the GSL and the UWL. Hence, the lexical profiles would present only the words falling into the first 2,000 words and into the UWL, but not words falling into the 3,000-, the 5,000-, or the 10,000-word levels. All these would be combined and presented as one group of 'not-in-list words, instead.

In my study, the three word lists that were used in the word sampling process for the VLT (Schmitt et al., 2001), which are Thorndike and Lorge (1944), West's (1953) GSL, and Coxhead's (2000) AWL word list, were used in the process of sorting lemmas produced by the participants (using the software called Text Lex Compare on Cobb's (n.d.) Lextutor website). By doing so, the data of free productive vocabulary could be compared to that of the receptive and the controlled productive tests. This makes the scoring system of the three tests more consistent.

However, the total size of the participants' free productive vocabulary could not be measured because the calculation could only be based on the number of lemmas produced by the participants, not on those they did not produce. This is still a limitation in the field and more appropriate measure is needed for the research in the future.

In addition to sorting produced lemmas with the same word lists, I also checked the accuracy of using the lemmas in the free productive test with the same scoring system as in the controlled productive test. They were graded by the same raters and with the same references. This has never been done in previous research studies in the field.

The second innovation of this study that is different from other studies is the adaptation of two scoring systems to analyse both receptive test and controlled productive test (both the dichotomous and the developmental approaches). Reviewing the literature, I found that there was no consensus of what scholars think about the relationship between receptive and productive vocabularies. The common findings – receptive vocabulary size is always larger than productive vocabulary size – can be interpreted in different ways.

     1) Receptive vocabulary is acquired before productive vocabulary.

     2) Productive vocabulary is a subset in receptive vocabulary.

     3) A productive test is more demanding than a receptive test.

All these make the productive vocabulary scores lower than receptive scores. Hence, I did not assume about the relationship between them. I included the receptive test and the controlled productive test as one self-reported knowledge scale (with verification) so that I could score the answers from the tests by using dichotomous scoring and developmental scoring systems.

I tried scoring the data by using the dichotomous approach and found that the answers from the three tests could be scored and compared with one another as if the three vocabularies were competitors. They tended to be higher when the participants were in higher years of study, with controlled productive vocabulary showing the biggest gap.

Similarly, when I scored them by using the developmental approach following Paribakht and Wesche (1997), I also see that the participants' vocabulary knowledge gradually shifted to another scale (both in upward and downward trends) when the participants were in higher years of study. It shifted as if it was moving on a continuum.

Therefore, the test formats researchers use appear to represent what they think about the relationship between receptive and productive vocabulary knowledge, and the findings support what they think. If they believe the two vocabularies are separate groups of words, they can test each of them and compare the results statistically. The receptive and controlled productive vocabulary growth would be analysed in terms of sizes and scholars will compare them. If scholars believe that vocabulary knowledge gradually develops form no knowledge to being able to use a word, then they will decide to use a knowledge scale. Researchers seem to have used test formats that match their thoughts and reported the results as what they thought.

However, both approaches can still lead to two ways of interpretation. In the dichotomous approach, calculating controlled productive/receptive ratios leads to the idea that receptive and productive vocabularies are different groups and ratios represent the gap between them. The ratios can also be interpreted as the chance of receptive vocabulary becoming productive vocabulary. Hence, this seems more like the developmental approach – receptive vocabulary becoming productive vocabulary. The ratios can show only the chance but cannot show at which point receptive vocabulary becomes productive vocabulary or whether the assumption is correct.

In the developmental approach, it is believed that receptive vocabulary needs to pass through many stages before it becomes productive vocabulary. This can be interpreted that productive vocabulary is a subset of the receptive vocabulary because the productive word can still be understood or used receptively. This can also be interpreted that words are stored as one group in mental lexicon and change their position along continuum from no knowledge to partial receptive knowledge, to precise receptive knowledge to partial productive knowledge and to precise productive knowledge.

When the findings of my study can be interpreted within both approaches, I then further qualitatively analysed the data. I exemplified the answers of one of the participants, as shown in the second example in Section 10.4.2. In the example, I found that the participants actually knew the meaning of that lemma but only supplied the wrong translation. He meant 'ten years' for the lemma 'decade' in the sentences supplied, but he supplied 'ศตวรรษ' (sat-ta-wat, meaning 'a hundred years') instead of 'ทศวรรษ' (tot-sa-wat, meaning 'yen years') in the receptive test – translation task. This mistake must have been a result of the similarity in sounds of the two Thai translations, and the participant got them mixed up.

This sample shows one limitation of the scoring method, that is, researchers should not assume that once the wrong translation is supplied and such participants will not be able to use the word in sentences. Therefore, I suggest two points for future research: 1) researchers not assume from the answer of the receptive test but the sentences also be checked carefully even though the wrong translation was provided and 2) one more step of verification or an interview be used for the receptive test be applied so that it is more certain whether the participants actually knew its meaning.

In conclusion, the same data in my study can be both analysed and interpreted with both the dichotomous and the developmental approaches. Using the same scoring system as the other two tests, sorting words from the free productive test and checking the accuracy of word use are the innovations of the current research study.

The qualitative analysis of answers from the three tests was also done and showed some points which can be used to support the relationship between receptive and productive vocabulary knowledge. I discussed what I found when qualitatively analysed the data and presented them in Section 11.3.4.

11.3 Findings

Following the order of the research questions, I started with the discussion of the findings about each group of vocabulary (Section 11.3.1 – Section 11.3.3) before the findings about the relationship among them (Section 11.3.4). The findings presented in Section 11.3.1 to Section 11.3.3 are within the dichotomous approach. I discussed them in terms of possible explanations for the findings, limitations of the research methodology, and pedagogical implications. In Section 11.3.4, I discussed the findings from the quantitative analysis (within both the dichotomous approach and the developmental approach) and the qualitative analysis.

11.3.1 Receptive vocabulary knowledge

In this section, I discuss the findings of the first research question, receptive vocabulary knowledge at different frequency levels and the receptive vocabulary knowledge at different years of study. The possible explanations include chance of encountering words, learning contexts, participants' interests, and the ceiling effect. Limitations of methodology include versions of vocabulary tests and the number of participants per variable. Pedagogical implications include using English as a medium and explicitly teaching register-based words for EFL students.

Receptive vocabulary knowledge in this study refers to the ability to recognise L2 written form and to recall one of its L1 meanings. The knowledge is relevant to the aspect of form-meaning link in Nation's (2013) word knowledge taxonomy. The term 'word' used in discussion refers to 'word families' for other studies but to 'lemmas' for my study. For comparability, I compared percentages and focused more on vocabulary knowledge patterns or trends. Percentages were rounded for the ease of comparison.

Firstly, the findings of the present study suggest that the participants' receptive vocabulary was significantly affected by frequency levels. It decreased when the word frequency became lower. The participants from all the four year cohorts had the highest scores at the highest frequency level (the 2,000-word level) and the

lowest scores at the lowest frequency level (10,000-word levels). However, the receptive vocabulary scores at the 3,000-word level and those at the academic word level were not significantly different.

This is consistent with the findings of many studies in different contexts such as studies in China (e.g., Fan, 2000; Li and MacGregor, 2010; Zhang and Lu, 2013; Zhong and Hirsh, 2009), Malaysia (e.g., AbManan et al., 2017), Japan (e.g., Waring, 1997a; Webb, 2008), Korea (e.g., Shin et al., 2011), Afghanistan (e.g., Amin, 2020), Turkey (e.g., Ozturk, 2015), Israel (e.g., Laufer, 1998), and ESL in Canada (e.g., Laufer and Paribakht, 1998). Despite the different number of the frequency levels tested, the different receptive test formats, and different groups of participants, the patterns of vocabulary knowledge at each frequency level are similar. Therefore, this vocabulary knowledge pattern in relation to frequency levels seems to be universal.

For example, Fan (2000) studied first-year non-English students' receptive knowledge of only three frequency levels (the 2,000-, the 3000, and the academic word levels) using the VLT. The participants had the highest scores at the 2,000-word level (89%). The scores at the 3,000-word level were lower (78%), and the scores at the academic word level were slightly lower than that at the 3,000-word level (75%).

Shin et al. (2011) examined high school students' receptive vocabulary knowledge by using the VST, which includes ten frequency levels. They also found that the participants knew around 94% of the 1,000-word level, 90% at the 2,000-word level, and the vocabulary scores decreased to 40% at the 10,000-word level.

Amin (2020) explored the fourth-year English-major students' receptive knowledge by using the VST. The highest scores were at the 1,000-word level and the lowest was at the 10,000-word level as well.

This is probably because there is more chance for learners to encounter high-frequency words, the learners, therefore, are familiar with them and knowing their meanings. Webb (2007a) found that in one of his classroom experimental studies the more learners encounter words, the more knowledge of those words that learners gain.

Besides, the participants in my study from all the four year cohorts reached the mastery levels at the 2,000-word level. This is also true for many studies (e.g., AbManan et al., 2017; Li and MacGregor, 2010; Zhang and Lu, 2013). However, this finding is not in line with that of a few studies with participants at a university level (e.g., Amin, 2020; Ward, 2009b).

Ward (2009b) investigated the first-year students majoring engineering in Thailand He found that the participants knew around only 1200 words out of the most 2,000 frequent words, or only 60%. This implied that the participants' receptive vocabulary knowledge did not reach the mastery level even at the high frequency level. This might be because the participants' interest was not the English language as Ward mentioned in his paper, and they had been studying in Thai programme in the EFL context. Therefore, a small chance of encountering English in everyday life in the EFL context and lack of interest in English as non-English majors might lead to their low proficiency in the English language.

Similarly, the fourth-year participants in Amin's (2020) study in Afghanistan did not reach the mastery at the 2,000-word level although they did master the first 1,000 words. Hence, when the scores at the 1,000- and the 2,000-word levels were sum up and averaged, they knew roughly 69%. Even though the participants were English majors in their final year at the university and the receptive test format (the multiple-choice format) was less demanding than the one used in my study, the scores were still low. Therefore, the EFL context seemed to play an important role in the participants' vocabulary learning in Amin's study.

In my study, I also found that the participants' receptive scores at the academic word level were similar in size to those at a high-frequency level, the 3,000-word level. Many studies that employed the Schmitt et al.'s (2001) VLT (e.g., AbManan et al., 2017; Fan, 2000; Li and MacGregor, 2010; Ozturk, 2015; Zhang and Lu, 2013) showed similar results. This might be because the test words in my study were selected from test words in the VLT (Schmitt et al., 2001), so they were sampled from the same word lists.

The scores at the academic word level of the participants in Li and MacGregor's (2010) study were higher than the scores at the 3,000-word level. This showed that the university students seemed to know a lot of academic words. This might be because academic words have been used a lot in academic texts and students at a university level tend to be exposed to them often. This then implied that frequency is not the only factor affecting learners' receptive vocabulary knowledge but register also does have an impact on it. Therefore, explicitly teaching register-based words can be beneficial for EFL learners.

However, in some studies such as Laufer (1988) and Laufer and Paribakht (1998) the academic word scores were relatively low and roughly similar to those at the 5,000-word level, instead. This might have been affected by using different versions of the VLT – they used Nation's (1983; 1990) VLT, so different word lists were used in the word sampling process. While Schmitt et al. (2001) used Coxhead's (2000)

AWL, Nation (1990) used the UWL (Xue and Nation, 1984). Hence, tests based on different word lists also affects vocabulary knowledge scores.

Another possible explanation might be that the participants in these two studies included high school students and they might not have encountered a lot of academic words in class. Pedagogically, academic words should be given more attention in high school classes when one of the goals is to prepare learners for studying in a tertiary level.

In terms of the receptive vocabulary at different years of study, the differences of receptive scores between two proficiency groups and among the four year cohorts were interpreted as receptive vocabulary gains or attritions as my study is a cross-sectional study. I acknowledged that this should be interpreted carefully. The findings showed that the participants' receptive vocabulary knowledge tended to be larger for the students at higher years of study. Therefore, years of study seem to be one of the factors affecting receptive vocabulary knowledge.

This finding is consistent with that in both cross-sectional studies and longitudinal studies (e.g., Al-Masrai and Milton, 2012; Alonso and Fontecha, 2014; Gyllstad et al., 2015; Laufer, 1998; Sungprakul, 2016; Zhang and Lu, 2013; Zhong and Hirsh, 2009). The percentage of gains varied.

In my study, the gaps of the scores between the higher and the lower year groups ranged from 6-16%, with the largest gain (16%) at the academic vocabulary and the smallest gain (6%) at the 10,000-word level. In a longitudinal study by Zhang and Lu (2013), the largest gain was at the 5,000-word level and the smallest gain was at the 2,000-word level. The small gain at the 2,000-word level can probably be explained by the ceiling effect. The participants at their first year knew around 91% of the 2,000-word level (highest among the five levels). Hence, there was not much room for improvement.

In a cross-sectional study by Li and MacGregor (2010), the MA participants even had lower receptive score (97%) at the 2,000-word level than the first-year students (99%). This small difference is not significant, though. The scores at this level of both groups were very high, and this might be because they were in the ESL context and used the English language regularly. For pedagogical implication, if English is used more as a medium in EFL classroom, this might help improve EFL learners' receptive vocabulary knowledge.

In my study, a group in higher years also had lower scores at some frequency levels than a group from lower years. The data showed that the second-year group scored slightly lower than the first-year group at the 5,000- and the 10,000-word levels.

Their scores were not significantly different at all levels, except the academic word level. For my study, the following data from the questionnaire and some limitations of the study might be the reasons for the slightly lower scores of the second-year group and for that the receptive vocabulary between these two was nearly similar in size.

- More of the first-year students studied in an English-programme in their secondary schools (15% of the first-year students but only 8% of the second-year students).
- The first-year students on average spent more time abroad (around four months for the first-year group and two months for the second-year group).
- The number of years these two groups had been studying English was nearly the same (13.53 years for the first-year group and 13.87 years for the second-year group).
- At the time when this study was conducted, the second-year students had just moved up from their first year, in which only two basic English courses were taken (English 1 and English 2) throughout the year. These courses focused on only basic English skills – reading, writing, listening, and speaking. Accordingly, using the English language as a medium at their secondary schools and exposure to English in English-speaking countries seemed to affect their receptive vocabulary knowledge to some extent. Moreover, the students in Year 2 might have not acquired a lot more vocabulary than the first-year students due to having learned English in class for almost equal amount of time as those in Year 1 and having taken a few English courses in the previous year. Their vocabulary knowledge was then not much different.
- Methodologically, these ANOVA results were calculated from the participants of fewer than 30 per variable – an optimal number suggested by Cohen et al. (2017). This might have had an impact on the statistics. Therefore, it is recommended for future research that there be at least 30 students per year group when comparing their vocabulary knowledge.

