교육학석사학위논문

A Corpus-based Analysis of Collocations in Korean Middle and High School English Textbooks and Korean EFL Learner Writing

한국 중학교 및 고등학교 영어 교과서와 한국인 학습자 영어글쓰기에서의 연어 사용에 대한 코퍼스 기반 분석

2020 년 2 월

서울대학교 대학원

외국어교육과 영어전공

김 영 신

A Corpus-based Analysis of Collocations in Korean Middle and High School English Textbooks and Korean EFL Learner Writing

by
Young Shin Kim

A Thesis Submitted to
the Department of Foreign Language Education
in Partial Fulfillment of the Requirement
for the Degree of Master of Arts in Education

At the
Graduate School of Seoul National University

February 2020

A Corpus-based Analysis of Collocations in Korean Middle and High School English Textbooks and Korean EFL Learner Writing

한국 중학교 및 고등학교 영어 교과서와 한국인 학습자 영어 글쓰기에서의 연어 사용에 대한 코퍼스 기반 분석

지도교수 오 선 영

이 논문을 교육학 석사 학위논문으로 제출함
2019년 12월

서울대학교 대학원
외국어교육과 영어전공
김 영 신

김영신의 석사학위논문을 인준함
2020년 1월

위 원 장 _____

부위원장 _____

위 원 _____

A Corpus-based Analysis of Collocations in Korean Middle and High School English Textbooks and Korean EFL Learner Writing

APPROVED BY THESIS COMMITTEE:

BYUNGMIN LEE, COMMITTEE CHAIR

YOUNGSOON SO

SUN-YOUNG OH

ABSTRACT

Collocation, the habitual word association, has been thought of as a large and significant component of native speakers' language production. As with its pervasiveness in natural English, the idiomaticity of collocation explains its usefulness as a readily available word sequence. Referred to as "semi-preconstructed phrases that constitute single choice" (Sinclair, 1991, p. 110), collocation is known to have a pivotal role in achieving native-like fluency, to which native speakers develop their sensitivity through extensive and intensive exposures to language input.

EFL learners, however, often face challenges with this lexical category, due to a lack of exposure to authentic collocations. And yet very little is known how collocations are represented in the major source of language input in the EFL classrooms. Thus, there is a need to understand various aspects of collocation use in the English learning materials, especially in the curriculum-based English textbooks used in most of the Korean EFL English classes. Moreover, much uncertainty still exists about how Korean EFL students would use collocations after years of English learning through textbooks. Therefore, the present study will explore the collocation use in the middle and high school English textbook and Korean EFL college learners' writings.

Employing the Firthian notion of collocation, the present study identifies collocations based on the statistical measure and has investigated the use of collocations in the middle and high school English textbook and Korean EFL college learner writing corpora regarding four distributional patterns; collocation density, diversity, repetition rate, and association strength. Furthermore, in order to complement a single-word level wordlist based on the 2015 National English

Curriculum in Korea, 1,718 words from the newly revised curriculum wordlist were used as head nouns. Also, three types of collocation were specified as a target of analysis; verb-noun collocation (VNC), noun-noun collocation (NNC), and adjective-noun collocation (ANC).

First, collocation use in the textbook corpus was examined. As for collocation density, the result shows that the textbook corpus presents a significantly larger body of VNCs and ANCs than in the reference corpus, indicating that Korean learners would be exposed to a relatively higher proportion of collocational input from the textbook materials. No significant difference was found in the use of NNCs. Second, the analysis of collocation diversity reveals that in textbook materials, more variety of collocation repertoire is presented given the text length; the diversity rate gives higher marks for VNCs and ANCs in the textbook corpus than in native baseline, while showing no significant difference for NNCs. The higher collocation density and diversity of VNCs and ANCs together can be seen as the extensive collocation use in language input. Next, all three subtypes are found to be markedly less repetitive in textbook materials than in the reference corpus. While collocations are highly recurrent phenomena in the native corpus, the textbook materials tend to introduce a larger body of diverse items without due repetition, compromising the collocational formulaicity. Lastly, while the estimated association strength of collocations in the target corpus is found to be generally higher, collocations at low-mid level association strength are relatively scarce in comparison to native reference data. Furthermore, a correlation between association measure and frequency level of individual collocations in the target corpus is weaker than in the reference corpus. This result indicates that in the textbook materials, the frequency of co-occurrence

is far less predictive of the association strength, and thus learners are not likely to benefit from frequency effects that would help them to distinguish between a wide range of associative strengths or to consolidate memory traces of stronger associations.

Regarding the second major research question, the present study examines Korean EFL college learners' writings. In learners' production, VNCs and ANCs occur more frequently than in the reference corpus. Exceptional is the density of NNCs, which have shown a significantly lower density in the learner corpus. When it comes to collocation diversity, learner writing presents a larger number of collocation types with VNCs and ANCs, but not with NNCs. From the result, heavier reliance on VNCs and ANCs and the contrastingly underrepresented NNCs are hypothesized as a distinctive pattern of non-native-like collocation use. The third variable, the repetition, has shown that the repetition rate of VNCs and ANCs is markedly lower, and that of NNCs higher in the learner writings than those in the reference corpus. Lastly, when examining association strength, learner writings have shown significantly higher association scores than the reference corpus does. The result indicates that learners' collocation repertoire may be limited to more typical and likely associations, falling short of the knowledge of the less predictable associations at low-mid level strength. Furthermore, a weaker correlation between association measure and frequency level of individual collocations in the learner corpus suggests that learners' use of collocation only weakly correspond to the association strength, and the distributional patterns of collocation use in learner writings may deviate from that in the native data.

These findings provide meaningful implications for collocation learning in

the Korean EFL context. First, the present study supports the view of ‘more is less’ in that the intensity of collocation use, the formulaic nature of repeated co-occurrence, may have been compromised by the extensive coverage of a larger number of collocations in the current Korean textbook materials. To represent authentic collocational distribution, it is thus recommended to increase the number of repetitions given to each lexical combination. Next, the current data also suggest that learners' sensitivity to a collocational relationship could be fostered if there is more correspondence between the level of co-occurrence frequency of collocational input and the association strength. Furthermore, pedagogical attention is called for to address learners’ restrictive collocational repertoire which we found to be highly limited to stronger associations. Lastly, there is a definite need for differentiated instructions on specific subtypes which could be particularly challenging for learners (ie., NNCs).

Keyword : Collocation, collocation density, collocation diversity, repetition, association strength, co-occurrence frequency, textbook materials, learners’ writing corpus, formulaicity

Student Number : 2018-22214

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1. Introduction

This study aims to investigate the use of collocation in the Korean middle and high school English textbooks, in conjunction with the written productions of Korean EFL college students. This chapter outlines the purpose and organization of the thesis. The statement of purpose is introduced in Section 1.1, followed by research questions in Section 1.2, and organization of the thesis in Section 1.3.

1.1. Purpose of the Study

The last decades of linguistics and English Language Teaching (ELT) research have established the significant role of formulaic language in native-like written productions and language acquisition. Abundant evidence as to the pervasiveness of formulaic language in native English has been accumulated. Native speakers make extensive use of word sequences, which are stored and retrieved as a unit in their memory without having to generate or analyze the sequences into segments (Cowie, 1992; Erman & Warren, 2000; Howarth, 1996; Sinclair, 1991; Wray, 2005). Cowie (1992), for example, measured collocational density in journals and found that more than 40% of verb-noun pairs are already well-established collocations. This finding is supported by Howarth (1998), who reported that up to 41% of verb-noun pairs are conventional collocations in academic writing. Furthermore, Erman and Warren (2000) showed that more than half of a written text consists of prefabricated language, concluding that raising awareness of this abundance of existing prefabrications would improve students' learning strategies and command of English; they also noted that teaching materials should be adapted

to represent the native-like use of language more precisely. Thus, a formulaic tendency is indeed an essential part of native speakers' language use; as stated by Howarth, "there is in native writing an identifiable core of collocational conventionality" (1996, p. 171).

While such pervasiveness represents the distribution of collocation category as a whole in native English, recursion, a repetitive association of words, characterizes the use of collocation at the individual level. According to Ellis (2001), native learners are given repetitive exposure to co-occurring pairs in native language input, and thus become increasingly efficient in processing word combinations as a single entity. The more frequently that learners encounter collocational pairs, the more likely they are to accumulate a vast amount of probabilistic information of the co-occurrence, and eventually consolidate the associative relation into their long-term memory.

As such, L1 learners seem to develop collocational competence through extensive and repeated exposure to word associations in natural input and achieve native-like fluency with the help of these readily available sequences. Similarly, Foster (2013) found that native speakers tend to increase collocation use under unprepared conditions, demonstrating that native speakers rely on their collocational repository which permits cognitive advantage and increased fluency. It is therefore becoming clear that greater levels of collocational density and repetition lead to the deeper entrenchment of the target collocations in learners' mental lexicon and to increased fluency.

Such acceptance of the role of prefabricated language in language acquisition and native English in general is a fundamental departure from the rule-

based theory where chunks are often thought as peripheral to a creative language production based on generative grammar. Phraseologists began to notice the productivity of collocation and other formulaic languages, which distinguish them from the dead-end, frozen idioms. In fact, they form a fruitful category as "mini-grammar" (Pawley & Syder, 1983, p. 216) serving as an alternative principle for language production. Halfway between "the extreme ends of the spectrum [of] free combinations and idioms" (Cowie, 1998, p. 186), collocations are said to form "the large and complex middle ground" (Howarth, 1998, p. 42) which exhibits a moderate level of variability and restrictiveness. It has been acknowledged that, therefore, native-like proficiency is strongly related to the skillful control of collocational fuzziness, that is, a half-fixed and half-flexible association of words.

As such, the significance of collocations has been established in terms of the categorical level of density and repetitiveness at the individual level. Also, mid-strength formulaicity defines collocational productivity and variability. These attributes of collocation altogether have offered advantages of fluent language processing and efficient L1 development for native speakers. Since the early 1980s, the significant role of collocation and formulaic language in native-like production has been recognized from the point of view of second language acquisition (eg., Cowie, 1992; Nattinger & DeCarrico, 1992; Pawley, Syder, Richards, & Schmidt, 1983) and EFL coursebooks development (eg., Flower, Berman, & Powell, 1989; Harmer & Rossner, 1997; McCarthy & O'Dell, 1994; Redman, Ellis, & Viney, 1989; Rudzka, Channell, Putseys, & Ostyn, 1981). Phraseology is critical in L2 development, Nattinger and DeCarrio (1992) contend, in that lexical phrases as unanalyzed chunk provides a bedrock for learners' language production until they

are synthesized into the emerging syntactic rules.

With growing attention having been paid to collocations, it has been pointed out by many researchers and teachers that EFL learners are currently provided insufficient collocational input, which may impede their language development. As suggested by Durrant and Schmitt (2009), a lack of exposure to collocational items may be profoundly responsible for L2 learners' lag in the development of collocational competence. In a similar vein, many researchers have pointed out the need for an integrated vocabulary syllabus containing words, collocations and lexical chunks, even from the very earliest level of learning (O'keeffe, McCarthy, & Carter, 2007). O'Keeffe, et al. (2007) maintained that the vocabulary syllabus is incomplete without due attention to the lexical chunks. Howarth (2013) also noted that published teaching materials do not recognize the nature of collocations or offer help.