In conclusion, word frequency levels and years of study play an important role in learners' receptive vocabulary knowledge and vocabulary acquisition. Learners had large vocabulary size of high-frequency words and small vocabulary size of the words rarely used and encountered in everyday life or in classrooms. University students are in academic fields, so they tended to know a lot of academic words. Additionally, learners' receptive vocabulary tends to be larger for the participants in higher years of study. Some variations might have been caused by versions of tests that have been used or the number of participants per variable. Future research can further the investigation on these points.

11.3.2 Controlled productive vocabulary knowledge

This section discusses the participants' controlled productive vocabulary knowledge. Controlled productive vocabulary knowledge in the current study refers to the ability to use the prompts in a sentence with semantic appropriateness and/or grammatical correctness. The points I discussed include words encountering and chance to use them. I also discussed the limitations of the methodology and suggestions for future research in terms of a cross-sectional research design and the number of participants per year group.

The findings in my study showed that the participants' controlled productive vocabulary decreased as the word frequency levels became lower. The participants were able to use the most words at the 2,000-word level and the fewest words at the 10,000-word level. The percentage scores of their knowledge of academic words were roughly equal to those of words at the 3,000-word level. They were not statistically different.

The findings are concordant with similar studies in other countries both in the EFL and ESL learning context such as Israel, Canada, China, Malaysia, Iran, Turkey, and Afghanistan (e.g., AbManan et al., 2017; Amin, 2020; Laufer and Nation, 1999; Laufer and Paribakht, 1998; Zhang and Lu, 2013; Zhong and Hirsh, 2009) even though they employed a different test format such as a cued-recall test – Laufer and Nation's (1999) PVLT test. Controlled productive vocabulary knowledge for this test referred to the ability to recall L2 words to fit the context provided. Webb (2008) used an L1 to L2 test format and sampled words from a different corpus, but the finding also shows the underlying fact that learner's controlled productive vocabulary knowledge is affected by frequency levels.

The percentages at each of the five frequency levels ranged widely. University students' controlled productive vocabulary scores at the 2,000-word level ranged from only 37% (the fourth-year students in Amin's (2020) study) to as high as 94% (the first-year students in Laufer and Nation's (1999) study).

Both high school EFL students and university students can master controlled productive vocabulary at this level. For example, in a cross-sectional study by Laufer and Nation (1999), the 11th and the 12th graders scored very high at 83-90%, and the 11th graders in a longitudinal study by Zhong and Hirsh (2009) scored 85% and improved to 88% ten months later.

In the second study (a longitudinal study) by Ozturk (2015), the first-year participants scored 85%. These students scored higher than the fourth-year students in my study

(84%) although they were also in the EFL context and they were in lower years of study. In my study, only the fourth-year students mastered controlled productive vocabulary at the 2,000-word level. The difference in test formats might have been involved in the differences in scores. The cued-recall format seems to be less demanding than a sentence-writing task used in my study.

Therefore, for both pedagogical and research purposes, test formats should be consistent when comparing vocabulary knowledge of different groups of learners. Test score interpretation should be done in a careful way as controlled productive vocabulary knowledge referred to different abilities. It can be either the ability to recall words in context or the ability to use words given in a sentence level. Perhaps, these two abilities should be classified differently, not only simply as controlled productive vocabulary knowledge.

At the 3000-word level, scores became lower and fewer groups of students in previous studies mastered this level. Among the studies I reviewed, only the first-year students in Laufer and Nation's (1999) study scored more than 80%. None of the participants in my study mastered controlled productive vocabulary at this level. At the 5,000-word level, the scores did not reach 50% at each frequency level, ranging from 4-49%. In previous studies, the vocabulary scores at the 10,000-word level were the lowest and could range from 4% to 27%.

The controlled productive vocabulary at the academic word level tends to be similar in size to the knowledge at the 3,000-word level. For example, the first-year participants in Ozturk (2015) knew 46% of academic words and around 45% of vocabulary at the 3,000-word level. Their knowledge of academic words is relatively high, compared to other studies. This can possibly be explained that the participants in Ozturk's study "were highly advanced in English as they had to pass a very competitive national English test to be admitted to the programme" (Ozturk, 2015, p. 96).

Hence, context is an important factor of vocabulary knowledge. Controlled productive vocabulary scores might vary from context to context, but the overall trends seem to be similar. This shows that it is an underlying fact about the relationship between controlled productive vocabulary knowledge and word frequency levels. Moreover, learners' general language competence is directly linked to controlled productive vocabulary knowledge level. Pedagogically, vocabulary tests can also be used as a quick placement test to classify students.

Concerning the controlled productive vocabulary knowledge at different years of study, the higher year group had significantly larger controlled productive vocabulary than the lower year group at all the frequency levels with large effect sizes. This

implied that there was a growth of this vocabulary over the years of study. However, the findings were from a cross-sectional design, and a longitudinal study might provide more solid findings.

Their knowledge of academic words showed the biggest gap (23%), reflecting the learning environment in a university level and the writing assignments they had been required to do. However, the gap of the scores between year groups was small at the 10,000-word level. This is similar to a cross-sectional study by Ozturk (2015) where there was no significant difference of controlled productive vocabulary knowledge between the fourth-year students and the first-year students at this level because these words were infrequent. The participants then did not have much opportunity to encounter them. As a result, only a few words were then acquired and used in a sentence level correctly. Also, teachers have probably not taught them explicitly or not asked learners to use them in sentences. This is an assumption that future research can investigate.

My students in higher years of study generally had a larger controlled productive vocabulary than those in lower years of study, but the second-year students scored lower than the first-year students at certain frequency levels (the 3,000-, 5,000-, and 10,000-word levels). These lower scores in their controlled productive vocabulary were not statistically significant, but they might be attributed to the same four points I discussed in the section of their receptive vocabulary: 1) fewer of them were from English-programme schools, 2) they spent shorter time abroad, 3) they spent roughly equal number of years learning English as the first-year students, and 4) they took only a few English courses in their first year. These four points are all contextual factors.

Methodologically, the lower scores of the Year 2 students is also possibly due to the limitation of the cross-sectional study with participants of fewer than 30 per variable. Hence, in order to gain more insight whether this is a real vocabulary decline along vocabulary development, tracking individual's vocabulary development at these five frequency levels longitudinally would be ideal if time is not constrained.

In conclusion, word frequency levels and years of study have an impact on L2 learners' controlled productive vocabulary knowledge. Their controlled productive vocabulary at high frequency levels and at the academic word level was larger than that at lower frequency levels. However, not many L2 learners could reach the mastery level of this knowledge, even at the 2,000-word level. Therefore, having learners to use target words in class might yield benefits for teaching and learning English. Learners' vocabulary expanded as they moved up to higher years of study. The academic vocabulary increased the most reflecting the register to which they

294

were exposed at university; register also has an impact on vocabulary knowledge. Finally, the low of controlled productive vocabulary at some year cohorts might have been due to the limitation of a cross-sectional research design or the time the participants spent using English.

11.3.3 Free productive vocabulary knowledge

Free productive vocabulary knowledge means the ability to produce and use words with correct meaning and appropriate grammar in compositions. The participants' free productive vocabulary was calculated based on the number of lemmas produced by each of the participants, not the number of all the lemmas they could produce at free will with correct meaning and grammar. Therefore, the mastery level could not be analysed. Like the two previous sections (Section 11.3.1 and Section 11.3.2), I discussed major findings and pointed out possible explanations, pedagogical implications and limitations of methodology. The limitations lead to suggestions for research in the future.

To answer the first sub-question, the differences in the percentage scores of the free productive vocabulary were statistically significant. This showed that frequency levels affected free productive vocabulary knowledge. However, the percentage scores at the 2,000-word level and those at the academic word level were not significantly different. This might be because the participants used many lemmas from these two levels in their free writing task and used them correctly in meaning and in grammar. Hence, academic words appeared to have been exposed frequently and the participants were familiar with using them.

The third-year participants produced the most lemmas (falling into the five frequency levels) in their compositions. The lemmas they produced were also graded in terms of accuracy of usage, similar to those in the controlled productive vocabulary test. This was done for the consistency of the research design, and it has not been done before by previous research studies. Therefore, research in the future can adopt this method. The participants on average produced more lemmas at the 2,000-, the 3,000-, and the academic words. They did not produce many lemmas at the two lower frequency levels, the 5,000- and the 10,000-word levels.

These are congruent with the findings of many previous studies (e.g., Azodi et al., 2014; Djiwandono, 2016; Laufer, 1998; Laufer and Nation, 1995; Zheng, 2012). However, these studies used the LFP (Laufer and Nation, 1995), which showed the percentages of produced tokens, types, and word families into wider frequency-based groups of words: the 1,000-, the 2,000-word levels, the academic word level,

and the 'Not-in-list' group. They all showed that the participants produced most word families at the 2,000-word level and the smaller number of word families at the other groups. The percentages of the lemmas at the rest of frequency levels such as the 3,000-, the 5,000-, and the 10,000-word levels, cannot be compared nor can the scores of word usage. Even though different research tools were used by studies, the findings were similar. This means the underlying reality is being captured by the research instruments in my study. Therefore, I suggest that future research sort words into the same frequency-based groups for comparability of the findings.

When I compared the scores of the lemmas that were correctly used by the participants, I found that the higher year group had significantly higher scores than the lower year group. I also found that the largest gap of free productive vocabulary between the year groups was at the academic word level. Again, the possible explanation for this is the learning context; university students have encountered academic words frequently and are required to use them in many of their assignments. Accordingly, teachers can assign learners to use the vocabulary in their fields if such groups of vocabulary are necessary for their study objectives and their career in the future.

Furthermore, that the higher year group on average produced more lemmas in their essays than the lower year group did not affect the scores of correct lemmas. This is because the scores were calculated based on the possible total score of all the lemmas that a participant produced in his composition. The scores were then averaged before comparison. However, another way that research in the future might choose to do for the validity of the test would be to analyse the same number of running words produced by each of the participants. For example, only the first 300 running words in each of the essays will be analysed, as conducted in the study by Laufer and Nation (1995).

A comparison of the free productive scores (at all the five frequency levels) among the four year cohorts showed that the participants from the higher years of study had higher scores. However, when I examined the data at different five levels, I found no common patterns. The participants in higher years of study did not always have the highest free productive scores. The third-year students or the second-year students could have higher scores than the fourth-year students at some frequency levels. The differences in scores were not statistically significant. This can be explained that the scores were relatively low and roughly the same for all year groups.

The analysis also showed that the largest gap of scores among the year groups was at the academic word level between the third-year and the first-year participants (MD = 1.94%). The smallest gap was at the 10,000-word level between the third-year and

the fourth-year group (MD = 0.02%). Even though the ANOVA analysis showed that there were some significant differences in scores between some year groups, the effect sizes of these differences did not exceed Cohen's (1988) convention of a small effect (0.20).

In conclusion, frequency levels, years of study, register, learning context (the university level), word encounters and chance of writing words also affected the participants' free productive vocabulary knowledge, but only to a trivial degree. More pieces of compositions from the participants might lead to more striking findings.

Additionally, for the consistency of the analysis, I recommend future research (comparing receptive and productive vocabularies) to use the methodology of scoring the correctness of usage and to profile produced words into the same frequency levels as in the receptive and the controlled productive tests so that the findings from the three tests can be comparable.

11.3.4 Relationship among receptive, controlled productive and free productive vocabulary knowledge

This section discusses the overall findings about the relationship among receptive, controlled productive, and free productive vocabularies using the quantitative analysis with the dichotomous and the developmental approaches, together with using the qualitative analysis. The discussion involves possible explanations, pedagogical implication, limitations of the methodology, and suggestions for future research.

Quantitative analysis within the dichotomous approach

The quantitative analysis within the dichotomous framing compares the sizes of receptive, controlled productive, and free productive vocabularies. The analysis showed that receptive vocabulary was the largest, followed by controlled productive vocabulary and free productive vocabulary.

Shin et al. (2011) write:

> "This pattern…is not surprising since words are easier to access receptively than to use productively, and being able to use words productively involves many complex aspects of knowing words" (Shin et al., 2011, p. 133).

The three types of vocabulary correlated with one another. The strongest correlation was between receptive vocabulary and controlled productive vocabulary and the

highest correlation was at the academic word level. The weakest one was between controlled productive vocabulary and free productive vocabulary, with the significant and the highest correlation at the academic word level as well. This might be because they participants needed to read a lot of academic texts and write academic assignments. Hence, both the receptive and the productive vocabulary knowledge at the academic level was then acquired the most. When the receptive vocabulary was acquired the controlled productive vocabulary and the free productive vocabulary would be acquired as well.

Additionally, within this approach, ratios among the three types of vocabulary were examined. The underlying idea for this is receptive vocabulary becoming productive vocabulary even though the point of the transition is unknown. For example, the controlled productive/receptive ratio refers to the percentage of chance that a participant's receptive vocabulary would become controlled productive vocabulary. The controlled productive/receptive ratio of 60% means 60% of the words known receptively tend to be used when forced. The idea of transition sounds more like the idea of continuum in the developmental approach. This has been examined by research studies in this approach, though. I discuss the relationship among the three groups of vocabulary later in this section.

I examined the three ratios (the controlled productive/receptive ratios, the free productive/receptive ratios, and the free productive/controlled productive ratios). The findings showed that the controlled productive/receptive ratios were the highest among the three types of ratios and the free productive/receptive ratios were the lowest. All the three types of ratios were high at higher frequency levels and at the academic word level. This, again, might be because the words at these levels had been exposed to frequently. The participants then implicitly or explicitly acquired them and tended to use them more. Pedagogically, more exposure to target words embedded in context might be beneficial the teaching and learning. It can help improve how learners use target words.