However, this may give rise to another issue; that is, identifying and selecting collocation based on reliable criteria. If exposure is important in developing collocational knowledge, which collocations should be prioritized for incorporation into the language input? In establishing rank between collocational pairs, the last distributional property, association strength, must be considered. Collocations are distinguished from the single-word lexis in that they are defined by the associative relationship between two component words. In representing "the relationship that the lexical items appear with greater than random probability in its context" (Hoey, 1991, p. 7), the simple co-occurrence count may not be the most accurate measure, since a collocational status in the minds of native language users may not always agree with the frequency of the word pair (e.g., "is a," "at the"). Instead, a relative probability, in which the item co-occurs more often than expected in the individual

frequency of constituents (e.g., “strong tea,” “cutting-edge”), can better measure the strength of association between words. Therefore, the probabilistic definition of collocation based on “mutual expectancy” (Firth, 1957, p. 195) and statistical measurement of associative strength should be adopted.

Since association strength is a more abstract construct, which is less observable from the language input, numerous studies have reported that most L2 learners are not sensitive to association strength in the same way as native speakers. In the recent literature, it has been found that native speakers’ judgment of collocability is more related to association strength, while non-native speakers tend to rely on the familiarity of items which correlate strongly with co-occurrence frequency (Durrant & Schmitt, 2010; Siyanova & Schmitt, 2008).

Notably, collocational fuzziness, the mid-level formulaicity between free associations and obviously idiomatic expressions, is argued to present difficulties for L2 learners in collocation learning (Cowie, 1998; Howarth, 1998; Schmid, 2003). This ambiguity of being in the middle of the associative spectrum makes a collocational category less salient and distinguishable than those collocations at the extreme poles. This suggestion poses the question of how middle-strength collocations can be identified, and what kinds of collocation are of middle-strength in the native language model. Since this eventually leads to the matter of measuring association strength based on reliable criteria, it is necessary to examine collocation association strength in reference to the native corpus.

Taken together, we have argued that collocations are a distinctive and significant element of the lexis in English due to their characteristically prevalent, recursive distribution as well as mid-level formulaicity. These distributional features

are seen to play a facilitative role in L1 speakers' fluency and language development in general. Thus, in order to achieve native-like collocational competence, EFL learners may need sufficient language input which fully represents these unique patterns of collocation, similar to that in native English. In addition, the distributional profiles of collocations used in Korean EFL learners' writing will help teachers and material developers to make informed decisions on collocation teaching and selection of lexical items.

Despite its necessity, collocation learning has not been fully incorporated in lexical syllabus in the National English Curriculum in Korea. While the curriculum wordlist has been continuously revised to improve its representativeness based on highly objective criteria, such as word frequency, range, and teacher ratings for item familiarity (Lee & Shin, 2015), the entries still need to be supplemented with lexical items beyond the single-word level. Along with the absence of a curriculum-related, statistically verified collocation list exploiting a large-scale reference corpus data, few collocation studies in the Korean ELT context have attempted to profile collocation use in textbook materials and Korean EFL learner writings. Although some researchers employed the frequency data to identify collocation, we are concerned that this measure may not fully account for the probabilistic nature of word association. For progress to be made, identifying useful collocations compatible with the curriculum word list is mandated. In addition, consultation to a large-scale reference corpus data with statistical measure is crucial for reliable analysis of collocation use in the target corpus.

To complement the previous research, we carried out extensive computations in order to identify statistically verifiable collocations from large-scale

reference corpus data. Furthermore, to investigate how collocations are presented in language input within the Korean EFL classroom, and how Korean EFL learners make use of collocations, we analyzed the corpora of major English teaching materials used in Korean middle and high schools, together with the writings of Korean EFL college freshmen as learning outcome of the middle and high school English curriculum. In choosing these target corpora, we assumed that each corpus represents the input in the EFL context, and learner output. Additionally, to represent the general English and native input, we consulted to multi-billion word native reference corpus with a wide coverage of English texts in varied contexts.

With this advanced methodology, the current thesis aims to examine how collocations are used in the Korean textbook materials and Korean EFL learners' writings. In detail, regarding the extent of use of collocational categories in language input and output in the Korean EFL context, the categories' density and diversity will be examined. In addition, to investigate the intensity of collocational input, the repetition rate will be estimated. Lastly, the last variable, association strength, will examine how strongly associated, probable collocations are used in the English textbook materials and Korean college learners' writings.

1.2. **Research Questions**

The primary focus of the study is to examine the use of collocations in the current English textbook materials currently used in middle and high school and in Korean EFL learner writings. Using a set of target collocations established based on the curriculum wordlist, and statistically verified by the native reference corpus, we will investigate how these collocations are distributed in the textbook corpus and

learner writings. To these aims, the present study was designed to address the following two broader research questions (RQ 1 and 2), each of which will be studied in relation to the four specific subquestions (Subquestions *a* to *d*).

Research Question 1: How are collocations presented in the Korean middle and high school English textbook materials?

Research Question 2: How do Korean EFL college students use collocations in their writings?

Subquestion *a*: With how much density do collocations appear in corpora?

Subquestion *b*: How diverse are the types of collocations used in corpora?

Subquestion *c*: To what extent are collocations repeated through corpora?

Subquestion *d*: How strongly are collocations associated in corpora?

1.3. **Organization of the Thesis**

This thesis consists of three major chapters. Chapter 2 presents the general theoretical background; it first explains the three mainstreams of collocation studies and specifies the primary approach on which this thesis is established. This chapter will further elaborate on methodologies adopted in the corpus-based studies of collocation, followed by research findings on the teaching and learning of collocation. Chapter 3 includes the procedures of corpus compilation and collocation extraction, followed by the statistical measures taken for the data analysis. Chapter 4 and 5 in turn address and discuss the findings with regard to each of the research questions described above. Lastly, Chapter 6 will summarize and suggest the pedagogical implications of the research findings of this thesis.

Chapter 2. Literature Review

This chapter introduces a basic notion of collocation and reviews the methodologies and findings in the previous literature. Section 2.1 outlines the three main frameworks of collocation studies, and the concept of collocations defined based on each approach. Section 2.2. explains the methodologies used in the corpus-based analysis, while Section 2.3 discusses the research findings on the collocations in language input and learner writings.

2.1. Theoretical Approaches to Collocation

Following Gablasova et al. (2017) and Henriksen (2013), a theoretical approach to collocation is categorized as a 'frequency-based approach', 'phraseological approach' and 'psychological approach'. In Section 2.1.1., a definition of collocation based on the frequency-based approach and major interest area of research is introduced. In the following Section 2.1.2. and 2.1.3. describe the phraseological and psychological approach of collocation study, respectively, followed by a statement of the approach selected for the present thesis.

2.1.1. Frequency-based Approach

In a frequency-based approach, with its theoretical ground credited to Firth, collocation is conceptualized as recurrent word combinations. Firth (1957) argued that the meaning and behavior of each word is, to some degree, determined by its collocates, stating that "a word by the company it keeps" (p. 179). In Firthian sense, word choice in natural language is not entirely random, and each word has a different

level of "mutual expectancy" (p. 181) to the other. The probability of co-occurrence of their constituent words and its statistical significance is central to distinguish collocations from free associations. Identified based on frequency data rather than semantic properties, collocation in Firthian tradition may include not only the semantically opaque, idiomatized combinations (e.g., *sweeten the pill, see reason*), but also transparent and less restricted items (e.g., *make a decision, drink coffee, apply for a job, submit a proposal*) (Laufer & Waldman, 2011; Paquot & Granger, 2012).

This frequency-based view has become one of the major traditions of collocation, which motivated quantitative collocation studies. This stream of research has become more fruitful, with the increasing availability of advanced computation and larger corpora since the 1990s. The computation of electronic corpus data necessitates sophisticated statistical methods and further revision of the notion of collocation (Bartsch & Evert, 2014). One methodological breakthrough was Sinclair's 'window-based approach,' which operationalized collocation as "the occurrence of two or more words within a short space of each other in a text" (Sinclair, 1991, p. 170). Based on this approach, the linguistic term is now translated into a few quantifiable parameters, such as *search span* and *co-occurrence frequency*; and collocational relationship is assumed if the two constituents appear together within a span of three to five words to either side of the keyword. Based on this definition, the number of recurrence for each word-pair, the co-occurrence frequency count, is obtained from the corpus.

Most importantly, the Firthian notion of collocation; the habitual, recurrent word combinations with mutual expectancy, is now identified by the level of

association strength. This measure quantifies the attraction between words by comparing the observed co-occurrence frequency against independent frequencies of the constituent words (Bartsch & Evert, 2014). More precisely, the observed frequency is compared in relation to the expected frequency, under the null-hypothesis that the two words might co-occur purely by chance, so as to measure how much the observed co-occurrence frequency exceeds the expected frequency or how likely the null-hypothesis of independence is to be denied (Evert, 2009). As such, collocations are modeled based on the statistical notion of significance, and a number of statistical formulas, also referred to as association measures, have been developed over the years to best estimate the probability of the co-occurrence (e.g., log-likelihood ratio, t-score, Dice-coefficient, and Mutual Information).

2.1.2. Phraseological Approach

In the phraseological approach, collocation is distinguished from the other type of formulaic expressions by its semantic/structural unity and fixedness of form. In this sense, collocations are lexicalized word combinations characterized by its non-substitutability (Manning & Schütze, 1999). That is, the substitution of constituents is restricted, as the noun *shoulders* co-occur with the limited number of verbs (i.e., *shrug*), and the use of the word *decision* is restricted to a certain type of verbs (i.e., *make/reach*) to be used in an appropriate sense. Collocations are not entirely fixed but are subject to some degree of limitation in the choice of components with which they can co-occur, as a constituent of "the large and complex middle ground" between "the extreme ends of the spectrum, free combinations and idioms" (Howarth, 1998, p. 42). According to Nesselhauf (2003), free combinations

are the word pairs in which the two constituents can be freely replaced by the other word (i.e., *want a car*). Collocations, on the other hand, pose restrictions on one of their components; for example, in verb-noun collocations such as *take a photograph* or *take a picture*, the choice of verbs is highly restricted, while nouns are unrestricted. By contrast, idioms are the most restricted type of combination, in that neither verbs nor nouns can be substituted, as seen in *sweeten the pill*. As such, the phraseological view provides a qualitative analysis of collocational units concerning the semantic relations between constituents. Although critics often question the objectivity of such semantic notions as "restrictiveness," this approach cannot be completely disregarded as it may complement the weakness of the other quantitative accounts. Since frequency-based studies may be blind to the underlying semantic and psychological principles, and thus need to be complemented to gain a clearer picture of linguistic patterns associated with certain types of word co-selections with respect to the degree of restrictions (Gablasova, Brezina, & McEnery, 2017).

2.1.3. Psychological Approach

Studies in psycholinguistic approach have attempted to explain how linguistic units and categories emerge in the learners' language system with an influence of instance of usage (eg., Bybee, 1998; Goldberg, 1995; Tomasello, 2009). In such models, language patterns emerge from speakers' usage history and are entrenched to our memory by repeated occurrences. According to Ellis (2002), learners' interactions with linguistic input shaped their L2 system as learners generate regularities from the frequencies exposure to language items. As a result of the frequent encounter with language input, learners acquire the common, highly

predictable sequences of words which mediate language reception and production. This formulaic language, which is also referred to in many ways as “chunks” (Ellis, 2002; 1996) or “semi-preconstructed phrases that constitute single choice” (Sinclair, 1991, p. 110), forms a basic psychological unit which is stored as a whole in the mental lexicon of language users (eg., Hoey, 2005; Schmitt, 2010; Stubbs, 2001). As such, by placing collocations within the domain of speakers’ mental processing, researchers have investigated how formulaic language helps speakers reduce working memory storage and leads to automated processing, fluent production and native-like idiomaticity (Pawley & Syder, 1983; Segalowitz, 2010; Wray, 2005). Furthermore, Ellis (2002) observed that children’s grammar acquisition is a gradual process, which begins with picking up frequent formulas, through low-scope patterns, to ultimately generalizing more abstract constructions. The ready-made units are then hypothesized as a seed to further language development, since they provide ease of processing by reducing working memory demands.

To measure the cognitive construct such as entrenchment, recent collocation studies have adopted corpus-derived metrics of frequency and association measures, and compared L1 and L2 participants in the processing of the distributional data as a signal of collocation (e.g., Durrant, 2008; Ellis, Simpson-Vlach, & Maynard, 2008; Siyanova & Schmitt, 2008). As such, corpus-based research has provided the alternative means to what psycholinguists traditionally had to test relying on the indirect measures of reaction time or eye movement; which Schmid (2016) states as, “If frequency drives entrenchment, the number of times a particular phenomenon occurs within the corpus should be a direct measure of its entrenchment in the cognitive system” (p. 105).

Thus far, three theoretical approaches to collocation studies have been reviewed. Despite the different interest areas of each strand of research, they all provide valuable insight to understand the notion of collocations and its acquisition. First, the frequency-based approach allows a more meticulous, statistical description of the distributional patterns of collocations by analyzing the large-sized language data which is representative of natural English. On the other hand, the phraseological approach provides a semantic profile of collocational association as an intermediary of semantic idiomaticity. Lastly, psycholinguistics gives an explanation of the collocational distribution in relation to language processing and memory storage. This being said, we will now address the main approach taken in the current study and the type of collocations targeted for the investigation.

For the purpose of profiling collocations used in textbook and learner corpora, a quantitative approach was selected to ensure the objectivity of identifying collocations from large-sized corpus data. Using association strength as a key indicator, the current study will rest upon the Firthian definition of collocation. Collocation defined in the current study, therefore, refers to a recurrent combination of words which co-occur more often than their individual frequencies would predict. To identify collocations from the textbook and learner writings, the association strength of word pairs appearing in those target texts will be checked against the reference corpus data; by which means we hope to analyze collocation use with greater statistical accuracy. In addition, we are convinced that an association strength represents the typicality and commonness of the items, and thus will provide data more suitable for pedagogical purposes.

With a frequency-based approach as its theoretical background, the current

study aims to describe the distributional patterning of collocation use in Korean English textbook materials and Korean EFL learner writings. The discussion of semantic restrictiveness of the units, or relationship between language input and output, will be kept very marginal but not entirely excluded, since there are some overlaps between the three approaches which provide pedagogically meaningful insights to understand our data.

2.2. Corpus-based Analysis of Collocation

In the previous section, we stated our choice of a quantitative approach to collocations based on Firthian tradition. This section will further illustrate the statistical measures used for collocation identification (Section 2.2.1.) and major distributional variables investigated in previous literature of corpus-based collocation studies (2.2.2).

2.2.1. Collocation Identification

This section reviews methodologies for collocation identification adopted in the previous literature. The two major approaches to collocation identification will be first reviewed in subsection 2.2.1.1. Then statistical criteria and classification categories will be discussed in the following subsections (2.2.1.2., 2.2.1.3.).

2.2.1.1. Approaches to Collocation Identification

In verifying the collocational status of the word pairs extracted from the target corpus, there are two major approaches commonly used in the previous literature; that is, the top-down approach and bottom-up approach. First, in the

bottom-up approach, researchers extracted word combinations of certain grammar categories (eg., V-N combinations) and checked each candidate against the reference corpus data or collocation dictionaries to confirm its collocability (eg., Laufer & Waldman, 2011; Men, 2018; Parkinson, 2015). On the other hand, a top-down approach, exemplified by the study of Tsai (2015), provides an alternative way for an effective corpus search based on a pre-determined set of target items. Instead of making individual queries against the reference corpus, Tsai first compiled a collocation list for the target 4,000 node words using the statistical process, and then used the list as a reference for collocation identification. A choice of the top-down approach was made based on the researcher's context of teaching: In Taiwan, ELT materials for the English curriculum are designed on the basis of the word list mandated by the national curriculum. Using the items in the list as a search word, a researcher can investigate collocations which are closely related to the curriculum and ELT materials. Furthermore, the collocation list based on the prescribed national curriculum word list may provide a supplementary resource for a single-word level lexical syllabus. Besides, the list of more than 40,000 collocations identified based on the association measure enables a larger-scale, reliable analysis of collocation. Given these advantages, the top-down approach was chosen for the current study, as we aim to analyze collocation use in textbooks and learner writings, based on the 2015 revised national curriculum of English and the revised Korean Basic English Word List (Lee & Shin, 2015).

2.2.1.2. Statistical Criteria for Collocation Identification

A corpus-based analysis of collocation usually involves automatic

collocation identification by requesting a computer to search for targeted node words across corpora and to retrieve possible collocate candidates which neighbor with the node word within a span of n-words. By the term "node" refers to the head of a search word or an item whose collocations we are studying (eg,. *strong coffee* consists of the collocate *strong* and the node *coffee*), while "collocate" means the co-occurring item that is dependent on the node. "Candidate" refers to syntactic patterns based on part-of-speech and chunk annotations, or direct extraction syntax trees (Evert, 2002). In sorting out collocational pairs from the random combinations, statistical methods were used to quantify the collocability of the pairs in combination with other computational methods, such as automatic POS tagging[1], to restrict the search window. To this aim, collocation candidates are ranked based on many parameters, such as frequency thresholds, choice of association measures, and type of syntactic relation to which the item belongs.

One of the biggest challenges in identifying collocation from corpus data may lie in fine-tuning different kinds of frequency-based parameters to reliably predict our intuition of "collocability." To better operationalize this psychological reality of collocation; that is, the collocability which best predicts the perceived collocational units, many different kinds of statistical measures have been developed.

First, the simplistic way of identifying collocation is to count the number of times that two component words co-occur with each other in the corpus; namely, the raw frequency of co-occurrences. The more frequently the words co-occur, the more common or conventional is the collocation. This measure has limitations, however;

[1] A POS tag (or a part-of-speech) tag is a lable assigned to each word to indicate its grammatical categories such as case, or part of speech (noun, verb etc.). Automatic POS tagging is an annotation of corpus data with POS tags using software.

although raw frequency may be directly linked to the notion of collocational recurrence, it does not consider the frequencies of the individual words or corpus size. The measure alone cannot assess whether the observed co-occurrence might have come about by chance (Evert, 2009) or explain the "exclusivity" of collocates (Gablasova et al., 2017). Expressing their dissatisfaction with a rank ordering of multiword phrases based solely on frequency or intuition, for example, Simpson-Vlach and Ellis (2008) have underscored the necessity of combining frequency rankings with a statistically robust measure.

The alternative estimate for collocability is "association strength," which expresses the relationship between the number of co-occurrences as opposed to the individual frequency of each component word. This measure is often assumed to represent the "salience" which is more easily recognized, acquired and stored as a unit (Gablasova, Brezina, & McEnery, 2017) and thus have more "pedagogical relevance" (Ackermann & Chen, 2013). While there are a number of association measures, three indices; t-score, Mutual Information (MI) score, and logDice score, are particularly noteworthy.

First, the t-score has been one of the most widely-cited measures in corpus-based collocation research. It is calculated as an adjusted value of collocation frequency: random co-occurrence frequency is subtracted from raw co-occurrence frequency and then divided by the square root of the raw frequency (Gablasova et al., 2017). The score may often be biased toward the frequency of individual words and highlights frequent combinations of words; as Durrant and Schmitt (2009) noted, t-score rankings are similar to the rankings from raw frequency. This is evidenced by the fact that the bigrams (two-word combinations) with the highest t-scores in the

BNC are “is the,” “to a,” and “and a” (Gablasova et al., 2017). Similar to raw frequency, t-scores are directly dependent on corpus size and t-score cannot be used as a standardized scale for the purpose of direct comparison of different corpora (Hunston, 2002).

The MI-score has been another commonly used standard measure for word associations (e.g., Hunston, 2002; Manning & Schütze, 1999; Siyanova & Schmitt, 2008). With measures ranging from zero to 17, the MI-score informs how strongly and likely the words are associated with one another. Manning and Schütze (1999) also noted that the MI-score is particularly useful for detecting “interesting collocations,” which are less frequent but strongly associated word combinations. Accordingly, many researchers have reported that the MI-score may perform as a better predictor of native-like command of collocations or advanced level of phraseological complexity (Bestgen & Granger, 2018; Durrant & Schmitt, 2010; Paquot, 2018), despite its well-known bias toward low-frequency data (Bartsch & Evert, 2014).

Although it has gained popularity as an association measure for retrieving “rare, native-like” items in many collocation studies, the MI-score is not without shortcomings. Since the measure highlights the exclusive “rarity” of the co-selection, it tends to downgrade the high-frequency items which are still high in usefulness, especially for L2 learners. Moreover, as the MI-score tends to overestimate infrequent combinations, this measure may risk the pedagogical value by emphasizing lexical knowledge of infrequent, specialized individual lexical items which L2 learners may not yet have acquired (Gablasova et al., 2017). For example, the collocate candidates of “baby” with the highest MI-scores are “womb,” “aborted,”

"breastfeeding," "diaper," and "abort." If learners' collocation knowledge is tested based on such items, the test scores may be misleadingly affected by single-word vocabulary size, rather than reflecting learners' collocational knowledge.

As an alternative to the MI-score, the logDice measure has been attracting attention in recent corpus-based studies (Frankenberg-Garcia, 2018; Frankenberg-Garcia, Lew, Roberts, Rees, & Sharma, 2019; Gablasova et al., 2017; Schmid, 2016), since it highlights "exclusivity in users' co-selection of words" without penalizing the high-frequency collocates. Rychlý (2008) has also noted that the logDice measure allows a "reasonable interpretation, and scales well on different corpus size" (p. 7) and can be used to identify good collocations. Examples of collocate candidates with the highest logDice scores are: "born," "birth," and "boomers," which may provide more relevant items than those retrieved from the MI-score. In the current study, therefore, logDice scores will be used as the major statistical criteria to verify the saliency of the collocation candidates retrieved from the reference corpus and as a measure for their association strength. The current study, for a pedagogical reason, would avoid the excessive emphasis on the rarity of combinations or large vocabulary knowledge expressed by high MI-scores. Instead, the primary concern will be the more reasonable collocational relationship based on the logDice measure, which seems to represent more desirable and realistic objective for collocation learning in the general language course for EFL Middle to High school students.