The high controlled productive/receptive ratios at high-frequency levels are consistent to those in many studies, both cross-sectional studies and longitudinal studies (e.g., Amin, 2020; Ozturk, 2015; Zhong and Hirsh, 2009; Shin et al., 2011). In Ozturk's (2015) study, the controlled productive/receptive ratios at the five levels (the 2,000-, the 3,000-, the 5,000-, and the 10,000-word levels and the academic word level) of the first-year participants in a cross-sectional study were 79%, 50%, 43%, 38%, and 47%, respectively. Those in her longitudinal study were 89%, 55%, 48%, 45%, and 54% (Ozturk, 2015). The ratios for the 11th graders in Zhong and Hirsh (2009) were in the same pattern. The ratios at the 2,000-, the 3,000-, the 5,000-word

levels and the academic word level were 97%, 58%, 56% and 66%. These showed that word frequency levels had an impact on chance of receptive words becoming controlled productive words.

In my study, I also found that the first-year participants had the highest controlled productive/receptive ratios at the 5,000- and the 10,000-word levels, compared to the other three year cohorts. This might be because both of the scores in the receptive test and the controlled productive test were relatively low. Therefore, the gap was not wide.

Despite this common pattern, some controlled productive/receptive ratios at some frequency levels did not always behave in that way. In Shin et al.'s (2011) study, the ratios at the 6000- and the 8000-word levels (40% and 19%, respectively) were higher than those at the 5,000- and the 7000-word levels (27% and 17%), respectively. This is because an unexpected rise in controlled productive scores at the 6,000-word level and the low scores in both receptive and controlled productive tests at the 8,000-word level. The gaps then became closer, and the ratios were then unexpectedly higher. However, the researchers did not mention further the possible explanations of these unexpected scores.

This is similar to some controlled productive/receptive ratios in some studies (e.g., Amin, 2020; Laufer, 1988; Webb, 2008). In Webb's study, the controlled productive/receptive ratio was 94% at the third band (3401st – 6600th words from COBUILD corpus), higher than the ratio (91%) at the second band (1901st – 3400th words from COBUILD corpus). This happened when he scored the receptive and the controlled productive tests with a sensitive scoring method but not when he scored them with a strict scoring method. Hence, scoring methods affect the difference between them.

In Amin's (2020) study, the controlled productive/receptive ratio at the 2,000-word level (54%) was lower than that at the 3,000-word level (81%), possibly because Amin used different tests for receptive vocabulary (the VST) and controlled productive vocabulary (the PVLT); the two tests were based on different word lists. I, therefore, suggest the future research to use the vocabulary tests that were based on the same word lists.

These overall findings of my study were compared with those of other previous studies in the field. However, only the controlled productive/receptive ratios could be compared with the ratios presented in other studies because most of previous studies compared only receptive vocabulary to controlled productive vocabulary (e.g., AbManan et al., 2017; Ozturk, 2015; Waring, 1997a; Zhong and Hirsh, 2009). Also, previous studies comparing the three groups of vocabulary (e.g., Laufer, 1998;

Laufer and Paribakht, 1998) did not grade them with the same scoring system. Therefore, the findings of free productive vocabulary were not comparable to them. I then suggest the research in the future adopt the analysing method (sorting and grading the correctness of word use in the free productive test) as conducted in the present study. It is for the consistency of the research methodology and for the comparability of the findings.

Additionally, the findings in my study showed that the ratios tended to be higher for the higher year group. The gap between the receptive and the controlled productive vocabularies was smaller. Cross-sectionally, this implied that the controlled productive vocabulary grew at a higher rate than the receptive vocabulary and its level came closer to the level of receptive vocabulary. These findings also support the findings of other studies (e.g., Waring, 1997a; Zhong and Hirsh, 2009; Ozturk, 2015).

Ozturk (2015) also found this pattern when she investigated the growth of university learners' receptive and controlled productive vocabularies both cross-sectionally and longitudinally. When she cross-sectionally examined this, the controlled productive/receptive ratios at the five frequency levels were 87%, 55%, 51%, 39%, and 49%, while the ratios of the first-year students in the longitudinal study three years later were also higher: 93%, 70%, 55%, 63%, and 66%. Hence, the ratios appeared to be related to years of study. For vocabulary assessment, this pattern confirms the validity of the receptive and the controlled productive vocabulary tests. Hence, these two tests can be used for pedagogical purposes.

The findings from longitudinal and classroom experimental studies by Vincy (2020) and Oberg (2012) confirm that the more learners were exposed to the words, the higher receptive and controlled productive scores they got. The receptive vocabulary score nearly reached the maximum. Therefore, there was not much room for improvement of the receptive vocabulary. With more room for improvement, the controlled productive vocabulary then increased at a higher rate. The controlled productive/receptive ratios kept increasing after each treatment. This implied that the gap between the two groups of vocabulary became narrower.

Moreover, that the controlled productive vocabulary showed the biggest gap might be related to that the learners needed to write a lot during their study at the university. Therefore, more vocabulary drills in a sentence level and other writing assignments might help improve learners' controlled productive vocabulary knowledge.

Quantitative analysis within the developmental approach

The quantitative analysis within the developmental approach examined the percentages of receptive and controlled productive vocabulary knowledge at the five different scores:

- score of 1, referring to "no knowledge of the lemmas at such frequency level"
- score of 2, referring to "recognising the written form of the lemma"
- score of 3, referring to "recognising the written form of the lemma and know the lemma's meaning"
- score of 4, referring to "recognising the written form, know the lemma's meaning, and being able to use it with semantic appropriateness in a sentence"
- score of 5, referring to "recognising the written form, know the lemma's meaning, and being able to use it with semantic appropriateness and grammatical accuracy in a sentence"

The analysis showed that most of the vocabulary at higher frequency levels was known at a score of 5. The percentages of the vocabulary at the 2,000-word level were higher than those at the 3,000-word level. Fewer words were known at the 5,000-word level; the majority of them were known at score of 2. Most of the lemmas at the 10,000-word level was known at either score of 1 or 2. Most of the academic lemmas were known at score of 5 for all the year cohorts, except the second-year group, knowing most of them at score of 3. Vocabulary knowledge moves forward to the productive states when the frequency level was high. Academic words can be grouped as high-frequency words in academic registers. The findings from the analysis also confirmed that frequency levels affected vocabulary knowledge.

Within this approach, I cannot compare the differences between the participants' receptive vocabulary and their controlled productive vocabulary at different years of study as I did within the dichotomous approach. However, the percentages at the score of 5 were higher for the higher year groups. Those at the score of 1 and 2 was lower for the higher year group. This supports the findings in the earlier approach.

Moreover, the differences of the percentages between the two continuous year groups were small. The trends of the percentages for each frequency level shifts at the score of 5 for the higher year group as well. Cross-sectionally, this can imply that way that vocabulary knowledge moved from one score to another was not an abrupt change but rather a shift for some degrees or gradually changed both forward and backward for the participants at higher years of study. That vocabulary knowledge moved forward and backward might have been a result of vocabulary retention and

attrition or the limitation of a cross-sectional study. Future research might employ a longitudinal research design to investigate vocabulary acquisition as well as attrition.

Qualitative analysis

After quantitatively analysing the data within both the dichotomous and the developmental approaches, I found that the findings did not clearly show the relationship among receptive, controlled productive, and free productive vocabularies. I then analysed the data from the three tests qualitatively as well.

When qualitatively analysing the data from the three tests, I found that the participants' answers in the three tests involved many aspects of vocabulary knowledge even though my instruments did not track all aspects of word knowledge.

I noticed that, out of the lemmas that each of the participants produced in their free writings, a few lemmas were used incorrectly. Most of these lemmas were used with correct meaning, correct grammatical functions and correct derived forms but with incorrect collocates. Many participants knew the meaning of a lemma, they could recall the lemma in the free productive task but used it with at least one inappropriate aspect. This can be interpreted that the students must have had receptive knowledge of the lemmas before they could recall and used in their free writings. It also implied that the students preferred to use the lemmas they were confident when using in sentences.

In this way, it is then impossible to completely and neatly categorise words into either receptive vocabulary or productive vocabulary. This is because some aspects of the word are at the precise productive stage, some are at the partial productive stage, some aspects are still at the receptive stage and some aspects have not even been acquired. However, the three tests in my study did not focus on exploring the knowledge of different word aspects, and this needs to be explored in more detail by research in the future.

Regarding conceptualising vocabulary knowledge, when scholars divide words into either receptive or productive vocabularies, they seem to divide them by some aspects only, for example, by the knowledge of form and meaning, as in Webb (2008). In that way, it is certainly clear-cut. That is, if a participant knows the meaning of a word, that word is part of his/her receptive vocabulary. If this word can be recalled, this word is then categorised as part of his/her productive vocabulary. However, if other aspects are involved when such lemma is graded, the categorisation might not be that simple.

Therefore, I suggest that each word that a learner acquires is stored as one item, but its receptive and productive uses may develop to different degrees at different times. Therefore, it is difficult to define when and where should be the dividing line between receptive vocabulary and productive vocabulary. The knowledge of vocabulary aspects moves along the cline. It develops and retreats as we can see in studies on vocabulary acquisition and attrition (e.g., Dabaghi and Rafiee, 2012; File and Adams, 2010).

11.4 The findings and the underlying concepts of vocabulary knowledge

As the present study did not assume the relationship among receptive, controlled productive, and free productive vocabulary, the data was then quantitatively analysed within two approaches: the dichotomous approach and the developmental approach. However, when the relationship was still not clear after the two ways of analyses, the framework approach was also used at a later stage. This section presents how the findings support each of the underlying concepts of vocabulary knowledge.

Within the dichotomous approach, the receptive test and the controlled productive test provided the estimate of the sizes of receptive and controlled productive vocabulary. The size of the free productive vocabulary cannot be measured from the free productive test in the present study. However, the trends of the free productive vocabulary at the five frequency levels and at the four year groups can be compared with those for the receptive vocabulary and the controlled productive vocabulary. These three groups of vocabulary were affected by frequency levels and years of study. They tended to be high at the high-frequency levels and the academic level, and low at the low-frequency level. The data showed that the scores of the receptive vocabulary was always the highest, followed by those of the controlled productive vocabulary and the free productive vocabulary. Also, they were all high for the higher year group and low for the lower year group.

The sizes of the receptive vocabulary and the controlled productive vocabulary were compared, in terms of the differences and the ratios. The findings showed that the participants' receptive vocabulary was always larger than their controlled productive vocabulary. Regardless the frequency levels, the controlled productive vocabulary of the fourth year cohort and the first year cohort showed the biggest gap. If these two groups of vocabulary are competitors, the controlled productive vocabulary might

start later than the receptive vocabulary, but then it grew faster after the receptive vocabulary is large enough or reaches the mastery level. However, when examining the sizes at each of the five frequency levels, the gap of the receptive vocabulary between the year groups at the 5,000- and the 10,000-word levels was wider than that of the controlled productive vocabulary. Cross-sectionally, the receptive vocabulary developed faster, and the controlled productive vocabulary lagged behind. The scores of the free productive vocabulary at each frequency level were roughly equal, but the scores at the 2,000-word level and at the academic word level were high. Hence, the trends were similar to those of the receptive and the control productive vocabulary.

By using these tests and the scoring systems, the controlled productive/receptive ratios can be calculated as well. This is what studies within the dichotomous approach usually do. The ratios can be interpreted in two ways. First, the ratios refer to gaps. Then, the gaps between receptive and controlled productive vocabulary were small at the high-frequency levels. If the ratios refer to the chance of receptive vocabulary becoming controlled productive vocabulary, there was a high chance at the high-frequency levels. The chance is high for academic words too. However, this idea sounds like the idea within the developmental (or continuum) approach, with an unclear diving line. Accordingly, the findings within the dichotomous approach support the idea that receptive vocabulary and productive vocabulary are separate groups of words. They compete each other, with the receptive vocabulary was always the largest.

Within the developmental approach, the same data was analysed by using the developmental scale. The findings can again support the idea that the vocabulary knowledge was moving on a continuum, with one end is the receptive stage and the other end is the productive stage. Regardless of years of study, the vocabulary scores at the 2,000-, the 3,000-word level and the academic word level of the participants were at the scores of 4 and 5, and the percentages at the scores of 4 and 5 were higher for the higher year group. The scores of 1 and 2 and 3 was lower and they were higher at the score of 5. This was not an abrupt shift but the percentages at the scores of 1, 2, and 3 were not completely zero. At the mid- and the low-frequency levels, the scores of 5 were still higher for the higher year group but the majority of the scores were at the scores of 1 or 2. So, the scores tended to move towards the end of the productive end of the continuum for the higher year group. This implied that learners developed the ability to comprehend words and use words with correct meaning and grammar when they were in higher years of study or when they had been exposed to words frequently.

The findings showed that vocabulary knowledge gradually moved or shifted from one stage to another. The high scores for the higher year groups were not much higher than those for the lower year groups, but the scores were slightly higher from Year 1 to Year 2, Year 2 to Year 3, and Year 3 to Year 4. This supports the idea that vocabulary development is incremental, and the vocabulary knowledge seems to move towards the productive end of the continuum. It can also shift back to the receptive end, also. However, I acknowledged that the data was analysed cross-sectionally and a longitudinal study will be able to reveal how vocabulary knowledge shifts better.

The same data can be analysed within these two approaches and the findings then support these two ideas. The findings do not confirm which is the relationship between receptive and productive vocabulary. It might be because the data was analysed by using major categories of vocabulary knowledge. Accordingly, I attempted to qualitatively analysed it by using Nation's framework of vocabulary knowledge aspects. Aspects of an individual word was the focus. I acknowledged that my study was not a multicomponent study. The tests used in the present study cannot tap onto all the aspects involved in knowing a word. They allowed the examinations of some aspects, though. The findings showed that different aspects developed at different stages, from no knowledge to precise knowledge and from receptive to productive knowledge as explained in Section 11.3.4. One aspect can be at a precise productive stage, some can be at a partial receptive stage, and the others might not develop at all.