2.2.1.3. Classification of Collocation Subtypes

When searching collocation from the corpus data, syntactic relations between node and collocate is widely used as linguistic parameter (e.g., node as noun modified by adjective collocate, node as noun which is a subject/object of verbal collocate; node noun with nominal premodifiers as its collocate). According to Bartsch (2004), collocations are "lexically and/or pragmatically constrained recurrent co-occurrences of at least two lexical items which are in direct syntactic relation with each other" (p.76). While there are many ways to categorize collocations based on syntactic category, as exemplified in Table 2.1 below, the verb+noun collocations (VNCs), noun+noun collocations (NNCs), and adjective+noun collocations (ANCs) are the most commonly used syntactic combinations. In corpus-based studies, collocations of different syntactic relations are usually treated as an independent category.

Table 2.1
Classification of Collocation Subtypes

BBI dictionary	Hausman	Wordsketch
Adjective+Noun	Adjective+Noun	Adjective+Noun Adjective+Preposition
Noun+Verb	Noun+Verb	Noun+Verb
Noun+Preposition+Noun	Noun+Noun	Noun+Noun, Noun+Preposition+Noun Noun+Conjunction+ Noun
Adverb+Adjective	Adverb+Adjective	
Verb+Adverb	Verb+Adverb	Verb+Adjective
Verb+Noun	Verb+Noun	Verb+Noun Verb+Preposition

(Adapted from Chang, 2018; Seretan, 2005)

The first reason why most corpus-based studies have categorized collocations based on grammatical relation is its practicality. Since the corpus is usually annotated with information of the word classes of each word, researchers can automatically filter out the indirect, accidental combinations by specifying the grammatical categories of its sub-type (Evert, 2009).

In addition to this practical motivation, syntactic categorization is important because they differ in frequency, saliency and learnability (Henriksen, 2013). Granger and Bestgen (2014) also pointed out that merging the categories would run the risk of overlooking the subtle differences in collocation use. For instance, VNC type has been reported as the primary source of errors for many L2 learners (Laufer & Waldman, 2011; Men, 2018; Nesselhauf, 2003; Peters, 2016) while learning of ANCs is known to take place at the early stages of language development, preceding that of NNCs (eg., Biber, Gray, & Poonpon, 2011). On the other hand, writers' choice of the syntactic sub-types may differ on stylistic grounds and the genre of the texts. It has been acknowledged that nominal collocations (ANCs and NNCs) are of growing significance in modern English (Biber & Clark, 2002; Biber & Conrad, 2009; Biber & Gray, 2011). In their cross-sectional analysis of English text, Biber and Gray (2011) found that there is a rapid increase in the frequency of noun phrases (eg., noun-noun sequence) and a relative absence of verbs in modern academic prose and news articles. They explained that this trend reflects present-day informational writing discourse in which writers need to deliver the vast amount of information within a page limit, and thus compress information within the complex NPs to achieve economy of language. Similarly, Ackermann and Chen (2013) found that noun phrases dominate academic registers, forming the largest category.

Thus far, we reviewed the sub-categories of collocations based on the word classes. Since many studies have acknowledged the pedagogical value of VNCs, NNCs, and ANCs, in terms of its learning difficulty, growing utility in modern discourse, or its large population among the whole set of collocations, the current study will aim to investigate the use of these three sub-types in the textbook materials and learner writings.

2.2.2. Distributional Variables of Collocation Use

Before we discuss the distributional profiles of collocation, it is necessary to explain the underlying assumptions. First, as a corpus-based study of collocation, the methodological framework of the present study is established on the notion of corpora-as-input and corpora-as-output (Schmid, 2016). "Corpora-as-input" represents a sample of the language use that members of a particular speech community are exposed to during the acquisition, based on which, learners' mental representation is shaped and reshaped through the lifetime of language acquisition. As a model of linguistic input, a corpus is assumed to approximate the language input ranging from those adapted for children to those encountered in adulthood, for which the author argued that large, mixed-register corpus such as BNC has some limitation, but reasonably close to the idealized member of the community. Conversely, "corpora-as-output" means a corpus is a sample of the language use of a particular group of speakers representative of a particular speech community. The observation of corpora, therefore, enables us to model mental linguistic representations.

Another assumption for the corpus-based analysis of collocation is that the human mind is sensitive to the statistical information of language input, including

frequency, variability, distribution, and co-occurrence probability (e.g., Ellis, 2012; Erickson & Thiessen, 2015; Yi, 2018); and language acquisition is construed as a part of the general learning process in which learners update their current knowledge by discovering the systematic regularities embedded in the input and its distributional properties (Frost, Armstrong, Siegelman, & Christiansen, 2015).

Although these usage-based and cognitive SLA theories may not be the central framework of the current study, they provide important insight on how we should analyze and understand corpus data in the pedagogical context of English teaching and learning. Hence, while primarily focusing on describing collocation use in the two independent corpora compiled at different time periods, the present thesis attempts to interpret the textbook corpus as a sample of major language input Korean EFL learners are exposed to and the Korean EFL college freshmen students' writing corpus as a sample of learner production by those who are assume to complete six years' of middle and high school English curriculum. Additionally, native reference corpus will represent a sample of natural English as well as L1 input based on which native speakers would develop their language competence.

Based on this assumption, we will examine how extensively and intensively collocations are represented in the language input in the regular English classrooms and how productively and intensely Korean EFL students are able to use collocations in their writings after finishing the regular English courses in the middle and high schools. As Boers and Lindstromberg (2009) contend, the extensive exposure to many different types of collocations seems to be related to fluent processing, or reception, of collocational inputs, whereas the intensive usage is crucial to entrenching the word pairings durably in the memory, and increases the fluency of

production. In many previous studies, the extensive/productive use of the word combination is termed as collocation density and diversity, while the intensive use is generally represented by repetition and association strength of collocations.

The first variable, density, is one of the most commonly investigated distributional features. It represents how many collocation tokens are presented in the corpus. The higher density indicates that the text contains a large number of collocational pairs. In most literature, the measure is operationalized by a proportion of collocations among the various grammatical combinations, or the token frequency of collocations within the whole corpus, even though its operation is not always understood in the same terms by different authors[2].

The next variable is collocation diversity; it represents how many different collocation types are introduced in the text. Higher collocational diversity indicates a wider range of collocational repertoire introduced in the texts. In measuring collocation diversity, some researchers have calculated the number of unique collocation types in relation to the word counts of the corpus, while others have modified the measure of lexical diversity rate, type-to-token ratio (TTR), into a measure for collocation (CTTR)[3], by calculating the ratio of collocation types per collocation token counts. Paquot (2018) has used "Root type-token ratio (RTTR)" to

[2] Authors differ regarding to the operationalization of "proportion of collocations in the texts". Laufer and Waldman (2011) counts the number of collocations in relation to 1) token frequency of running words and 2) total noun tokens. The former is also used in the study of Kjellmer (1991), while the latter was exemplified by Stubbs (2001) who identified 47 word counts with an initial 'f' among 1000 word samples as phraseological units.

[3] Durrant (2008) points out that "Collocation type token ratio (CTTR) = Collocation type counts / Collocation token counts" indicates how many times each collocation recurred throughout the text, rather than how many different collocation types appeared within a set corpus size. It is suggested that CTTR is more related to the level of repetitions, rather than collocation diversity.

minimize the corpus size effect.

These two measures represent the level of extensive use of collocations by counting the relativized token or type frequencies of the "overall" collocations (e.g., how many instances of collocational pairs appear in the text? How many collocation types appear in the text?), but they do not necessarily represent the intensity of use of "each" co-occurring pattern (e.g., How "repetitively" do textbook materials and learner corpus provide/produce individual collocation type? How are strongly associated collocation types used in the text?), which is primarily a matter of specific expression.

The third major variable which previous research has evaluated is the repetition rate. Repetition rate indicates the degree to which how intensely the corpus represents individual collocation types. Estimating how many times the individual collocation type recurs throughout the entire text, the role of repetition has been highlighted especially in usage-based studies, which assume that greater exposure to language input leads to a deeper entrenchment of the target items (e.g., Arnon & Snider, 2010; Conklin & Schmitt, 2012; Ellis, 2002) and learners' increased sensitivity to collocation (Durrant, 2014; Durrant & Doherty, 2010; Durrant & Schmitt, 2010). In addition, repetition has drawn attention of the researchers who investigated collocation use in teaching materials (e.g., Jui-Hsin, Wang, & Good, 2007; Koya, 2004; Jinkyong Lee, 2015; Tsai, 2015) or the researchers who examined input modification or repeated exercise for implicit and explicit learning of collocation knowledge (e.g., Boers, Demecheleer, Coxhead, & Webb, 2014; Boers & Lindstromberg, 2009; Sonbul & Schmitt, 2013; Toomer & Elgort, 2019).

However, the simple measure of raw co-occurrence frequency count may

not be a sufficient indicator to gauge whether the target items are “truly collocational” in the system of natural English (Evert, 2009). There would therefore seem to be a definite need for employing more sophisticated assessments based on statistical measures to compare them on the wide spectrum of ‘association strength’, the last measure.

The last feature, the association strength, has been used in much of collocation research to estimate the intensity of association between the component words, indicating how strongly each node word is associated with its collocates to constitute a collocation. In psycholinguistic terms, association strength is often translated into “predictability”, “probability” or “salience” of the formulaic language which allows processing advantage; The more predictable the association is, the faster can the formulaic sequence be processed (e.g., Conklin & Schmitt, 2012; Durrant & Schmitt, 2010). As stated in the previous section, it is operationalized and ranked by the probability of two words co-occurring together against the likelihood of their occurring separately (Schmitt, Candlin, & Hall, 2010).

Association measure is different from the previously-mentioned distributional properties, in several ways. First, it estimates the relationship between the constituent words “within the unit” rather than collocation “as a unit.” It thus ranks the target items based on the continuum of the probabilistic scale, instead of categorizing them into the dichotomy of collocation/non-collocation. The association scores, therefore, could provide a more detailed profile of stronger/weaker collocations (Chen, 2019), indicating the intensity of the word-pairing ranging from the most strongly associated to completely independent combinations.

Next, established based on the large reference corpus in most of the studies, the association measure usually serves as a reference scale which accesses the collocability of target candidates based on the native standard. To the knowledge of the author, there have been three major methodologies in previous research that have utilized association data. The most common methodology is to use this measure as a threshold to identify statistically verifiable collocations (e.g., Tsai, 2015). Secondly, more recent studies have looked more closely into the level of association scores by analyzing their distribution (e.g., Bestgen & Granger, 2018; Chen, 2019; Durrant & Schmitt, 2009b; Granger & Bestgen, 2014). In doing so, "band-based" method has often been used, where collocations in the corpus are divided into a certain number of association score bands and the proportions of collocations in each level are compared. Alternatively, the average association scores of collocations have been compared between groups. Lastly, different kinds of association measures have allowed researchers to test related psycholinguistic constructs or linguistic competence (e.g., Durrant & Doherty, 2010; Durrant & Schmitt, 2010; Paquot, 2019; Siyanova-Chanturia, 2015; Siyanova-Chanturia & Janssen, 2018; Siyanova & Schmitt, 2008b); for example, participants' subjective judgment of the collocability have been compared with the MI-scores or raw frequencies established in the reference corpus to measure the participants' sensitivity to frequent or rare collocations.