In conclusion, if one or two aspects such as form and meaning link are involved when making a decision whether an individual word is receptive vocabulary or productive vocabulary, it is rather quick and easy to classify them. However, when considering all the aspects for such word, it will be much harder and much more challenging for researchers to classify them into either receptive, controlled productive, or free productive vocabulary. As individual words carry many aspects or what Jiang call specifications with them, it is unavoidable and unfair to completely consider at one aspect only when classifying them. Therefore, the findings from these analyses support that vocabulary knowledge incrementally developed and the aspects move along a continuum, including no knowledge, partial knowledge, and precise knowledge. The aspects move forward and backward at different stages. It is hard to define the dividing line, where receptive vocabulary becomes productive vocabulary, because many aspects are actually involved. If a dividing line does exist, there might be too many dividing lines for different aspects and for different groups of words and further research investigating these dividing lines.

11.5 Pedagogical implications

The section suggests some pedagogical implications. According to the findings from all the three tests, the participants regardless of their years of study seemed to know and be able to use a lot more words at high-frequency levels than those at low-frequency levels. The possible explanation for this is that the learners have been exposed to high-frequency words many times and have become more familiar with them. They have opportunities to use those words for their class assignments, so they used them with confidence and gained high scores in both the productive tests. The scores for the academic words were relatively high or approximately as high as those for the 3,000-word level. This might be because learning at a university level required the participants to read and write various texts in an academic genre.

Consequently, encountering words for multiple times is an important factor of vocabulary knowledge development. Pedagogically, teachers might need to expose their students to some useful words. They can be high-frequency words, academic words, register-based words, or specialised words for their fields of study. If being exposed and practiced much enough, any words will become part of learners' mental lexicon and will be used receptively and productively with fluency and confidence. Therefore, it is recommended that teachers select the group of words that might be beneficial for the learners in such fields and let them drill those words.

EFL contexts, however, do not provide much chance for English learners to use the English language and a few hours of English class also one obstacle for acquiring and practicing English words. Using English as a medium in English class is then ideal because this supports the exposure to English.

For high-school teachers, teaching a lot of academic words might be able to prepare high-school students for learning at a university level. This specially benefits the students who would like to major in English or to enter any programmes or faculties where English is used as a medium in class. Using English as a medium still allows focusing on content, implicit vocabulary learning, as well as explicitly teaching useful words.

For high-school students who do not aim for an English major or for English-programme university, teachers can still use English as a medium in class for the learners' exposure. Interesting and authentic texts might help maintain their interests and make them understand more about the necessity of using the English language.

In this way, they can have goals for learning the language and will then learn more words as a result.

Because the scores of the three tests showed similar findings to those of previous studies, both cross-sectional and longitudinal studies, the three tests can be used as a diagnostic test to examine which group of vocabulary needs more attention and practice. It can also be used as a quick placement test to roughly divide students into groups based on their vocabulary knowledge so that vocabulary or language tasks can be best selected to suit learners' levels.

Also, teachers who aim to track their learners' knowledge development of some target words can also use the receptive and the controlled productive test used in the present study because the tests can tap onto small shifts from one stage to another. These shifts roughly show at which stage (form, meaning, or use) the learners and the teachers should improve so that the learners' knowledge of such words can be pushed towards the productive end of the continuum.

11.6 Limitations of the present study

This section concludes the limitations of the present research study. This study aimed to explore the English major students' receptive and productive vocabulary knowledge at different frequency levels and at different years of study. As collecting the data for four years was not impossible for this study, the research design was then a cross-sectional study. As a cross-sectional study, it is recommended by Cohen et al. (2017) that there be at least 30 participants per variable. In this study, variables were the four year cohorts. However, there were not enough participants from some year groups: 34 first-year students, 24 second-year students, 29 third-year students, and 24 fourth-year students. I then grouped the first- and the second-year groups as the lower year group and the third- and the fourth-year groups as the higher year group for more reliable statistical results. When I compared the vocabulary scores of the four year groups, the effect sizes were not large, though.

Moreover, as a cross-sectional study, the differences of the vocabulary scores cannot completely represent the vocabulary growth over the years of study. Likewise, the higher scores and the lower scores for the higher year group need to be carefully interpreted as the increase or decrease of vocabulary knowledge throughout their study at the university.

Another limitation of the present study is using test items from old wordlists as used in the VLT (Schmitt et al., 2001). As these test items have been used by many

previous studies from various EFL contexts and the VLT has been proved to be valid and reliable, these test items were then used in this study for the comparability. However, this affected the participants' vocabulary knowledge slightly as the knowledge of a few of the test words such as 'lieutenant' did not behave like the others at the same frequency level.

In terms of the number of test words at each frequency level and the number of frequency levels included, the current study included only ten words or ten lemmas at each of the five lists (the 2,000-, the 3,000-, the 5,000-, the 10,000-word levels, and the academic word level). The vocabulary scores then cannot represent the vocabulary knowledge at the 4,000-, the 6,000-, or the 9,000-word levels. Only roughly estimation can be done. This was because the test battery was relatively demanding and ten words for each level were tested to avoid the participants' fatigue effect.

For test formats, despite many vocabulary knowledge aspects, the test formats used in this study cannot tap onto all the aspects at once. However, the tests can still cover all the three main categories: form, meaning, and use as suggested by Nation (2013). The receptive test assesses the knowledge of form and meaning, or the ability to recognise an L2 word form and to supply its meaning. The controlled productive test tests the ability to use a provided L2 form in a sentence level. The controlled productive vocabulary test in this study does not elicit the same ability as the PLVT, a controlled productive test with a cue-recall format, by Laufer and Nation (1999), though. The free productive test examines the ability to recall an appropriate word that matches the concept or the idea that a participant wants to express and the ability to use that word form with a correct meaning and a correct grammar in the form of an essay. The two productive vocabulary tests include the knowledge of form, meaning, and use. They were not able to explore specific aspects thoroughly.

Additionally, while the scores from the first two tests can imply the participants' receptive and controlled productive vocabulary sizes, the free productive test cannot serve as a measure of free productive vocabulary size. This is because the test cannot examine words that the participants knew but did not produce in their essays. Despite this, the three tests to some extent show the overall trends of the participants' receptive, controlled productive, and free productive vocabulary knowledge in relation to word frequency levels and years of study.

In the process of grading the participants' answers, Paribakht and Wesche's (1997) scoring scale was adopted in the present study. Once a participant supplied an incorrect meaning for a provided word, no point was given to that answer on the receptive test. However, when I qualitatively examined the answers from the

receptive test and the controlled productive test, I found in one example that the meaning used in the sentence was actually correct but the translation was incorrect. Therefore, only one verification for the receptive test might not be completely relied on and double checking with the sentence supplied by the participants needed to be checked for correctness.

Furthermore, this study focused on comparing the scores of the three groups of vocabulary at different frequency levels and at different years of study. The study did not focus on reasons why the scores of one year group was lower than those of the other group. Some information from the questionnaire about their educational background and their exposure to English outside class can be used to explain some unexpected findings. Nevertheless, there might be some other factors involved and this study did not cover all those factors.

11.7 Suggestions for future research

This section presents some suggestions for future research in the field. Regarding the research design, if the research design needs to be a cross-sectional study, recruiting at least 30 participants per variable might provide more solid statistical results. If the aim of the research is to track the vocabulary knowledge over the four years of study and time is not constraint, then a longitudinal design is strongly recommended. It can track the vocabulary knowledge of the same participants and it can track the knowledge of the same tested words. In this way, the increase or the decrease of the vocabulary knowledge can definitely represent the learners' vocabulary development and the vocabulary attrition.

In vocabulary assessment, test items are also important. Sampling words from more recent wordlists might reflect the participants at the time better than those from old wordlist. Therefore, researchers might consider using recent wordlists like BNC/COCA (Nation, 2017) in their research in the future. Importantly, the innovations of using the same wordlists and the same scoring methods for all the three tests in this study are worth replicating for the consistency of the analysing methods.

Also, to better represent the knowledge of vocabulary, testing more than ten test items and including more than these five frequency levels are recommended if the test battery is not demanding as this might help researchers to have a clearer picture of vocabulary as a whole. Devising vocabulary tests that are able to include more test items is also needed.

For checking the correctness of translations and word use, it is better to use a three-scale scoring system or a multi-scale scoring system that allows partial knowledge. Knowing a word is not an all-or-nothing phenomenon (Henriksen, 1999). Partially correct answers should be considered as part of someone's vocabulary knowledge.

Regarding the test formats, as they cannot explore specific aspects of vocabulary knowledge, the test formats in this study might not be suitable for the study which aims to explore each aspect in detail. To examine vocabulary knowledge as a whole, it is inevitable to explore each aspect and to understand its smaller parts. Hence, there should also be more studies focussing on vocabulary aspects (multicomponent studies) and selecting test formats suitable for doing so is necessary. A free productive test that can serve as a free productive vocabulary size test is still needed.

As to the use of terms, while the 'controlled productive' test in this study does not involve the ability to recall a word but does include the ability to use a word in a sentence, a 'controlled productive' test in Laufer and Nation's (1999) study includes the ability to recall an L2 word that fit the meaning in the context and the form hinted. These are two different abilities but are called the same. Theoretically, I, therefore, suggest that a new term be coined for one of these two abilities to differentiate them and to avoid confusion.

Concerning grading the receptive vocabulary test, researchers should not assume that participants did not know the meaning of a provided word from only one translation task. Double checking the meaning in the supplied sentence and the addition of one more verification are required for the receptive test.

For grading the free productive test, the percentages of the lemmas used at each frequency level in this study were calculated from the number of all the lemmas the participants produced in one essay. Even though the length of the essays did not seem to affect the participants' free productive vocabulary scores, future research might try assessing only the first 300 tokens of each essay for each of the participants in order to examine whether the scores between year groups would support the findings of the present study or behave differently. Also, when comparing the vocabulary scores between groups, the same test formats used for one group should be the same for the other groups for comparability.

Moreover, this study did not focus on factors behind some unexpected findings such as that some higher year groups scored lower than the lower year group at some frequency levels and that few words were used incorrectly in the free productive test. For the sake of pedagogy, it is recommended that an interview or a more detailed questionnaire that can include more factors regarding vocabulary learning might help

explain these better and should be included in future research. The information can probably show ways of learning vocabulary that is best suitable for such group of learners and finally help improve English vocabulary teaching and learning. It might also reveal the underlying facts about learners' receptive and productive vocabulary knowledge.

Finally, as vocabulary are words and there are many words in a language. Examining vocabulary from a bigger view is important, so studies on vocabulary breadth is important. However, checking it from a smaller view is also needed as knowing a word involves knowing many vocabulary aspects. Hence, to understand more about vocabulary knowledge, studies on the depth of vocabulary or multicomponent studies are needed. We now know from previous research (e.g., González-Fernández and Schmitt, 2020; Schmitt, 1998; Webb, 2007b) that aspects are interrelated, but there is still so much more to explore. Moreover, a language consists of both single words and multiword units, it might be beneficial for future research to investigate the aspects of multiword units as well.

References

AbManan, N.A., Azizan, N. and Nasir, N.F.W.M. 2017. Receptive and Productive Vocabulary Level of Diploma Students from a Public University in Malaysia. *J. Appl. Environ. Biol. Sci.* **7**(1S), pp.53-59.

Aitchison, J. 1994. Say, say it again Sam": The treatment of repetition in linguistics. *Repetition.* **7**, p15.

Aizawa, K., Ochiai, N. and Osaki, S. 2003. The Effects of Teaching on Vocabulary Knowledge : Receptive vs. Productive. 全国英語教育学会紀要. **14**, pp.151-160.

Akbarian, I.h. 2010. The relationship between vocabulary size and depth for ESP/EAP learners. *System.* **38**(3), pp.391-401.

Al-Masrai, A. 2009. Measuring the English vocabulary size of Saudi university students: Validating a new 10,000 word vocabulary size test. *Unpublished Master's Thesis, Swansea University, Swansea, UK.*

Al-Masrai, A. and Milton, J. 2012. The Vocabulary Knowledge of University Students in Saudi Arabia. *Perspectives (TESOL Arabia).* **19**(3).

Alonso, A.C. and Fontecha, A.F. 2014. Motivation and L2 Receptive Vocabulary Knowledge of Spanish EFL Students at the Official School of Languages. *Miscelánea: A Journal of English and American Studies.* **49**.

Alonso, A.C. and Garcia, M.A. 2014. Productive vocabulary knowledge of Spanish EFL learners. *RAEL: revista electrónica de lingüística aplicada.* **13**(1), pp.39-56.

Amin, M.A. 2020. The Comparison of Receptive and Productive Vocabulary Size of Afghan Tertiary Students. *Vocabulary in Curriculum Planning.* Springer, pp.29-55.

Amy, S. 2006. A study of the English vocabulary knowledge of university students in Hong Kong. *Asian Journal of English Language Teaching.* **16**, pp.1-23.

Aviad-Levitzky, T., Laufer, B. and Goldstein, Z. 2019. The new computer adaptive test of size and strength (CATSS): Development and validation. *Language Assessment Quarterly.* **16**(3), pp.345-368.

Azodi, N., Karimi, F. and Vaezi, R. 2014. Measuring the Lexical Richness of Productive Vocabulary in Iranian EFL University Students' Writing Performance. *Theory and Practice in Language Studies.* **4**(9), p1837.

Bahramlou, K. and Esmaeili, A. 2019. The Effects of Vocabulary Enhancement Exercises and Group Dynamic Assessment on Word Learning Through Lexical Inferencing. *Journal of Psycholinguistic Research.* pp.1-13.

Bakla, A. and Çekic, A. 2017. Using an Online Vocabulary Memorization Tool versus Traditional Vocabulary Exercises. *Ana Dili Eğitimi Dergisi.* **5**(4), pp.948-966.

Bao, G. 2015. Task type effects on English as a Foreign Language learners' acquisition of receptive and productive vocabulary knowledge. *System.* **53**, pp.84-95.

Barcroft, J. 2004. Second language vocabulary acquisition: A lexical input processing approach. *Foreign Language Annals.* **37**(2), pp.200-208.

Barrow, J., Nakanishi, Y. and Ishino, H. 1999. Assessing Japanese college students' vocabulary knowledge with a self-checking familiarity survey. *System (Linköping).* **27**(2), pp.223-247.

Bauer, L. and Nation, I.S.P. 1993. Word Families. *International Journal of Lexicography.* **6**(4), pp.253-279.

Bayazidi, A. and Saeb, F. 2017. Assessing Reliability of Two Versions of Vocabulary Levels Tests in Iranian Context. *Advances in Language and Literary Studies.* **8**(1), pp.30-43.

Beaton, A., Gruneberg, M., Hyde, C., Shufflebottom, A. and Sykes, R. 2005. Facilitation of receptive and productive foreign vocabulary learning using the keyword method: The role of image quality. *Memory.* **13**(5), pp.458-471.

Beglar, D. and Hunt, A. 1999. Revising and validating the 2000 word level and university word level vocabulary tests. *Language testing.* **16**(2), pp.131-162.

Cambridge Dictionary (English-Thai Dictionary). 2020. [Online]. [Accessed 1 January 2019]. Available from: https://dictionary.cambridge.org/dictionary/english-thai/

Cameron, L. 2002. Measuring vocabulary size in English as an additional language. *Language Teaching Research.* **6**(2), pp.145-173.

Caspi, T. and Lowie, W. 2013. The Dynamics of L2 Vocabulary Development: A Case Study of Receptive and Productive Knowledge. *Revista Brasileira de Linguística Aplicada.* **13**(2), p437.

Castro-García, D. 2017. Receptive vocabulary measures for EFL Costa Rican high school students. *International Journal of English Studies.* **17**(2), pp.81-99.

Catalán, R.M.J. and De Zarobe, Y.R. 2009. The receptive vocabulary of EFL learners in two instructional contexts: CLIL versus non-CLIL instruction. *Content and language integrated learning: Evidence from research in Europe.* pp.81-92.

Catalán, R.M.J., de Zarobe, Y.R. and Iragui, J.C. 2006. Vocabulary profiles of English foreign language learners in English as a subject and as a vehicular language. *Current Research on CLIL.* p23.

Catalan, R.M.J. and Gallego, M.T. 2005. The Receptive Vocabulary Of English Foreign Language Young Learners. *Journal of English Studies.* **5-6**(5-6), pp.173-191.

Channell, J. 1988. Psycholinguistic considerations in the study of L2 vocabulary acquisition. *Vocabulary and language teaching.* pp.83-96.

Cobb, T. 2017. *Web Vocabprofile, an adaptation of Heatley, Nation & Coxhead's (2002) Range.*

Cobb, T. n.d. Web Vocabprofile [accessed 13 June 2016 from http://www.lextutor.ca/vp/], an adaptation of Heatley. *Nation, & Coxhead's.*

Cohen, J. 1988. *Statistical power analysis for the behavioral sciences.* Second edition. ed. Hillsdale, N.J: L. Erlbaum Associates.

Cohen, L., Manion, L. and Morrison, K. 2007. *Research Methods in Education 6th Edition.* USA and Canada: Routledge.

Cohen, L., Manion, L. and Morrison, K. 2011. *Research methods in education.* 7th ed. London: Routledge.

Cohen, L., Manion, L. and Morrison, K. 2017. *Research methods in education.*

Corson, D. 1983. The Corson measure of passive vocabulary. *Language and Speech.* **26**(1), pp.3-20.

Coxhead, A. 2000. A New Academic Word List. *TESOL Quarterly.* **34**(2), pp.213-238.

Cronbach, L.J. 1942. An analysis of techniques for diagnostic vocabulary testing. *The journal of educational research.* **36**(3), pp.206-217.

Crow, J.T. 1986. Receptive vocabulary acquisition for reading comprehension. *The Modern Language Journal.* **70**(3), pp.242-250.

Dabaghi, A. and Rafiee, M. 2012. Incidental Vocabulary Learning and the Development of Receptive and Productive Vocabulary: How Gloss Types Work. *Suvremena lingvistika.* **38**(74), pp.175-187.

Deese, J. 1965. The Structure of Associations in Language and Thought. *Associations in Language and Thought.*

Dizon, G. and Tang, D. 2017. Comparing the Efficacy of Digital Flashcards versus Paper Flashcards to Improve Receptive and Productive L2 Vocabulary. *The EuroCALL Review.* **25**(1), pp.3-15.

Djiwandono, P.I. 2016. Lexical richness in academic papers: a comparison between students' and lecturers' essays. *Indonesian Journal of Applied Linguistics.* **5**(2), pp.209-216.

Dolch, E.W. and Leeds, D. 1953. Vocabulary tests and depth of meaning. *The Journal of Educational Research.* **47**(3), pp.181-190.

Dörnyei, Z. and Katona, L. 1992. Validation of the C-test amongst Hungarian EFL learners. *Language Testing.* **9**(2), pp.187-206.

Dubiner, D. 2017. Using vocabulary notebooks for vocabulary acquisition and teaching. *ELT JOURNAL.* **71**(4), pp.456-466.

Færch, C., Haastrup, K. and Phillipson, R. 1984. *Learner language and language learning.* Multilingual Matters Limited.

Fan, M. 2000. How big is the gap and how to narrow it? An investigation into the active and passive vocabulary knowledge of L2 learners. *RELC Journal.* **31**(2), pp.105-119.

File, K.A. and Adams, R. 2010. Should vocabulary instruction be integrated or isolated? *Tesol Quarterly.* **44**(2), pp.222-249.

Fitzpatrick, T., Al-Qarni, I. and Meara, P. 2008. Intensive vocabulary learning: a case study. *The Language Learning Journal.* **36**(2), pp.239-248.

Fitzpatrick, T. and Clenton, J. 2017. Making Sense of Learner Performance on Tests of Productive Vocabulary Knowledge. *TESOL Quarterly.* **51**(4), pp.844-867.

Fontecha, A.F. 2014a. Motivation and gender-related variation in EFL receptive vocabulary acquisition in Primary Education.

Fontecha, A.F. 2014b. Receptive vocabulary knowledge and motivation in CLIL and EFL. *Revista de lingüística y lenguas aplicadas.* **9**, pp.23-32.

Gallego, M.T. and Llach, M.d.P.A. 2009. Exploring the increase of receptive vocabulary knowledge in the foreign language: A longitudinal study. *International Journal of English Studies.* **9**(1).

González-Fernández, B. and Schmitt, N. 2020. Word knowledge: Exploring the relationships and order of acquisition of vocabulary knowledge components. *Applied Linguistics.* **41**(4), pp.481-505.

Gyllstad, H., Vilkaitė, L. and Schmitt, N. 2015. Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates. *ITL-International Journal of Applied Linguistics.* **166**(2), pp.278-306.

Hashemi, M.R. and Gowdasiaei, F. 2005. An attribute-treatment interaction study: Lexical-set versus semantically-unrelated vocabulary instruction. *RELC journal.* **36**(3), pp.341-361.

Hatano, F. 2008. A Study of the Development of Productive Vocabulary : Research on Written Production of University Students. *関東甲信越英語教育学会研究紀要.* **22**, pp.47-58.

Hatch, E.M. and Lazaraton, A. 1991. *The research manual: Design and statistics for applied linguistics.* New York, NY: Newbury House Publishers.

Hazenberg, S. and Hulstijn, J.H. 1996. Defining a minimal receptive second-language vocabulary for non-native university students: An empirical investigation. *Applied linguistics.* **17**(2), pp.145-163.

Heidari, K. 2019. Willingness to Communicate: A Predictor of Pushing Vocabulary Knowledge from Receptive to Productive. *Journal of Psycholinguistic Research.* pp.1-18.

Helms-Park, R. and Perhan, Z. 2016. The role of explicit instruction in cross-script cognate recognition: The case of Ukrainian-speaking EAP learners. *Journal of English for Academic Purposes.* **21**, pp.17-33.

Henriksen, B. 1999. The Dimensions of Vocabulary Development. *Studies in Second Language Acquisition.* **21**(2), p303.

Hirsh, D. and Nation, I.S.P. 1992. What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language.* **8**, pp.689-696.

Hsu, W. 2014. The effects of audiovisual support on EFL learners' productive vocabulary. *ReCALL.* **26**(1), pp.62-79.

Isaac, S. and Michael, W.B. 1995. *Handbook in research and evaluation: A collection of principles, methods, and strategies useful in the planning, design, and evaluation of studies in education and the behavioral sciences.* Edits publishers.

Ishii, T. and Schmitt, N. 2009. Developing an integrated diagnostic test of vocabulary size and depth. *RELC Journal.* **40**(1), pp.5-22.

Jiang, N. 2000. Lexical Representation and Development in a Second Language. *Applied Linguistics.* **21**(1), pp.47-77.

Johnson, M.D., Acevedo, A. and Mercado, L. 2016. Vocabulary Knowledge and Vocabulary Use in Second Language Writing. *TESOL Journal.* **7**(3), pp.700-715.

Kawauchi, C. 2010. The Effects of CALL-Alone and CALL-Plus : Developing Receptive and Productive Vocabulary. *外国語教育メディア学会機関誌.* **47**, pp.75-92.

Kilgarriff, A. 2006. *BNC database and word frequency lists. Retrieved 24 August, 2006.*

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. 2014. The Sketch Engine: ten years on. *Lexicography.* **1**(1), pp.7-36.

Kilic, M. 2019. Vocabulary Knowledge as a Predictor of Performance in Writing and Speaking: A Case of Turkish EFL Learners. *PASAA: Journal of Language Teaching and Learning in Thailand.* **57**, p133.

Kotchana, S. and Tongpoon-Patanasorn, A. 2015. EFL Learners' Vocabulary Size: A Case in the Northeastern Region of Thailand. *ASEAN Journal of Education.* **1**(1), pp.9-26.

Kremmel, B. 2016. Word families and frequency bands in vocabulary tests: Challenging conventions. *tesol QUARTERLY.* **50**(4), pp.976-987.

Kroll, J.F., Michael, E., Tokowicz, N. and Dufour, R. 2002. The development of lexical fluency in a second language. *Second language research.* **18**(2), pp.137-171.

Kučera, H. and Francis, W.N. 1967. *Computational analysis of present-day American English.* Dartmouth Publishing Group.

Laufer, B. 1988. The concept of 'synforms' (similar lexical forms) in vocabulary acquisition. *Language and Education.* **2**(2), pp.113-132.

Laufer, B. 1992. Reading in a Foreign Language: How Does L2 Lexical Knowledge Interact with the Reader's General Academic Ability? *Journal of Research in Reading.* **15**(2), pp.95-103.

Laufer, B. 1994. The lexical profile of second language writing: Does it change over time? *RELC journal.* **25**(2), pp.21-33.

Laufer, B. 1998. The development of passive and active vocabulary in a second language: Same or different? *APPLIED LINGUISTICS.* **19**(2), pp.255-271.

Laufer, B., Elder, C., Hill, K. and Congdon, P. 2004. Size and strength: do we need both to measure vocabulary knowledge? *Language Testing.* **21**(2), pp.202-226.

Laufer, B. and Goldstein, Z. 2004. Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language learning.* **54**(3), pp.399-436.

Laufer, B. and Nation, I.S.P. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics.* **16**(3), pp.307-322.

Laufer, B. and Nation, I.S.P. 1999. A vocabulary-size test of controlled productive ability. *Language testing.* **16**(1), pp.33-51.

Laufer, B. and Paribakht, T.S. 1998. The relationship between passive and active vocabularies: Effects of languagelearning context. *Language learning.* **48**(3), pp.365-391.

Lee, S.H. and Muncie, J. 2006. From Receptive to Productive: Improving ESL Learners' Use of Vocabulary in a Postreading Composition Task. *TESOL Quarterly.* **40**(2), pp.295-320.

Lemmough, Z. 2008. The Relationship between Grades and the Lexical Richness of Student Essays. *NJES: Nordic Journal of English Studies.* **7**(3), p163.

Li, L. and MacGregor, L.J. 2010. Investigating the receptive vocabulary size of university-level Chinese learners of English: how suitable is the Vocabulary Levels Test? *Language and Education.* **24**(3), pp.239-249.

Llach, M.d.P.A. 2016. Does L1 make a difference? Evidence from the receptive vocabulary size of Spanish and German EFL primary school learners. *The Language Learning Journal.* **44**(2), pp.214-235.

Longman Dictionary of English Language and Culture. 1992. s.v.

Maftoon, P. and Sharif Haratmeh, M. 2013. Effects of input and output-oriented tasks with different involvement loads on the receptive vocabulary knowledge of Iranian EFL learners. *Research in English Language Pedagogy.* **1**(1), pp.24-38.

Mahdavy, B. 2011. The role of topic familiarity and rhetorical organization of texts in L2 incidental vocabulary acquisition. *Procedia-Social and Behavioral Sciences.* **29**, pp.208-217.

Malmström, H., Pecorari, D. and Shaw, P. 2018. Words for what? Contrasting university students' receptive and productive academic vocabulary needs. *English for Specific Purposes.* **50**, pp.28-39.

Malvern, D., Richards, B., Chipere, N. and Durán, P. 2004. *Lexical diversity and language development.* Springer.

McLean, S. 2018. Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics.* **39**(6), pp.823-845.

Meara, P. 1982. Word associations in a foreign language. *Linguistics Circular.* **11**, pp.29-38.

Meara, P. 1984. Teaching vocabulary: Wallace, Michael, London: Heinemann Educational Books, 1982, 144 pp.(Practical Language Teaching No. 10). *System.* **12**(2), pp.185-186.

Meara, P. 1990a. A note on passive vocabulary. *Interlanguage studies bulletin (Utrecht).* **6**(2), pp.150-154.

Meara, P. 1990b. Some notes on the Eurocentres vocabulary tests. *Foreign language comprehension and production.* pp.103-113.

Meara, P. 1992. *EFL vocabulary tests.* Citeseer.

Meara, P. 1996a. The dimensions of lexical competence. *Performance and competence in second language acquisition.* **35**.

Meara, P. 1996b. *The third dimension of lexical competence*. Unpublished.

Meara, P. 1996c. The vocabulary knowledge framework. *Vocabulary Acquisition Research Group Virtual Library.*

Meara, P. 1997. *Towards a new approach to modeling vocabulary acquisition. InN. Schmitt & M. McCarthy (Eds.), Vocabulary: Description, Acquisition and Pedagogy (pp. 109-121)*. Cambridge: Cambridge University Press.