2.3. Collocation in English Teaching and Learning

To investigate how collocations are represented in a pedagogical context, this section overviews previous research findings on the collocation use in English teaching materials (Section 2.3.1.) and L2 learner production (2.3.2.).

2.3.1. Collocation Use in English Teaching Materials

This section introduces previous findings on the collocation use in the English teaching materials. Prior to discussing collocational input presented in the teaching materials, it will be helpful to briefly review distributional features of collocation in natural English which was touched on in Chapter 1. First, it is now well established from a variety of studies that natural English has a large coverage of collocation and other formulaic languages. For example, in their analysis of the proportion of the four types of prefabs (e.g., *a waste of time*, as lexical prefabs, *in spite of* as grammatical prefabs, *yes I think so* as pragmatic prefabs, *isn't* reducible prefabs) in LOB corpus, Erman and Warren (2000) revealed that prefabs constitute up to 58.6% of the spoken English discourse and 52.3% of the written discourse. According to Biber and Conrad (2009), collocations accounted for 30% of the spoken corpus, and 21% of the written academic corpus. Similarly, Altenberg (1991) also noted: "roughly 70% of the running words in the corpus form part of recurrent word combinations of some kind" (p. 128). As such, collocation is a part of the ubiquitous phenomenon of formulaic language in the natural English and the extensive exposure to collocational input in various contexts is known to promote strong associative links in the native language system (e.g., Ellis, 2001; 2002; 2003; 2006).

Next, repetition is another distributional feature which represents collocational idiomaticity in natural English. According to Ellis (1996), with repetition, sequences of words that were previously independent come to be processed as a single unit or "chunk". Once the memory trace of the word association is formulated in language learners' minds, his or her collocational knowledge becomes entrenched through repeated encounters with target items. In the similar vein, Sinclairs' "idiom principle" accounts for native speakers' reliance on "semi-constructed phrases that constitute single choice" (Sinclair, 1991, p. 320) and has given psychological interpretation to the high co-occurrence frequency as evidence of native-like idiomaticity. If the repetitive exposure to collocation is critical in developing fluency of its production and processing (Pawley & Syder, 1983) and thus in achieving the basic communicative purpose (Wray, 2005), then teaching materials for EFL learners also need to be assessed on the repetition and input frequency they give to individual collocation types.

These studies highlight the prevalence of collocational input in natural English and suggest idiomaticity as a major principle of natives' language production. As much as collocational knowledge is central to the native speakers' attainment of fluency and idiomaticity (Pawley & Syder, 1983), its significance applies the same to L2 learning as well; collocational competence enables L2 learners to make native-like idiomatic choices while focusing the remaining cognitive energy into more creative production, and to understand the polysemous, connotational meaning (Henriksen, 2015). Despite this advantage, most EFL learners do not seem to gain sufficient exposure to the target language as do L1 speakers (Fan, 2009).

If sufficient exposure to collocations through input is beneficial for L2

learning, one may ask how collocation-rich the text should be to approximate native norms and meet learners' needs. However, the findings have remained inconclusive. While some researchers highlighted the importance of the amount of exposure to collocational items in developing learners' sensitivity to native-like collocational patterns (e.g., Durrant & Schmitt, 2010; Ellis, 2002), others hold that repetition of individual items rather than the conflated frequency of single exposures is more crucial for consolidating the memory trace (e.g., Koya, 2004; Tsai, 2015). Similarly, previous research varied in their position as to the level of collocational diversity in the ELT materials. As Lewis (1997) pointed out "if there are too many examples, too many possible answers or the items are badly ordered, the exercise implies a rather perverse activity based mainly on guesswork" (p. 88). Many authors contend that more diversity may not always be pedagogically beneficial (eg., Boers & Lindstromberg, 2009; Groom, 2009; Koprowski, 2005; Tsai, 2015). For example, Boers and Lindstromberg (2009) suggest: "it would be naive to count on students being able to acquire large numbers of chunks incidentally" (p. 68), highlighting that it was the repetitive exposure to the target items that may be a more significant determiner of a successful intake than the total number of unspecified items.

In addition to the amount of collocation use in the ELT materials, the quality of collocation may be in question. For example, Koprowski (2005) investigated the usefulness of lexical phrases across the coursebook written by major international publishers. When measured by the usefulness score based on frequency and range data, a quarter of collocations presented in the textbook were found to be less useful in natural language. Interestingly, the developers promoted these coursebooks based on good coverage of multi-word items, and pedagogical usefulness may be

compromised as a result of the tendency to introduce wide coverage of items and theme-centered organization of lexical items. Furthermore, the inconsistent and ambiguous selection criteria led to a lack of agreement to the extent to which less than 1% of selected items are shared by three coursebooks. The study suggests that while many coursebook designers recognize the significance of collocation in language learning, the near absence of unified lexical syllabus focusing on the multi-word unit made it harder for them to represent such ideas into practice in a systematic way. Collocation items, therefore, are often chosen by the subjective judgment of publishers, and thus may not provide a quality input essential for language learners.

In the Korean EFL context, there have been fewer studies which focused specifically on collocation use in the textbooks. Among the few are Kim (2004), Choi and Chon (2012), who examined the collocation use in 10th grade English textbooks. As weak association strength of the most frequent collocations in the textbooks such as “Good boy, school students, really enjoy,” and “volunteer work, good grade, get grades, use cellphone, etc.,” indicates, it was shown that textbooks over-represented free-association collocations with little pedagogical value. As textbooks play a dominant role as a major learning resource in Korean middle and high schools, more quantitative analysis of how collocation is used in the materials will be of definite necessity.

2.3.2. Collocations in Learner Writing

This section reviews research findings on the use of collocation in learner writings. Regarding distributional patterns of density and diversity, reports on learners’ tendency to overuse and underuse collocations will be discussed

(Subsection 2.3.2.1., 2.3.2.2.). The overview of recent findings on association strength will be then presented (2.3.2.3.), followed by different patterns in distribution by collocation subtypes (2.3.2.4.).

2.3.2.1. The Overuse and Underuse of Collocations

Literature has emerged that examines whether collocations are as prevalent a phenomenon in the non-native learner corpus as they are in native writings; however the result has remained inconclusive. Some studies report higher collocation density in learner writings, while others have found the opposite trend.

First, for those who found severe underuse of native-like phrases in learner production, it was suggested that L2 learners tend to process words in a fundamentally different way from their native counterparts, and that learners may not develop native-like collocational capacity. Kjellemer (1990), for example, claimed that even advanced learners construct messages from individual words rather than from prefabricated patterns. Wray (2002) also maintained that learners do not have native-like sensitivity to a collocational relationship in the language input, and highlighted the intentional learning of L2 collocations. Supporting this view, several corpus-based studies reported on learners' underuse of collocations. Howarth (1998) examined the V-N collocations (VNCs) and found that only 25% of collocation in learner writing was restricted collocation, which fell below the 31% in the native corpus. Likewise, Laufer and Waldman (2011) found that Israeli learners tend to use VNCs less frequently (5.9%) than their native counterparts (10.2%). In a similar vein, Korean EFL learners are reported to have the same problems, especially with VNCs. Learners' use of VNCs is not only restricted in its

type diversity, but could also be the most erroneous lexical domain for Korean learners (eg., Choi, Chon, & Han, 2015; Sung, 2017).

By contrast, many studies have reported conflicting evidence against the claims that learners are deficient in collocational capacity; for instance, Chang (2018) and Tsai (2015) found no significant difference in the proportion of VNCs in the native and non-native writings, disproving the previously held view on lack of idiomaticity in learner productions. In their investigation of ANCs used in Russian learners' writings, Siyanova and Schmitt (2008) identified around half the learner production as frequent and strongly associated collocation, which is almost congruent with the native data. This result shows that learners do have productive collocational knowledge and that the learners' language system may be identically idiomatic to that of their native counterparts.

The amount of collocation in learners' writings could be related to many possible variables. It may differ on the amount of exposure to collocational items through language input and learners' familiarity with the item; for example, in his analysis of the use of ANCs in the writings of undergraduate and postgraduate non-native students in EAP courses, Durrant (2008) reported on learners' tendency to use higher-frequency collocations more productively. The result indicates that they may have sensitivity to co-occurrence frequency and that the amount of exposure to target collocations is significant in developing L2 learners' collocational competence. On the other hand, L1-L2 similarity may also affect the learners' production of certain collocation types. Parkinson (2015) found that NNCs were used with the highest frequency in Mandarin EFL learners, explaining that this might be due to similar N-N constructions in their L1. It was noted that L1 can both positively and negatively

affect the use of collocation. The findings showed that the overuse of the N-N collocation led to the underuse of the other subtypes even in the inappropriate context.

2.3.2.2. Limited Collocation Diversity

Several lines of evidence suggest that learners' use of collocation is characterized by a limited repertoire. Through the comparative analysis of the native and non-native corpus, Granger (1998) and Lorenz (1999) reported that non-natives tend to overuse certain words (e.g., say, think) in an active structure compared to natives, while clinging to a few fixed phrases with which they feel confident. Similarly, Durrant (2008) and Men (2018) compared the diversity rate between learner groups of different proficiency and reported that higher collocation diversity rate was found in the higher proficiency learner groups.

Countering the claim held by Durrant and Men, different trends have also been reported that lower collocation diversity represents non-native-like competence. In the data reported by Durrant and Schmitt (2009), for example, collocation diversity differed on the disciplines. In the disciplines of arts and humanities, journal articles have shown a lower diversity rate than 1st year undergraduate students, while the opposite trend was found in science writing. In the study of the use of collocation in the academic register (Frankenberg-Garcia, 2018), no significant differences were found in the number and type of collocations available to L1 and L2 EAP writers, demonstrating that native writers do not always present more productive collocation repertoire than L2 writers. Without sufficient opportunities to assimilate the lexical conventions, L1-English undergraduates supplied far fewer collocations than the experienced L2 academics. Additionally, no systemic increase in collocation

diversity was observed from one proficiency level to the next one, in the study by Paquot (2019). The author, however, did not entirely deny the relevance of the diversity variable as a measure for L2 development. Since unsophisticated metrics used to quantify collocation diversity might be the major cause of inconsistent results, more empirical evidence is required to describe the development of collocation complexity measured by other metrics, such as association strength.