Meara, P. and Buxton, B. 1987. An alternative to multiple choice vocabulary tests. *Language testing.* **4**(2), pp.142-154.

Meara, P. and Fitzpatrick, T. 2000. Lex30: An improved method of assessing productive vocabulary in an L2. *System.* **28**(1), pp.19-30.

Meara, P. and Jones, G. 1988. Vocabulary size as a placement indicator.

Meara, P. and Jones, G. 1990. Eurocentres vocabulary size test 10KA. *Zurich: Eurocentres.*

Meara, P. and Milton, J. 2003. *X-lex: the Swansea levels test.* Express Publishing.

Meara, P. and Miralpeix, I. 2006. Y_Lex: The Swansea advanced vocabulary levels test. v2. 05. *Swansea: Lognostics.*

Meara, P. and Miralpeix, I. 2015. *V_YesNo v1. 0*. Citeseer.

Melka, F.J.T. 1982. Receptive versus productive vocabulary: A survey. *Interlanguage Studies Bulletin.* pp.5-33.

Melka, F.J.T. 1997. Receptive vs. productive aspects of vocabulary. *Vocabulary: Description, acquisition and pedagogy.* **33**(2), pp.84-102.

Milton, J. and Fitzpatrick, T. 2014. *Dimensions of vocabulary knowledge.* Basingstoke: Palgrave Macmillan.

Min, H.T. 2008. EFL Vocabulary Acquisition and Retention: Reading Plus Vocabulary Enhancement Activities and Narrow Reading. *Language Learning.* **58**(1), pp.73-115.

Mondria, J.-A. and Wiersma, B. 2004. Receptive, productive, and receptive+ productive L2 vocabulary learning: What difference does it make. *Vocabulary in a second language: Selection, acquisition, and testing.* **15**(1), pp.79-100.

Morgan, B. and Oberdeck, L.M. 1930. Active and passive vocabulary. *Studies in modern language teaching.* pp.213-221.

Morshedian, M. 2008. The Role of Initial English as Foreign Language Proficiency in Lexical Attrition/Retention of Iranian Learners: Is Productive or the Receptive Word Knowledge of Learned Nouns More Likely to Be Lost? *Linguistics Journal.* **3**(1).

Mousavi, F. and Gholami, J. 2014. Effects of watching flash stories with or without subtitle and reading subtitles on incidental vocabulary acquisition. *Procedia-Social and Behavioral Sciences.* **98**, pp.1273-1281.

Muncie, J. 2002. Process writing and vocabulary development: Comparing lexical frequency profiles across drafts. *System.* **30**(2), pp.225-235.

Nation, I.S.P. 1983. Learning vocabulary. *New Zealand Language Teacher.* **9**(1), pp.10-11.

Nation, I.S.P. 1984. *Vocabulary lists: Words, affixes, and stems.* English Language Institute, Victoria University of Wellington.

Nation, I.S.P. 1990. *Teaching and learning vocabulary.* Boston, Mass: Heinle.

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

Nation, I.S.P. 2004. A study of the most frequent word families in the British National Corpus. *Vocabulary in a second language: Selection, acquisition, and testing.* **10**, pp.3-13.

Nation, I.S.P. 2006. How Large a Vocabulary Is Needed for Reading and Listening? *The Canadian Modern Language Review / La revue canadienne des langues vivantes.* **63**(1), pp.59-81.

Nation, I.S.P. 2008. *Teaching vocabulary: strategies and techniques.* Boston, MA: Heinle.

Nation, I.S.P. 2010. *Analysing vocabulary teaching techniques.* Equinox.

Nation, I.S.P. 2012. *Vocabulary Size test.* [Online]. [Accessed 6 November 2017]. Available from: https://www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I.S.P. 2013. *Learning vocabulary in another language.* 2nd ed. Cambridge: Cambridge University Press.

Nation, I.S.P. 2017. *The BNC/COCA Level 6 word family lists (Version 1.0. 0)[Data file].*

Nation, I.S.P. and Beglar, D. 2007. A vocabulary size test. *The language teacher.* **31**(7), pp.9-13.

Nation, I.S.P. and Coxhead, A. 2014. Vocabulary Size Research at Victoria University of Wellington, New Zealand. *Language Teaching.* **47**(3), p398.

Nation, I.S.P. and Wang Ming-tzu, K. 1999. Graded readers and vocabulary.

Nation, I.S.P. and Waring, R. 1997. Vocabulary size, text coverage and word lists. In: Schmitt, N. and McCarthy, M. eds. *Vocabulary Decription, Acquisition and Pedagogy.* Cambridge: Cambridge University Press, pp.6-19.

Nirattisai, S. 2014. *Vocabulary size and vocabulary learning strategies of Thai university students.* thesis, Prince of Songkla University.

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

Nizonkiza, D. 2016. First-year university students' receptive and productive use of academic vocabulary. *Stellenbosch Papers in Linguistics.* **45**, pp.169-187.

Nurweni, A. and Read, J. 1999. The English Vocabulary Knowledge of Indonesian University Students. *English for Specific Purposes.* **18**(2), pp.161-175.

Oberg, A. 2012. Receptive and productive vocabulary acquisition: examining processing time and memory threshold. *Indonesian Journal of Applied Linguistics.* **2**(1), pp.23-39.

Oxford Advanced Learner's Dictionary. 1995. s.v.

Oxford Advanced Learners' Dictionary. 2020. [Online].

Ozturk, M. 2015. Vocabulary growth of the advanced EFL learner. *The Language Learning Journal.* **43**(1), pp.94-109.

Palmberg, R. 1984. Solving communicative problems in interlanguage. *Interlanguage Studies Bulletin.* pp.35-49.

Palmberg, R. 1987. Patterns of Vocabulary Development in Foreign-Language Learners1. *Studies in second language acquisition.* **9**(2), pp.201-219.

Paribakht, T.S. and Wesche, M. 1993. Reading comprehension and second language development in a comprehension-based ESL program. *TESL Canada journal.* pp.09-29.

Paribakht, T.S. and Wesche, M. 1997. Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. *Second language vocabulary acquisition: A rationale for pedagogy.* **55**(4), pp.174-200.

Pathan, Z.H., Ismail, S.A.M.M. and Soomro, N.H. 2019. An Investigation into Receptive Vocabulary Growth and Its Predictability for Reading Development of University Students in a Semester Course. *International Journal of Instruction.* **12**(1), pp.797-808.

Pellicer-Sánchez, A. 2019. Examining second language vocabulary growth: Replications of Schmitt (1998) and Webb & Chang (2012). *Language Teaching.* **52**(4), pp.512-523.

Plauen, E.O. 1996. Vater und Sohn, Band 2. *Ravensburger Taschenbuch.*

Qian, D.D. 1999. Assessing the roles of depth and breadth of vocabulary knowledge in reading comprehension. *CANADIAN MODERN LANGUAGE REVIEW-REVUE CANADIENNE DES LANGUES VIVANTES.* **56**(2), pp.282-307.

Rashtchi, M. and Aghili, H. 2014. Computerized input enhancement versus computer-assisted glosses: do they affect vocabulary recall and retention? *Theory & Practice in Language Studies.* **4**(8).

Read, J. 1993. The development of a new measure of L2 vocabulary knowledge. *Language Testing.* **10**(3), pp.355-371.

Read, J. 1998. *Word associates test.*

Read, J. 2000. *Assessing vocabulary.* Cambridge: Cambridge University Press.

Richards, J.C. 1976. The role of vocabulary teaching. *TESOL quarterly.* pp.77-89.

Sahiruddin, S. 2008. ENGLISH IN INDONESIAN ISLAMIC HIGHER EDUCATION: Examining The Relationship between Performance in The Yes/No Test and Reading Skills. *JOURNAL OF INDONESIAN ISLAM.* **2**(2), p379.

Schmitt, N. 1998. Tracking the incremental acquisition of second language vocabulary: A longitudinal study. *Language learning.* **48**(2), pp.281-317.

Schmitt, N. 2000. *Vocabulary in language teaching.* Cambridge;New York;: Cambridge University Press.

Schmitt, N. 2008. Instructed second language vocabulary learning. *Language teaching research.* **12**(3), pp.329-363.

Schmitt, N. 2010. *Researching vocabulary: a vocabulary research manual.* Basingstoke: Palgrave Macmillan.

Schmitt, N. and McCarthy, M. 1997. *Vocabulary: description, acquisition and pedagogy.* Cambridge;New York;: Cambridge University Press.

Schmitt, N. and Meara, P. 1997. Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes. *Studies in second language acquisition.* **19**(1), pp.17-36.

Schmitt, N. and Schmitt, D. 2020. *Vocabulary in language teaching.* Cambridge university press.

Schmitt, N., Schmitt, D. and Clapham, C. 2001. Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing.* **18**(1), pp.55-88.

Schmitt, N. and Zimmerman, C.B. 2002. Derivative Word Forms: What Do Learners Know? *TESOL Quarterly.* **36**(2), pp.145-171.

Shin, D., Chon, Y.V. and Kim, H. 2011. Receptive and productive vocabulary sizes of high school learners: What next for the basic word list. *English Teaching.* **66**(3), pp.123-148.

Šišková, Z. 2016. The relationship between receptive and productive vocabulary of Slavic EFL learners. *Topics in Linguistics.* **17**(2), pp.26-40.

Smith, G.G., Smith, S.D., Li, M., Drobisz, J., Park, H.-R. and Kim, D. 2013. Play games or study? Computer games in eBooks to learn English vocabulary. *Computers & Education.* **69**, pp.274-286.

Stæhr, L.S. 2008. Vocabulary size and the skills of listening, reading and writing. *The Language Learning Journal.* **36**(2), pp.139-152.

Stewart, J., Batty, A.O. and Bovee, N. 2012. Comparing Multidimensional and Continuum Models of Vocabulary Acquisition: An Empirical Examination of the Vocabulary Knowledge Scale. *TESOL Quarterly.* **46**(4), pp.695-721.

Stoeckel, T., McLean, S. and Nation, P. 2020. Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition.* **43**(1), pp.181-203.

Sun, C.-H. 2017. The value of picture-book reading-based collaborative output activities for vocabulary retention. *Language Teaching Research.* **21**(1), pp.96-117.

Sungprakul, S. 2016. Measuring Vocabulary Size of Thai University

Students. *International Journal of Social Science and Humanities Research.* **4**(4), pp.608-624.

Susanto, A. 2017. Assessing the relationship between Vocabulary Level Test (VLT) and reading comprehension. *Studies in English Language and Education.* **4**(2), pp.157-171.

Sydorenko, T. 2010. MODALITY OF INPUT AND VOCABULARY ACQUISITION. *LANGUAGE LEARNING & TECHNOLOGY.* **14**(2), pp.50-73.

Tahmasbi, M. and Farvardin, M.T. 2017. Probing the Effects of Task Types on EFL Learners' Receptive and Productive Vocabulary Knowledge: The Case of Involvement Load Hypothesis. *SAGE Open.* **7**(3).

Takala, S.J. 1984. *Evaluation of students' knowledge of English vocabulary in the Finnish comprehensive school.* thesis, University of Illinois at Urbana-Champaign.

Thai-English Electronic Dictionary Lexitron. 2020. [Online]. [Accessed 1 Jan 2019]. Available from: https://lexitron.nectec.or.th/2009_1/

Thorndike, E.L. and Lorge, I. 1944. *The teacher's word book of 30,000 words.*

Treffers-Daller, J., Parslow, P. and Williams, S. 2016. Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. *Applied Linguistics.* **39**(3), pp.302-327.

Varandi, S.B. and Faezi, F. 2013. The Role of Dimensions of Lexical Repertoire in Reading Comprehension and Incidental Vocabulary Acquisition. *Theory & Practice in Language Studies.* **3**(11), pp.2078-2085.

Vincy, I. 2020. Examining the effect of explicit instruction on vocabulary learning and on

receptive-productive gap: An experimental study. *Journal of Language and Linguistic Studies.* **16**(4), pp.2040-2058.

Walters, J. and Bozkurt, N. 2009. The effect of keeping vocabulary notebooks on vocabulary acquisition. *Language Teaching Research.* **13**(4), pp.403-423.

Ward, J. 1999. How large a vocabulary do EAP engineering students need.

Ward, J. 2009a. A basic engineering English word list for less proficient foundation engineering undergraduates. *English for Specific Purposes.* **28**(3), pp.170-182.

Ward, J. 2009b. EAP reading and lexis for Thai engineering undergraduates. *Journal of English for Academic Purposes.* **8**(4), pp.294-301.

Ward, J. and Chuenjundaeng, J. 2009. Suffix knowledge: Acquisition and applications. *System.* **37**(3), pp.461-469.

Waring, R. 1997a. A comparison of the receptive and productive vocabulary sizes of some second language learners. *Immaculata.* (1), pp.53-68.

Waring, R. 1997b. A study of receptive and productive learning from word cards. *Studies in Foreign Languages and Literature.* **21**(1), pp.94-114.

Waring, R. 1999. *Tasks for assessing second language receptive and productive vocabulary*. thesis, University of Wales Swansea.

Waring, R. 2000. The'State Rating Task'—An alternative method of assessing receptive and productive vocabulary. *Kiyo, Notre Dame Seishin University: Studies in Foreign Languages and Literature.* **24**(1), pp.125-154.

Webb, S. 2005. Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition.* **27**(1), pp.33-52.

Webb, S. 2007a. Learning word pairs and glossed sentences: The effects of a single context on vocabulary knowledge. *Language teaching research.* **11**(1), pp.63-81.

Webb, S. 2007b. The Effects of Repetition on Vocabulary Knowledge. *Applied Linguistics.* **28**(1), pp.46-65.

Webb, S. 2008. RECEPTIVE AND PRODUCTIVE VOCABULARY SIZES OF L2 LEARNERS. *Studies in Second Language Acquisition.* **30**(1), pp.79-95.

Webb, S. 2009. The Effects of Receptive and Productive Learning of Word Pairs on Vocabulary Knowledge. *RELC Journal.* **40**(3), pp.360-376.

Webb, S. 2021. A different perspective on the limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition.* **43**(2), pp.454-461.