2.3.2.3. Weaker Association Strength

The association measures have been increasingly popular in L2 studies, since these measures allow researchers an empirical basis to define the psycholinguistic construct such as idiomaticity or commonness. For example, Durrant and Schmitt (2009) compared L2 writing in comparison to native writing, and found that L2 writers tend to underuse less common, strongly associated collocations which are identified by high MI-scores (e.g., densely populated, bated breath), while advanced learners tend to use collocations of high association level. When compared to native counterparts, non-native speakers' (NNS) writing gained higher t-score and lower MI-score than native writings. Similarly, Siyanova-chanturia (2015) also suggested association strength as a relevant variable in determining the level of difficulty. In the investigation and report of beginner Chinese learners of Italian, learners with lower proficiency showed preference in the use of ANCs with high frequency and low association strength, and progress was found in accordance with learners' proficiency.

In their comparison of native and non-native judgments of the frequency level of ANCs, Siyanova and Schmitt (2008) found that natives reliably

distinguished infrequent collocations from frequent ones, while non-natives were not as accurate as those of native speakers (NS); by underestimating the frequency of common collocations and overestimating the frequency of uncommon collocations. When it comes to the differentiation between mid- to high-frequency collocations, NS were able to make a finer distinction.

Another strand of study has used association measures to automatically distinguish learners' language proficiency in relation to the collocational complexity observed in their writing. Paquot and Naets (2015) hold the view that the phraseological dimension is significant in the development of language proficiency, especially for learners of the upper-intermediate to an advanced level. It was demonstrated in the study of Paquot (2018) that phraseological complexity measured by statistical indices of collocational strength in learner text increases steadily and significantly from upper-intermediate to advanced learners in higher education. His study reports that phraseological complexity explains 25% writing scores by human raters, which is more powerful than traditional measures of syntactic and lexical complexity.

2.3.2.4. **Difference in Collocation Use by Subtypes**

Among a wide range of lexical collocations, VNCs have been the primary concern in considerable literature, and have been reported as the primary source of errors for many L2 learners (eg., Chen, 2017; Laufer & Waldman, 2011; Men, 2018; Nesselhauf, 2003). Compared to well-established studies of VNCs, only marginal attention has been paid to ANCs and NNCs, of which pedagogical value should be revisited. In studies of writing assessment, complex nominals per clause are reported

as one of the best predictors for writing complexity or quality (Crossley & McNamara, 2014; Lu, 2011). While both ANCs and NNCs constitute the complex noun phrase, report findings have shown differences in the use of the two types by learner proficiency. Parkinson and Musgrave (2014) reported on proficient writers' preference on noun modifiers while less proficient groups relied heavily on adjective modifiers, confirming the hypothesis suggested by Biber et al. (2011) that attributive adjectives are acquired at the early stage of language development. Granger and Bestgen (2014) also found that intermediate and advanced learners are significantly different in their use of strongly associated, lower frequency NNCs. They added that the difference between the two groups is much more pronounced in the intermediate corpus, exhibiting the highest scores for ANC types. Regarding their function in writing and different usage at different proficiency levels, the NNCs and ANCs used in the reference, textbook, and learner writing corpus are therefore worthy of investigation in the current study.

As discussed in the current chapter, distributional information such as co-occurrence frequency and association strength has been used for the identification of collocation in the Firthian tradition. While co-occurrence frequency itself is indeed an important measure, its shortcomings should be complemented by the probabilistic measure to estimate association strength. Despite its statistical robustness, the measure has not yet been employed by many researchers.

Next, while studies on different kinds of learner production and L2 collocation use have been relatively prolific in the previous literature, empirical evidence is still in need of the data from language input which learners would be exposed to in naturalistic environment and EFL context. To fill the research gap,

there seems to be an urgent need to investigate the representation of collocation in general English as a natural input and in the teaching materials which is a major source of language input for most of the EFL students. Particularly, to represent a fuller picture of the use of collocation in English language in particular and use it as a native baseline, the current study attempts to profile the large-scale reference corpus, rather than the small-sized text samples from native writings.

Lastly, it has been pointed out that a relatively less interest has been paid to various subtypes of collocations. Since each subtype shows different patterns in its distribution and developmental stages in natives' and L2 learners' production, a comprehensive investigation is needed to avoid overgeneralization.

As a way of bridging the gap, the present study will identify collocations based on statistical criteria involving association measure, and investigate the use of three collocation subtypes in both target and reference corpus to represent both language input and output by native speakers and L2 learners.

Chapter 3. Methodology

The main focus of this corpus-based research is to examine collocation use in the English textbook materials and Korean EFL learner writings. This chapter outlines the methodology used to analyze collocation use in the textbook and learner corpus. Section 3.1. describes the general profiles of the three corpora used. Section 3.2. introduces measures and tools selected in the current study. In Section 3.3., the main procedures for data analysis will be explained, followed by the summary of the collocation database and process in Section 3.4.

3.1. Corpora

This section overviews the general profiles of three corpora used in the current study. First, the selection of the native reference corpus is explained in Subsection 3.3.1. The following Subsection 3.3.2. introduces the middle and high school English textbook corpus and then Subsection 3.3.3. describes the written productions of Korean EFL learners.

3.1.1. A Reference Corpus

Following the Firthian approach, this study will identify collocation based on the level of association strength and frequency of co-occurrence. The reference corpus was chosen to represent "general English", and to check the two collocation measures for sets of candidate word pairs found in the target corpus. The measures checked in the reference corpus will determine the collocational status of the word pairs by estimating how frequently and strongly the word pairs are associated with

one another in the general English texts.

In choosing the reference corpus, two main considerations guided the decision. First, the corpus should provide good examples of contemporary English used in everyday, standard, formal and professional context. Second, sufficient coverage of the English language should be ensured and thus the corpus is to be sizable enough to model the language input in a natural environment. Based on these criteria, a recently compiled corpus, Sketch Engine for Language Learning (SkELL), is selected for the reference corpus. Table 3.1 presents a general profile of the SkELL corpus.

Table 3.1
SkELL Corpus

Subcorpus	Tokens	Used	Percentage	Corpus Information
Wiki	1.6G	403M*	39%	Many articles cover geographical and historical domains. Thousands of articles not containing fluent text were filtered.
English Web Corpus 2013 (enTenTen 2013)	22G	321M	31%	Documents were obtained by a web crawler, which queried seed URLs and downloaded web pages. The search results are sorted using the technology specialized in collecting linguistically valuable web content, followed by the process of downloading, cleaning, and converting to plain text.
WebBootCated	105M	77M	7%	Texts in the web pages were queried from the search engine using approximately 100 million seed words.
Timestamped Web Corpus	900M	146M	14%	News source obtained from crawled a list of RSS feeds
BNC	112M	90M	9%	100 million words of British English, written and spoken. Contains written component (informative, imaginative), spoken component (private speech, public speech, monologue), etc.
Total	1G		100%	

(Adapted from “English corpus for SkELL | Sketch Engine,”n.d.)
* The figures are rounded up to millions

The corpus was compiled specifically for the purpose of English language learning and contains 57 million sentences with 1 billion word tokens and 3.6 million types from the web-crawled corpora of English which now becoming common place in recent corpus linguistics, featuring news, academic papers, Wikipedia articles, open-source fiction books, webpages, discussion forums, blogs (Baisa & Suchomel, 2014; Baroni, Kilgarriff, Pomikálek, & Rychlý, 2006). It also contains BNC corpus which has been commonly exploited as a reference in many previous studies. Scholarly interest in SkELL as a resourceful language database has been increasing, with its pedagogical usefulness being recognized by other researchers (e.g., Barrs, 2016; Frankenberg-Garcia et al., 2019; Hirata & Hirata, 2018; Williams, 2019)

3.1.2. Korean Middle and High School English Textbook Corpus

To analyze collocation use in the textbook corpus, we compiled a corpus with newly published textbooks based on the 2015 national curriculum. In compiling the ELT material corpus, the design of the corpus was determined by the two criteria. First, the corpus is established to represent the Korean EFL learners' language experience in English classrooms. Most Korean EFL learners generally use one textbook (up to 10th grade), or two in each grade level, as most High schools provide English 1 and English 2 in their local curriculum as a selective course. The second criterion was corpus size: to identify meaningful patterns, the corpus should not be too small. In deciding sampling size, we followed Durrant and Schmitt (2009) who stated that an "extended piece of writing is desirable, for statistically robust trends may only emerge in longer stretches of writing where larger numbers of collocation can be identified" (p. 161). It was thus considered sensible to classify textbook

materials by publishers and merge the reading passages of the same publishers from Middle school 1st grade to English 2 (High school) textbooks, to represent the lexical input that individual learners may encounter throughout the middle and high school curriculum.

The corpus includes approximately 307 reading passages from English textbooks covered in the Middle (Grade 1 and 2) and High schools (High School English, English 1·2). The textbooks under study were recently published based on the 2015 revised national curriculum. For comparability, we selected 8 out of 12 publishers that provide textbooks of all grade levels, and compiled them into a separate corpus. As summarized in Table 3.2, each of eight textbook corpus contains on average 22,300 words of 4,388 different types. Among these 4,388 word types, 713 items are the nouns from the curriculum-based wordlist and occurred 3,315 times in each textbook corpus. In other words, learners who use textbooks of each publisher may encounter on average 713 types of nouns from the curriculum wordlist for 3,315 times throughout the middle and high school curriculum.

Table 3.2
General Profile of Textbook Corpus

	Number of texts	Single-word		Node word	
		tokens	types	tokens	types
Textbook Corpus	37* (2.26)	22,230 (3256.14)	4,388 (530.69)	3,315 (552.36)	713 (72.91)
Reference Corpus	57,143,446	1,041,138,575	3,602,507	57,143,446	1,718

* Average (S.D)

3.1.3. Korean EFL College Students' Writing Corpus

In analyzing Korean EFL learners' written production, Yonsei English Learner Corpus (YELC) was chosen as the target corpus. Containing 1,085,828 words from 6,572 essays written by 3,286 college freshmen, the corpus consists of argumentative writings on topics of various social issues (e.g., smoking in public places, animal testing, using mobile phones while driving) and narrative writings on personal interests (eg., the favorite extracurricular activity in high school) from nine different levels (CEFR A1 to C2) of participants (Rhee & Jung, 2012). In terms of the sampling size, we decided to split the corpus into five sub-corpus in order to lessen the problem of disguising differences between individual texts, and also to utilize standard inferential statistics. As summarized in Table 3.3, 5,975 essays were randomly sampled from the writing collection, and grouped into five subcorpora, each of which consists of 1,195 texts with 196,433 tokens and 10,137 types on average. The average of 27,115 tokens of nouns with 1,025 different types were found to match with a curriculum wordlist and thus identified as node words for collocation. Among the 1,718 target nouns from the curriculum wordlist, the average of 1,025 types of nouns were used in the learner corpus, occurring 27,115 times per each sub-corpora.

Table 3.3
General Profiles of Learner Corpus

Corpus	Number of texts	Single-word		Node word	
		Tokens	Types	Tokens	Types
Learner Corpus	1,195	196,433 (2452.54)	10,137 (69.87)	27,115 (546.48)	1,025 (7.89)
Reference Corpus	57,143,446	1,041,138,575	3,602,507	57,143,446	1,718

* Average (S.D)

3.2. Measures and Tools

This section presents the measures and instruments used in collocation identification, and analysis of its distribution within a corpus; First, section 3.2.1. describes the statistical criterion used to identify collocations. Then section 3.2.2. explains the calculation formulae to operationalize the distributional variables of collocation density, diversity, repetition, and association strength. Methods for statistical analysis will be illustrated in section 3.2.3., followed by an introduction of software in the last section 3.2.4.