Webb, S. and Chang, A.C.-S. 2012. Second language vocabulary growth. *RELC journal.* **43**(1), pp.113-126.

Webster's Ninth New Collegiate Dictionary. 1987. s.v.

Wesche, M. and Paribakht, T.S. 1996. Assessing second language vocabulary knowledge: Depth versus breadth. *Canadian Modern Language Review.* **53**(1), pp.13-40.

West, M. 1953. *A general service list of English words, with semantic frequencies and a supplementary word-list for the writing of popular science and technology.* Rev. and enl. ed. London: Longmans.

Wilson, M.E. and Beineke, L. 1979. *Applications of graph theory.* Academic Press.

Xue, G. and Nation, I.S.P. 1984. The University Word List. *Language Learning and Communication.* **3**(2), pp.215-229.

Yamamoto, Y. 2011. Bridging the gap between receptive and productive vocabulary size through extensive reading. *The Reading Matrix.* **11**(3).

Yang, Y., Shintani, N., Li, S. and Zhang, Y. 2017. The effectiveness of post-reading word-focused activities and their associations with working memory. *System.* **70**, pp.38-49.

Yunus, K., Mohamad, M. and Waelateh, B. 2016. The breadth of receptive vocabulary knowledge among English major university students. *Journal of Nusantara Studies (JONUS).* **1**(1), pp.7-17.

Zechmeister, E.B., Chronis, A.M., Cull, W.L., D'Anna, C.A. and Healy, N.A. 1995. Growth of a functionally important lexicon. *Journal of Reading Behavior.* **27**(2), pp.201-212.

Zhang, X. and Lu, X. 2013. A longitudinal study of receptive vocabulary breadth knowledge growth and vocabulary fluency development. *Applied Linguistics.* **35**(3), pp.283-304.

Zhang, X. and Lu, X. 2015. The relationship between vocabulary learning strategies and breadth and depth of vocabulary knowledge. *The Modern Language Journal.* **99**(4), pp.740-753.

Zhao, T. and Macaro, E. 2016. What works better for the learning of concrete and abstract words: teachers' L1 use or L2-only explanations? *International Journal of Applied Linguistics.* **26**(1), pp.75-98.

Zheng, Y. 2012. Exploring long-term productive vocabulary development in an EFL context: The role of motivation. *System.* **40**(1), pp.104;119;-119.

Zhong, H. 2008. *Vocabulary size development: A study on Chinese high school students*. Unpublished.

Zhong, H. 2016. The relationship between receptive and productive vocabulary knowledge: a perspective from vocabulary use in sentence writing. *The Language Learning Journal.* pp.1-17.

Zhong, H. and Hirsh, D. 2009. Vocabulary growth in an English as a foreign language context. *University of Sydney papers in TESOL.* **4**(4).

Zhou, S. 2010. Comparing receptive and productive academic vocabulary knowledge of Chinese EFL learners. *Asian Social Science.* **6**(10), p14.

Zyad, H. 2017. Vocabulary Growth in College-Level Students' Narrative Writing. *International Journal of Languages' Education and Teaching.* **5**(3), pp.363-372.

Appendix A: Ethical approval

The Secretariat
University of Leeds
Leeds, LS2 9JT
Tel: 0113 343 4873
Email: ResearchEthics@leeds.ac.uk

UNIVERSITY OF LEEDS

Peeraya Utsajit
School of Education
University of Leeds
Leeds, LS2 9JT

ESSL, Environment and LUBS (AREA) Faculty Research Ethics Committee
University of Leeds

29 October 2018

Dear Peeraya

Title of study: A cross-sectional study of the receptive and productive vocabulary knowledge of English-major students in Thailand

Ethics reference: AREA 17-160 amendment October 2018

I am pleased to inform you that your amendment to the research application listed above has been reviewed by the ESSL, Environment and LUBS (AREA) Faculty Research Ethics Committee and I can confirm a favourable ethical opinion as of the date of this letter. The following documentation was considered:

Document	Version	Date
AREA 17-160 amendment October 2018 1 Ethics Amendment (Peeraya_with Signatures_02 Oct 2018).pdf	1	08/10/18
AREA 17-160 Ethical_Review_Form_V3 (AREA 17-160_Peeraya_Version 3_10 Oct 2018).pdf	1	10/10/18
AREA 17-160 Transfer Report_Peeraya_with Template (Revised Version 10 Oct 2018).pdf	2	10/10/18
AREA 17-160 5 Participant_Information_Sheet_for teachers_Class observation (Peeraya_Version 3_2 Oct 2018).doc	1	08/10/18
AREA 17-160 6 Information Sheet _for Grading Thai teacher (Peeraya_Version 1_02 Oct 2018).doc	1	08/10/18
AREA 17-160 4 Participant_Information Sheet _for Students in Main study (Peeraya_Version 3_2 Oct 2018).doc	1	08/10/18
AREA 17-160 Information Sheet_for Student in Pilot (Peeraya_Version 1_02 Oct 2018).doc	1	08/10/18
AREA 17-160 5 Participant_consent_formlowrisk_for teachers_Class observation (Peeraya_Version 3_2 Oct 2018).doc	1	08/10/18
AREA 17-160 4 Participant_consent_formlowrisk_for Students in Main study (Peeraya_Version 2_2 Oct 2018).doc	1	08/10/18
AREA 17-160 3 Consent Form_for Student in Pilot (Peeraya_Version 1_02 Oct 2018).doc	1	08/10/18
AREA 17-160 6 Receptive Test & Productive Test 1 (Day 1 & 2) (Peeraya_Version 2_02 Oct 2018)).xlsx	1	08/10/18
AREA 17-160 Observation Sheet for Vocabulary Teaching and Learning (Version 1_2 Oct 2018).docx	1	08/10/18
AREA 17-160 1 Ethical_Review_Form_V3 (AREA 17-160_Peeraya_Version 2_10 July 2018).doc	2	10/07/18
AREA 17-160 4 Participant_Information_Sheet (AREA 17-160_Peeraya_Version 2_12 July 2018_for teacher_class observation).doc	2	12/07/18
AREA 17-160 3 Participant_consent_formlowrisk (AREA 17-160_Peeraya_Version 2_10 July 2018_for teachers_class observation).doc	1	10/07/18
AREA 17-160 5 Questionnaire (AREA 17-160_Peeraya_Version 2_10 July 2018).doc	1	10/07/18

Please notify the committee if you intend to make any further amendments to the original research as submitted at date of this approval as all changes must receive

ethical approval prior to implementation. The amendment form is available at http://ris.leeds.ac.uk/EthicsAmendment.

Please note: You are expected to keep a record of all your approved documentation, as well as documents such as sample consent forms, and other documents relating to the study. This should be kept in your study file, which should be readily available for audit purposes. You will be given a two week notice period if your project is to be audited. There is a checklist listing examples of documents to be kept which is available at http://ris.leeds.ac.uk/EthicsAudits.

We welcome feedback on your experience of the ethical review process and suggestions for improvement. Please email any comments to ResearchEthics@leeds.ac.uk.

Yours sincerely

Jennifer Blaikie
Senior Research Ethics Administrator, the Secretariat
On behalf of Dr Kahryn Hughes, Chair, AREA Faculty Research Ethics Committee

CC: Student's supervisor(s)

Appendix B: Information sheet (participant – pilot study)

Research and Innovation Service

UNIVERSITY OF LEEDS

Participant Information Sheet (Pilot Study)

Below are examples of the main points an information sheet should include

The title of the research project

A Cross-Sectional Study of the Receptive and Productive Vocabulary Knowledge of English-Major Students in Thailand

(Note: Receptive vocabulary means words students know when they read, and productive vocabulary means words they use when writing.)

Invitation paragraph

You are being invited to take part in a research project (a pilot study). Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

What is the purpose of the project?

The purpose of the project is to study the receptive and productive vocabulary knowledge of Thai university students majoring in English and to examine the relationship or the development of their vocabulary knowledge over years of study cross-sectionally by comparing the test scores of the four year cohorts. The study focuses on high-frequency, mid-frequency, low-frequency, and academic vocabulary. The measurement can be done by using vocabulary tests and students' writing pieces. The students' biodata and their exposure to English outside class will be collected by using a questionnaire.

Why have I been chosen?

This mixed-method research study focuses on students in Thailand. You are chosen by purposive sampling method because you speak Thai as your mother tongue, and you come from different regions of Thailand. You use the English language regularly in class. Therefore, you can represent English-major students in Thailand.

What do I have to do? / what will happen to me if I take part?

You will not get any benefits or risks from taking part in this research study, but it is hoped that your test results can then be used to improve the teaching and learning for English-major students of the Faculty of Liberal Arts and for English-major students in Thailand, especially in terms of vocabulary knowledge development. During the data collection of the pilot study, you will sit two vocabulary tests.

There is no reimbursement for participants.

What are the possible disadvantages and risks of taking part?

The disadvantage is that the participants spend your time doing the test, but there are not any risks for taking part in the study.

What are the possible benefits of taking part?

Whilst there are no immediate benefits for those people participating in the project, it is hoped that this work will contribute to the improvement of teaching and learning academic subjects for English-major students in the future.

Do I have to take part?

It depends on your decision whether to take part. If you do decide to take part you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time without it affecting any benefits that you are entitled to in any way. You do not have to give a reason.

Will my taking part in this project be kept confidential?/ What will happen to the results of the research project?

I agree for the personal data collected and your taking part to be stored, anonymised by numbering and kept confidential.

Withdrawing

You can withdraw your data after taking part in the study, but it is impossible to withdraw the data after the analysis (after 31 December 2018).

Contact for further information

Researcher

Name: Miss Peeraya Utsajit

Address: School of Education, Faculty of Education, Social Science, and Law, Hillary Place, University of Leeds, LEEDS, LS2 9JT

Telephone number: 0113 243 1751

Email addresses: edpu@leeds.ac.uk

Supervisor

Name: Dr. Richard Badger

Address: School of Education, Faculty of Education, Social Science, and Law, Hillary Place, University of Leeds, LEEDS, LS2 9JT

Telephone number: 0113 3434644

Co-Supervisor

Name: Professor Alice Deignan

Address: School of Education, Faculty of Education, Social Science, and Law, Hillary Place, University of Leeds, LEEDS, LS2 9JT

Telephone number: 0113 3434920

Note:

This study has been reviewed and given a favourable opinion by Research Ethics Committee on [18 July 2018], ethics reference [AREA 17-160 Response 2]

You will be given a copy of the information sheet and a copy of the signed consent form to keep.

Appendix C: Consent form (participant – pilot study)

[School of Education/ Faculty of Education, Social Sciences, and Law]

UNIVERSITY OF LEEDS

For a relatively low risk piece of research, it may be more appropriate to add a shortened consent form below to the end of the information sheet.

Consent to take part in

[A Cross-Sectional Study of the Receptive and Productive Vocabulary Knowledge of English-Major Students in Thailand]

	Add your initials next to the statements you agree with
I confirm that I have read and understand the information sheet dated 02/10/2018 explaining the above research project and I have had the opportunity to ask questions about the project.	
I agree for the data collected from me to be stored and used in relevant future research in an anonymised form.	
I understand that relevant sections of the data collected during the study, may be looked at by auditors from the University of Leeds or from regulatory authorities where it is relevant to my taking part in this research. I give permission for these individuals to have access to my records.	
I agree to take part in the above research project and will inform the lead researcher should my contact details change during the project and, if necessary, afterwards.	

Name of participant	
Student ID	
Participant's signature	
Date	
Name of lead researcher	Miss Peeraya Utsajit
Signature	
Date*	

*To be signed and dated in the presence of the participant.

Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form, the letter/ pre-written script/ information sheet and any other written information provided to the participants. A copy of the signed and dated consent form should be kept with the project's main documents which must be kept in a secure location.

Project title	Document type	Version #	Date
A Cross-Sectional Study of the Receptive and Productive Vocabulary Knowledge of English-Majored Students in Thailand	Consent form for Thai students (Pilot)	Version 1	02/10/2018

Appendix D: Information sheet (participant – main study)

UNIVERSITY OF LEEDS

Participant Information Sheet

The title of the research project

A Cross-Sectional Study of the Receptive and Productive Vocabulary Knowledge of English-Major Students in Thailand

(Note: Receptive vocabulary means words students know when they read, and productive vocabulary means words they use when writing.)

Invitation paragraph

You are being invited to take part in a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information. Take time to decide whether or not you wish to take part.

What is the purpose of the project?

The purpose of the project is to study the receptive and productive vocabulary knowledge of Thai university students majoring in English and to examine the relationship or the development of their vocabulary knowledge over years of study cross-sectionally by comparing the test scores of the four year cohorts. The study focuses on high-frequency, mid-frequency, low-frequency, and academic vocabulary. The measurement can be done by using vocabulary tests and students' writing pieces. The students' biodata and their exposure to English outside class will be collected by using a questionnaire.

Why have I been chosen?

This mixed-method research study focuses on English-major students in Thailand. You are chosen by purposive sampling method because you major in English, you are studying in year 1-4, you speak Thai as your mother tongue, and you come from different regions of Thailand. Therefore, you can represent four years groups of English-major students in Thailand.

What do I have to do? / what will happen to me if I take part?

You will not get any benefits or risks from taking part in this research study, but it is hoped that your test results can then be used to improve the teaching and learning for English-major students of the Faculty of Liberal Arts and for English-major students in Thailand, especially in terms of vocabulary knowledge development.

During the data collection, you will be asked to fill out a questionnaire asking about your biodata (name, student ID, study year, age, gender), your contact details (email address and mobile number), your education background (your previous school, programme you learned in your secondary school) and your exposure to English outside classroom. Also, you will sit a vocabulary test for 50 minutes. After a ten-minute break, you will be asked to write a piece

of 300-word essay in 60 minutes. Your reading and writing classes will also be observed to examine the vocabulary teaching and learning.

There are not any lifestyle restrictions as a result of participating.

There is no reimbursement for participants.

What are the possible disadvantages and risks of taking part?

The disadvantage is that the participants spend their time doing the test, but there are not any risks for taking part in the study.

What are the possible benefits of taking part?

Whilst there are no immediate benefits for those people participating in the project, it is hoped that this work will contribute to the improvement of teaching and learning academic subjects for English-major students in the future.

Do I have to take part?

It depends on your decision whether or not to take part. If you do decide to take part, you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time without it affecting any benefits that you are entitled to in any way. You do not have to give a reason.

Will my taking part in this project be kept confidential? / What will happen to the results of the research project?

I agree for the personal data collected and your taking part to be stored, anonymised and kept confidential.

Withdrawing

You can withdraw your data after taking part in the study, but it is impossible to withdraw the data after the analysis (after 31 May 2019).

Contact for further information

Researcher

Name: Miss Peeraya Utsajit

Address in Thailand: 999 Faculty of Liberal Arts, Mahidol University, Phutthamonthon Sai 4, Nakhonpathom, Thailand 73170

Address in the UK: School of Education, Hillary Place, University of Leeds, Leeds, the UK, LS9 2JT

Telephone number: 0113 243 1751

Email addresses: edpu@leeds.ac.uk

Supervisor

Name: Dr. Richard Badger

Address: School of Education, Faculty of Education, Social Science, and Law, Hillary Place, University of Leeds, LEEDS, LS2 9JT

Telephone number: 0113 3434644

Co-Supervisor

Name: Professor Alice Deignan

Address: School of Education, Faculty of Education, Social Science, and Law, Hillary Place, University of Leeds, LEEDS, LS2 9JT

Telephone number: 0113 3434920

Note:

This study has been reviewed and given a favourable opinion by Research Ethics Committee on [18 July 2018], ethics reference [AREA 17-160 Response 2]

You will be given a copy of the information sheet and a copy of the signed consent form to keep.

Appendix E: Consent form (participant – main study)

[School of Education/ Faculty of Education, Social Sciences, and Law]

UNIVERSITY OF LEEDS

For a relatively low risk piece of research, it may be more appropriate to add a shortened consent form below to the end of the information sheet.

Consent to take part in

[A Cross-Sectional Study of the Receptive and Productive Vocabulary Knowledge of English-Major Students in Thailand]

	Add your initials next to the statements you agree with
I confirm that I have read and understand the information sheet dated 02/10/2018 explaining the above research project and I have had the opportunity to ask questions about the project.	
I agree for the data collected from me to be stored and used in relevant future research in an anonymised form.	
I understand that relevant sections of the data collected during the study, may be looked at by auditors from the University of Leeds or from regulatory authorities where it is relevant to my taking part in this research. I give permission for these individuals to have access to my records.	
I agree to take part in the above research project and will inform the lead researcher should my contact details change during the project and, if necessary, afterwards.	

Name of participant	
Student ID	
Participant's signature	
Date	
Name of lead researcher	Miss Peeraya Utsajit
Signature	
Date*	

*To be signed and dated in the presence of the participant.

Once this has been signed by all parties the participant should receive a copy of the signed and dated participant consent form, the letter/ pre-written script/ information sheet and any other written information provided to the participants. A copy of the signed and dated consent form should be kept with the project's main documents which must be kept in a secure location.

Project title	Document type	Version #	Date
A Cross-Sectional Study of the Receptive and Productive Vocabulary Knowledge of English-Majored Students in Thailand	consent form for English-majored students	Version 2	02/10/ 2018

Appendix F: Information sheet (rater)

UNIVERSITY OF LEEDS

Information Sheet (Rater)

Below are examples of the main points an information sheet should include

The title of the research project

A Cross-Sectional Study of the Receptive and Productive Vocabulary Knowledge of English-Major Students in Thailand

(Note: Receptive vocabulary means words students know when they read, and productive vocabulary means words they use when writing.)

Invitation paragraph

You are being invited to take part in a research project. Before you decide it is important for you to understand why the research is being done and what it will involve. Please take time to read the following information carefully and discuss it with others if you wish. Ask me if there is anything that is not clear or if you would like more information. Take time to decide whether you wish to take part.

What is the purpose of the project?

The purpose of the project is to study the receptive and productive vocabulary knowledge of Thai university students majoring in English and to examine the relationship or the development of their vocabulary knowledge over years of study cross-sectionally by comparing the test scores of the four year cohorts. The study focuses on high-frequency, mid-frequency, low-frequency, and academic vocabulary. The measurement can be done by using vocabulary tests and students' writing pieces. The students' biodata and their exposure to English outside class will be collected by using a questionnaire.

Why have I been chosen?

You are a native English speaker and you are experienced in teaching English to Thai university students. To avoid a bias from one rater, another rater is needed in this study.

What do I have to do?/ what will happen to me if I take part?

You will not get any benefits or risks from taking part in this research study. You have to check the answers of the two productive vocabulary tests. A sensitive scoring method is used. You give one point when the participant uses the test word correctly in meaning in the sentence and two points when he or she uses the test word with both semantic appropriateness and grammatical accuracy. You will not give any point when the meaning of the test word does not fit the context at all.

What are the possible disadvantages and risks of taking part?

The disadvantage is that you need to spend your time checking the test.

What are the possible benefits of taking part?

Whilst there are no immediate benefits for the rater, it is hoped that this work will contribute to the improvement of teaching and learning academic subjects for English-major students in the future.

Do I have to take part?

It depends on your decision whether to take part. If you do decide to take part, you will be given this information sheet to keep (and be asked to sign a consent form) and you can still withdraw at any time without it affecting any benefits that you are entitled to in any way. You do not have to give a reason.

Will my taking part in this project be kept confidential?/ What will happen to the results of the research project?

I agree for your taking part to be kept confidential. The name of the rater will be anonymised by using the term 'Rater A'.

Withdrawing

You can withdraw your data after taking part in the study, but it is impossible to withdraw the data after the analysis (after 31 December 2018).

Contact for further information

Researcher

Name: Miss Peeraya Utsajit

Address in Thailand: 999 Faculty of Liberal Arts, Mahidol University, Phutthamonthon Sai 4, Nakhonpathom, Thailand 73170

Address in the UK: School of Education, Hillary Place, University of Leeds, Leeds, the UK, LS9 2JT

Telephone number: 0113 243 1751

Email addresses: edpu@leeds.ac.uk

Supervisor

Name: Dr. Richard Badger

Address: School of Education, Faculty of Education, Social Science, and Law, Hillary Place, University of Leeds, LEEDS, LS2 9JT

Telephone number: 0113 3434644

Email addresses: R.G.Badger@education.leeds.ac.uk

Co-Supervisor

Name: Professor Alice Deignan

Address: School of Education, Faculty of Education, Social Science, and Law, Hillary Place, University of Leeds, LEEDS, LS2 9JT

Telephone number: 0113 3434920

Email addresses: A.H.Deignan@education.leeds.ac.uk

Note:

This study has been reviewed and given a favourable opinion by Research Ethics Committee on [18 July 2018], ethics reference [AREA 17-160 Response 2]

You will be given a copy of the information sheet and a copy of the signed consent form to keep.

Appendix G: Questionnaire

Questionnaire **แบบสอบถาม**	
Research topic: A Cross-Sectional Study of the Receptive and Productive Vocabulary Knowledge of English-Major University Students in Thailand **หัวข้อวิจัย: การวัดระดับและเปรียบเทียบความรู้ระหว่างคำศัพท์ที่รู้และคำศัพท์ที่นำไปใช้ของนักศึกษาสาขาภาษาอังกฤษในประเทศไทย**	

Section I: Biodata and Educational Background (The information will be kept confidential.)
ส่วนที่ 1: ข้อมูลส่วนตัวและพื้นฐานการศึกษา (ข้อมูลส่วนตัวจะไม่ถูกเผยแพร่)

No.	Item	Response
1	Full Name ชื่อจริง-นามสกุล	
2	Student ID รหัสนักศึกษา	
3	Year of Study ชั้นปีที่กำลังศึกษา	
4	Age อายุ	
5	Gender เพศ	
6	Email Address อีเมลที่ใช้ประจำ	
7	Mobile Number เบอร์มือถือที่สามารถติดต่อได้	
8	Type of Your Secondary School ประเภทของโรงเรียนมัธยมศึกษาของตนเอง	
9	Programme You Studied in Your Secondary School สายที่เรียนในระดับมัธยมศึกษา	

Section II: Exposure to English
ส่วนที่ 2: การใช้ภาษาอังกฤษของตนเอง

No.	Item	Response
1	How many years have you been studying English in your life? คุณได้เรียนภาษาอังกฤษมาทั้งสิ้นกี่ปี (ตั้งแต่เริ่มเรียนที่โรงเรียนครั้งแรก)	
2	Have you ever lived or studied in any English-speaking coutries? คุณเคยมีประสบการณ์อยู่อาศัยหรือศึกษาในประเทศที่ใช้ภาษาอังกฤษหรือไม่	
-	Country 1 ประเทศแรกที่เคยไปอาศัยหรือไปศึกษา	
	How many months? จำนวนเดือนที่ไปอยู่ในประเทศแรก	
-	Country 2 ประเทศที่สองที่เคยไปอาศัยหรือไปศึกษา	
	How many months? จำนวนเดือนที่ไปอยู่ในประเทศที่สอง	
-	Country 3 ประเทศที่สามที่เคยไปอาศัยหรือไปศึกษา	
	How many months? จำนวนเดือนที่ไปอยู่ในประเทศที่สอง	
3	What activities or hobbies do you often do to practice the English language outside-class? กิจกรรมยามว่างอะไรที่คุณทำเพื่อฝึกภาษาอังกฤษนอกห้องเรียน	
-	Activity 1 กิจกรรมฝึกภาษาอังกฤษที่ 1	
	How many times per month? ทำกิจกรรมนี้กี่ครั้งต่อเดือน (โดยประมาณ)	
	How many minutes each time? ทำกิจกรรมที่นี่นานกี่นาทีต่อครั้ง (โดยประมาณ)	
-	Activity 2 กิจกรรมฝึกภาษาอังกฤษที่ 2	
	How many times per month? ทำกิจกรรมนี้กี่ครั้งต่อเดือน (โดยประมาณ)	
	How many minutes each time? ทำกิจกรรมที่นี่นานกี่นาทีต่อครั้ง (โดยประมาณ)	
-	Activity 3 กิจกรรมฝึกภาษาอังกฤษที่ 3	
	How many times per month? ทำกิจกรรมนี้กี่ครั้งต่อเดือน (โดยประมาณ)	
	How many minutes each time? ทำกิจกรรมที่นี่นานกี่นาทีต่อครั้ง (โดยประมาณ)	
-	Activity 4 กิจกรรมฝึกภาษาอังกฤษที่ 4	
	How many times per month? ทำกิจกรรมนี้กี่ครั้งต่อเดือน (โดยประมาณ)	
	How many minutes each time? ทำกิจกรรมที่นี่นานกี่นาทีต่อครั้ง (โดยประมาณ)	

Appendix H: Receptive and controlled productive vocabulary tests

Receptive Test and Productive	Full Name: ชื่อ-นามสกุล	
	Student ID: รหัสนักศึกษา	
	Year (1/2/3/4): ชั้นปี	

Instructions: Choose the number 1-5 for each of the following test words. คำสั่ง: กรุณาใส่เลข 1-5 ที่แสดงถึงระดับความรู้ของตนเองที่มีต่อคำศัพท์ในแต่ละข้อได้อย่างเหมาะสมที่สุด	1	I don't remember having seen this word before. ไม่เคยเห็นคำศัพท์นี้มาก่อน
	2	I have seen this word before, but I don't know what it means. เคยเห็นคำศัพท์นี้ แต่ไม่รู้ความหมาย
	3	I have seen this word before, and I think it means _____. (translation/synonym) เคยเห็นคำศัพท์นี้ ไม่แน่ใจความหมาย น่าจะหมายถึง _____ (ใส่คำแปล/คำอธิบาย ที่เป็นภาษาไทยหรืออังกฤษก็ได้)
	4	I know this word. It means _____. (translation/synonym) รู้จักคำศัพท์นี้ มันหมายถึง _____ (ใส่คำแปล/คำอธิบาย ที่เป็นภาษาไทยหรืออังกฤษก็ได้)
	5	I can use this word in a sentence: _____ สามารถใช้คำศัพท์นี้ในประโยค ดังนี้ _____ (Write a sentence using the test word and the conjunction 'so', 'because', or 'so that' and please also give the translation/synonym.) (กรุณาเขียนประโยคที่ใช้คำศัพท์นั้นๆ และคำเชื่อม 'so', 'because', หรือ 'so that' พร้อมทั้งใส่คำแปล/คำอธิบาย ที่เป็นภาษาไทยหรืออังกฤษก็ได้)

Item คำศัพท์ที่	Test Word คำศัพท์ที่ทดสอบ	Write Number ระดับที่รู้คำศัพท์	Translation/Synonym คำแปล/คำอธิบาย (ไทยหรืออังกฤษก็ได้)	Write a sentence using the test word and one of the following conjunctions ('so', 'because', or 'so that'). หากรู้คำศัพท์ระดับ 5 กรุณาเขียนประโยคที่ใช้คำศัพท์นั้นๆ และคำเชื่อม 'so', 'because', หรือ 'so that'
1	sport			
2	notion			
3	pail			
4	antics			
5	labour			
6	victory			
7	palm			
8	novelty			
9	batch			
10	percent			
11	birth			
12	belt			
13	balloon			
14	connoisseur			
15	principle			
16	temperature			
17	chill			
18	phase			

19	jumble		
20	fund		
21	flesh		
22	ox		
23	mess		
24	saliva		
25	technique		
26	salary		
27	structure		
28	apron		
29	dregs		
30	philosophy		
31	education		
32	bench		
33	compliment		
34	casualty		
35	sum		
36	scale		
37	charity		
38	apparatus		
39	seclusion		
40	consent		
41	journey		
42	province		
43	revenue		
44	revelry		
45	investigation		
46	treasure		
47	lieutenant		
48	mare		
49	apparition		
50	decade		

Appendix I: Free productive test

Productive Test 2

Name		Student ID	

Topic (หัวข้อ): **"A person cannot be poor and happy, because money is always needed to gain something that is important to that person."**

Instruction (คำสั่ง): **Write an essay of 300 words in English on the topic provided. Argue for or against this idea.** (เขียนเรียงความเป็นภาษาอังกฤษ 300 คำ เพื่อสนับสนุนหรือคัดค้านกับหัวข้อที่กำหนดให้)