3.2.1. Statistical Criteria for Collocation Identification

In identifying collocations from corpora, statistical parameters are applied. On statistical co-occurrence data from the reference corpus, we established the cut-off points for collocation verification. The criteria were used to identify true collocation from the random combinations by measuring the overall probability to which the component words in each collocation co-select one another. To rank collocations, we have used three association measures of t-score, MI-score, and logDice score in combination with the raw frequency. Although the t-score and the MI-score have been the most widely used association measure in previous literature (e.g., Hunston, 2002; Manning & Schütze, 1999; Rychlý, 2008), the measures are highly biased toward high- and low-frequency collocations. The MI-score in particular tends to overestimate the "rarity" of the co-selection, and thus overly emphasize low-frequency words and their collocates, which is of little value for EFL learners. In this regard, the t-score and MI-score are to be used only as a baseline in the current study. Alternatively, logDice measure was chosen as a major index to

rank collocability, since the measure does not penalize the high-frequency collocates in estimating the exclusivity of the co-selection of words.

In this study, the threshold level for association measure was set at 5 for logDice score, 4 for MI-score, and 2 for t-score, with the minimum co-occurrence frequency set at 5. We applied a little more stringent criteria than a commonly cited threshold level for MI of 3 (eg., Hunston, 2002) in conjunction with a minimum t-score of 2 (Church & Hanks, 1990) or cut-off frequencies set at 3-5 co-occurrences (Church & Hanks, 1990; Stubbs, 1995). In determining the threshold level for logDice score, we followed Frankenberg-Garcia et al. (2019), who noted that collocations with logDice score below 5 are perceived as a free association rather than collocations. The candidates over these threshold levels will be verified as collocations, while those below the cut-off points are to be categorized as free combinations.

3.2.2. Distributional Variables for Collocation Use

Once collocations are verified based on the aforementioned parameters, the four distributional patterns of collocation used in the target corpus (i.e., collocation density, diversity, repetition and association strength) are quantified, using the formulae which are to be explained in this subsection.

The first distributional variable to be analyzed is “collocation density” which indicates the amount of collocation used in the target corpus. The current study will apply two commonly used formulae following Laufer and Waldman (2011). The density was first operationalized as “collocation density = collocation token counts/word token counts” to indicate how many collocations appear within a

text of the same length. We will also apply the second formula "collocation density = collocation token counts/node word counts" to measure the proportion of collocations in relation to the node word counts. The collocation density calculated based on this measure is interpreted as how often collocations appeared whenever the target words were used in the corpus.

Next, collocation diversity is another distributional feature which indicates the level of a variety of collocation types within a corpus. In order to estimate the level of collocation diversity in a set text length, collocation diversity was operationalized as the mean number of collocate types in relation to the corpus size. To reduce the corpus size effect, we take the square-root of the token counts, and thus the formula "collocation diversity = the type counts/√corpus tokens" was used to calculate the level of collocation diversity given a corpus size. This formula indicates how many collocation types were used in a set number of word counts. In addition, we also wanted to examine the diversity of the association; that is, the degree of variety in collocate types associated with each node word. To gauge how many different collocates co-occurred with each node word, the formula "Association diversity = Collocation types/√node word tokens" was used.

Thirdly, the repetition rate refers to the degree to which individual collocation types recur throughout the corpus. As stated in the previous section, the modified collocation type-to-token ratio (CTTR) is chosen to measure repetition instead of collocation diversity. As stated in Section 2, the collocation type-to-token ratio gives the reverse score to the degree how much a writer repeats individual collocations (Durrant, 2008). If CTTR gives higher marks, it means fewer repetitions made by each collocation type to explain a set number of tokens. One of the

drawbacks of this formula, however, could be that it penalizes the (generally longer) texts with higher collocation token counts. Since the collocation token counts in the reference corpus are much higher than those in the textbook and the learner corpus, some adjustment was needed to minimize the size effect. To reduce the effect of the denominator, the formula "RTTR = Total collocation type counts/the square root of the total collocation token counts" was adopted as suggested by Guiraud (1954) and Paquot (2018).

Finally, the association strength was the last variable to be analyzed. Following Durrant (2008), Siyanova-Chanturia (2015), and Paquot (2019), we examined the distribution of association scores of collocations in each corpus. What distinguishes the current study from the other association studies is the use of logDice score as a primary measure for association strength. The median of logDice score given to each collocation in the corpus was computed with a statistical test of the significance of the difference.

In addition, we examined the correlation between association strength and frequency of co-occurrence. It is assumed that frequency cues which strongly predict association strength are facilitative in developing learners' sensitivity to the associative relationship. Besides, it was expected that encountering stronger association with a higher frequency is beneficial for efficient learning of collocation. If there is a positive correlation, the rank by co-occurrence frequency measured in the target corpus is concordant with the rank of the association measure established in the reference corpus. By contrast, a negative correlation indicates that the level of frequency in the target corpus does not agree with its collocational strength.

Based on the logDice score, the SkELL corpus was queried with a statistical

method to generate the reference collocation lists of the target nodes. Using SketchEngine tools, all collocates of each of the target nouns were retrieved and the logDice scores of each pair were recorded. The current study, in a pedagogical sense, would rather take the moderate view on rarity, since the excessive emphasis on exclusive combinations and large vocabulary knowledge is not the primary concern in the current study. The representativeness of lexical items rather than their exclusiveness is considered desirable for the general language course for EFL middle to high school students.

3.2.3. Statistical Analysis

In testing the statistical significance of difference, non-parametric tests were carried out. Since most frequency-based data is either categorial or does not follow normal distribution, the present study ran nonparametric tests. In the analysis of collocation density, diversity, and repetition rates, Chi-test has been conducted. To compare median association scores in each corpus, the Kruskal-Wallis test which uses ranks of the data, not the data points, for the calculation of median, was run.

To measure the correlation between the co-occurrence frequency and the association strength, Kendall's correlation test was carried out. Kendall's *Tau* is a measure of rank correlation (Lafferty, Lebanon, & Lafferty, 2002; Lapata, 2003), which calculates how much the ranking orders of the target items differ or agree within the comparing groups. The metric initially estimates the difference between the probability that two variables are in the same order versus the probability that they are in different orders. It is then rescaled to range from -1 to 1, representing the degree of similarity between the ranks of the items between two groups.

3.2.4. Software

In searching for collocations, we utilized a web-based corpus analyzing tool, SketchEngine with Application Programming Interface (API) for efficient data retrieval. SketchEngine provides automatic processing, lemmatization and part-of-speech tagging of the corpus which is necessary to specify the grammatical categories and lexical level of words to be searched. Additionally, it measures the association strength between the node word and candidates based on the frequency data. Candidates are then ranked by the computed association measure.

To automize the data retrieval process, a customized program was utilized for API requests. The retrieved data contained the maximum 1,000 collocates for each 1,718 node nouns and the calculated co-occurrence frequency and association measures. This database retrieved from the reference corpus henceforth will be referred to as “the reference collocation database,” while those retrieved from the target textbook and learner writing corpus using the same method will be called “the candidate database.” The dataset from each corpus was merged for the further collocation identification process and written in the format of Excel spreadsheets by Python script. Finally, SPSS software was used to summarize the data and carry out the test of statistical difference.

3.3. Procedures

The following subsections describe the methods and procedures of collocation identification. First, the development of the reference collocation database is explained in Section 3.3.1. Then, details of the collocation identification and the subsequent analysis process are described in Section 3.3.2.

3.3.1. Developing Reference Collocation Database for Korean EFL Learners

As stated in Chapter 2, the top-down approach was taken to profile curriculum-related collocations in the target corpora. In this approach, collocations in the target corpus are identified based on a collocation list which is statistically verified by the reference corpus.

The first stage of collocation identification is therefore to generate a collocation list by retrieving candidates and statistical data, including co-occurrence frequencies and association measures. Among these extracted candidates, the items which meet the statistical criteria based on the association measures are confirmed as collocations. Based on this database, in the search stage, the reference corpus was queried for collocate candidates for each of the 1,718 target node words retrieved. To filter out irrelevant candidates, the search parameters were specified to the word classes (Part-of-Speech), the minimum frequency of co-occurrence, and the distance from the target word (the size of the search window). Requests were made for 1,718 lemmatized nouns with a search parameter specified by [pos “n.*”], the minimum co-occurrence frequency was set at 5 within a span of ±4 for VNCs and ±1 for NNCs and ANCs. After retrieving the maximum 1,000 candidates based on the described

parameter, all other grammatical categories except verbs, nouns and adjectives were removed from the database. The candidates retrieved from the reference corpus are checked against the minimum criteria based on its own association measures (t-score≥2, MI≥4, logDice≥5). Consequently, the reference collocation database is compiled, containing collocation candidates with statistical data on the co-occurrence and association strength (i.e., the raw co-occurrence frequency, logDice, MI, and t-scores). This reference collocation database is used for the subsequent assessment of the target corpus. Among the candidates, the items which meet the statistical criteria based on the association measures are confirmed as collocations.

3.3.2. Collocation Identification and the Analysis of Collocation Use in Corpora

Using the textbook and learner corpus compiled on SketchEngine, we extracted collocation candidates for 1,718 node words through the Application Programming Interface (API) method. As was done with the reference corpus, only N-N, A-N, V-N subtypes were queried within a window of ±4 for VNCs and ±1 for ANCs and NNCs. After extracting all existing combinations from the target corpus with a minimum frequency threshold set at 1, we first checked if the retrieved items were attested in the reference corpus. If the pairs were not found in the reference corpus database, they were considered as irrelevant and thus removed from our analysis. The only items attested in both the target corpus and the reference corpus were selected as a candidate for statistical verification. After retrieving candidates from the target corpus, each candidate was checked against the reference collocation database and assessed based on the statistical criteria described above.

3.4. The Collocation Database

Table 3.4 presents the resulting token counts of candidates and collocations identified in each corpus. It should be explained here that "collocation token" indicates the entire amount of collocations appearing in the corpus. The reference corpus includes the total 14,002,016 collocation candidates (94,666,254 V-N pairs, 19,143,137 N-N pairs, 26,568,942 A-N pairs). The number of statistically verified VNCs is 24,345,386 (25.72% of all candidates), NNCs is 8,498,302 (44.39%) and ANCs is 12,502,402 (47.06%), which yields the reference collocation database total of 45,346,090 (32.30%) collocations.

Table 3.4
Token Counts of the Candidates and Collocations in the Target and the Reference Corpora

		VNC		NNC		ANC	
		Candidate	Collocation	Candidate	Collocation	Candidate	Collocation
Textbook Corpus	Average (S.D)	2,999 (444.61)	812 (140.85)	316 (62.67)	162 (35.54)	721 (136. 29)	352 (74.89)
	Percent (%)	27.00		51.22		48.60	
	Per one nodeword	4.2	1.1	0.4	0.2	1.0	0.5
Learner Corpus	Average (S.D)	29,433 (613.96)	6,299 (171.12)	2,030 (66.91)	814 (25.88)	6,700 (179. 49)	3,656 (108.61)
	Percent (%)	21.4		40.1		54.6	
	Per one nodeword	28.7	6.2	2.0	0.8	6.5	3.6
Reference Corpus	Total	94,666,254	24,345,386	19,143,137	8,498,302	26,568,942	12,502,402
	Percent (%)	25.72		44.39		47.06	
	per one nodeword	81,710.3	26,394.7	11,142.6	4946.6	15465.0	7277.3

In textbook corpus, the average of 2,999 V-N pairs, 316 N-N pairs, and 721 A-N pairs were identified as candidates. Among these 4,037 candidate tokens identified in both textbook and reference corpus, 1,326 pairs were verified as collocations by passing the threshold score. The proportion of verified collocations among the candidates varied on subtypes; the collocations with the highest frequency were VNCs, followed by ANCs and NNCs in frequency order. The average of nearly 812 verb-noun pairs, 352 adjective-noun pairs, and 162 noun-noun pairs in each corpus were verified as collocations.

In the learner corpus, the average of nearly 38,163 candidates (29,432 V-N pairs, 2,030 N-N pairs, and 6,700 A-N pairs) were selected as candidates to be examined for collocability. After being checked based on the statistical criteria, 10,769 pairs (28.2%) were consequently qualified as collocations. By its subtypes, 6,299 VNCs (21.4% of V-N candidates), 814 NNCs (40.1%), 3,656 ANCs (54.6%) were found to meet the statistical criteria.

Table 3.5 illustrates the candidate and collocation type counts in each corpus. Representing the degree of variety of collocations, “a collocation type count” is estimated by a total number of different collocations used in the corpus. The reference corpus exhibits a total of 38,676 type counts for VNC, 22,920 NNC types, and 21,499 ANC types, each of which accounts for 12.3%, 5.1%, and 5.7% of candidate type counts. As for the target corpus, each textbook material contains on average 2,174 V-N candidate types, 248 N-N candidate types, and 584 A-N candidate types, among which were found verifiable 634 VNC types (29%), 122 NNC types (49.1%), and 257 ANC types (43.9%). The learner corpus presents 7,463 different types of V-N candidates, 22.7% of which were verified as collocation

(1,691 types), 217 NNC types (25% of 867 N-N candidate types), and 601 ANC types (27% of 2,230 A-N candidate types).

Table 3.5
Type Counts of Candidates and Collocations in the Taget and the Reference Corpora

		VNC		NNC		ANC	
		Candidate	Collocation	Candidate	Collocation	Candidate	Collocation
Textbook Corpus	Average (S.D)	2,174 (307.21)	634 (110.86)	248 (49.62)	122 (25.83)	584 (109.89)	257 (52.88)
	percent (%)	29.0		49.1		43.9	
	Per one nodeword	3.0	0.9	0.35	0.1	0.8	0.4
Learner Corpus	Average (S.D)	7,463 (76.90)	1,691 (44.75)	867 (19.32)	217 (12.46)	2,230 (53.07)	601 (12.82)
	percent (%)	22.7		25.0		27.0	
	Per one nodeword	7.3	1.6	0.8	0.2	2.2	0.6
Reference Corpus	Total	314,266	38,676	449,197	22,920	378,658	21,499
	percent (%)	12.3		5.1		5.7	
	Per one nodeword	182.9	20.2	261.5	13.3	220.4	12.5

This chapter has illustrated the methodology used in this thesis. First, we explained the corpora used in the current study. Then, we further described the measures and instruments, including statistical criteria for collocation identification, formulae to operationalize collocational distributions, and the statistical tests, and software used for data analysis. The last section described procedures for compiling reference collocation database, statistical identification, and analysis of collocations in corpora. The summary of the report is presented in Figure 3.1.

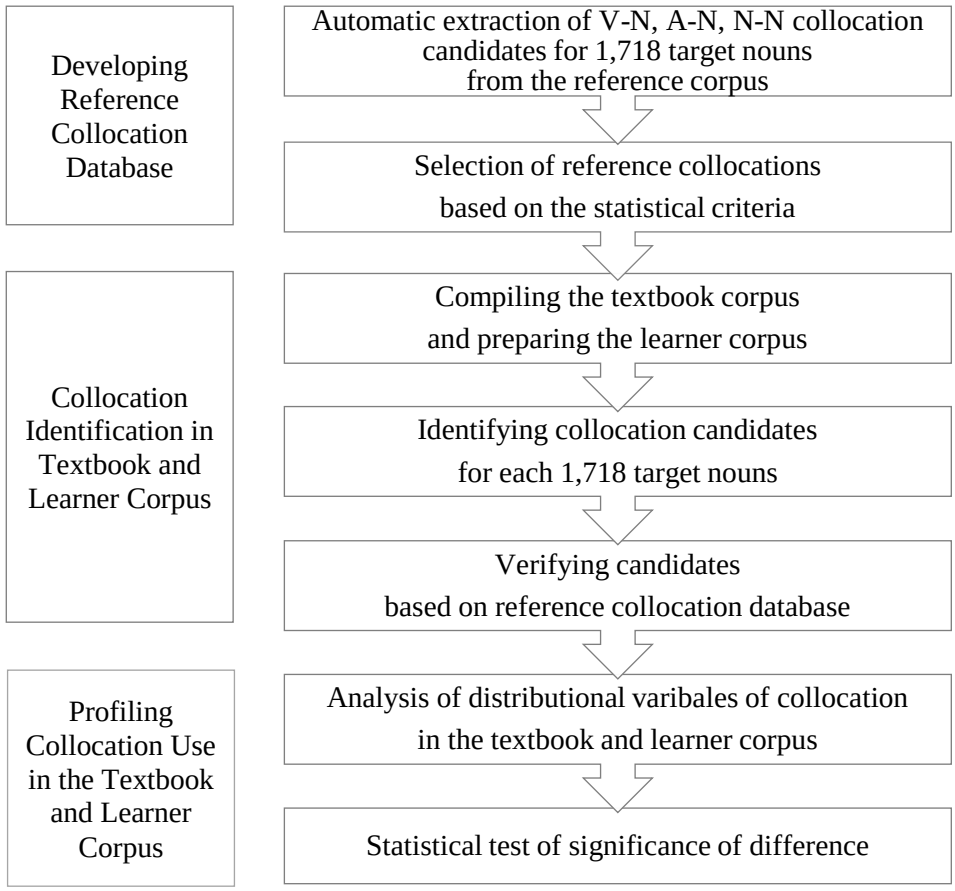

Figure 3.1 Summary of procedures

Chapter 4. Results

This chapter presents the findings. Section 4.1. reports on the findings regarding the use of collocation in Korean middle and high school English textbook materials. Then, section 4.2. discusses the use of collocation in Korean EFL learners' writing. The collocation use described in each section addresses the four distributional variables: collocation density, diversity, repetition rates, and association strength.

4.1. Collocation Use in the Korean Middle and High School English Textbook Corpus

This section reports on how Korean middle and high school English textbook materials represent VNCs, NNCs, and ANCs. Subsections 4.1.1. and 4.1.2. address the first two research questions on collocation density and diversity in the textbook corpus. The research questions on repetition rate and association strength are then discussed in the subsections that follow; 4.1.3. and 4.1.4.

4.1.1. Collocation Density

The first analysis regarding collocation density was aimed at estimating the proportion of statistically verified collocations in relation to: 1) text length (total word count), and 2) number of head nouns appearing in each corpus. First, the top half of Table 4.1 presents the total number of collocation tokens in each textbook corpus. On average, each corpus used nearly 1,326 statistically verified collocation tokens (812 VNCs, 162 NNCs, 352 ANCs) with 713 curriculum-related noun types as its node word.

The lower half of Table 4.1 shows the total number of collocations per 1,000 words. Noticeable findings are: 1) the current textbook materials provide approximately 60 collocation tokens per 1,000 words; and 2) the materials tend to present VNCs and ANCs more frequently than is presented by the native reference corpus.

Table 4.1
Collocation Density in the Textbook and Reference Corpora

	Subtype	Textbook Corpus	Reference Corpus
Collocation Token Counts [1)]	VNC	812.38	24,345,386
	NNC	162.13	8,498,302
	ANC	352.00	12,502,402
	Total	1326.24	45,346,090
Collocation Token Counts per 1,000 Words [2)]	VNC	36.43	23.38
	NNC	7.27	8.16
	ANC	15.79	12.01
	Total	59.49	43.55

[1)] Average collocation tokens per each textbook corpus by publishers
[2)] Collocation tokens/ Word counts *1,000

Inspecting the table data more closely, the estimated collocation token count per 1,000 words was 36.43 for VNCs; 15.79 for ANCs, and 7.27 for NNCs, giving a total collocation token count of 59.49 per 1,000 words which is 16 more collocation tokens than in the reference corpus of 43.55 co-occurrences in every 1,000 words (23.38 VNCs, 12.01 ANCs, and 8.16 NNCs). This result indicates that readers would generally encounter one collocation in every 16~17 word counts while reading the textbook materials, which rates higher than for the reference corpus in the same length (every 22~23 word counts), with the exception of NNCs. While VNCs and

ANCs tend to appear more frequently in the textbook materials than in the reference corpus, NNCs appear to be underrepresented compared to the native norms. Statistical significance in the difference was found in VNCs and ANCs, but not in NNCs (VNCs: $\chi 2$=140.099, df=1, p=***0.000, Phi(Φ)=0.000; NNCs: $\chi 2$=3.133, df=1, p=0.077, Phi(Φ)=0.000; ANCs: $\chi 2$=21.732, df=1, p=***0.000, Phi(Φ)=0.000).

One could suspect that the higher collocation density in the textbooks might be due to the bias in the more extensive coverage of the curriculum wordlist within the materials. To resolve this uncertainty, an additional analysis was undertaken using an alternative measure of the total number of collocation tokens in proportion to the head noun counts. The results of the supplementary analysis are summarized in Table 4.2, reaffirming a higher density of VNCs and ANCs in the textbook corpus than in the reference data; that is, a higher proportion of head nouns in the materials had collocational relationships with neighboring verbs and adjectives, while NNCs showed the opposite trend.

Table 4.2
Collocation Density in the Textbook and Reference Corpora

	Subtype	Textbook Corpus	Reference Corpus
Collocation Density by Head Noun Counts [1)]	VNC	24.51%	19.85%
	NNC	4.89%	6.93%
	ANC	10.62%	10.19%

1) Collocation tokens / Head noun token counts *100 (%)

In more detail; on average, one in four head nouns (24.51%) was found to have a collocational relationship with its neighboring verbs, while an average 15% of head nouns (10.62% for NNCs, 4.89% for ANCs) formed collocational units with

their adjectival and nominal modifiers. The proportion of collocations of all subtypes to node word counts is significantly different between the two corpora (VNCs: $\chi 2$=45.171, df=1, p=***0.000, Phi(Φ)=0.001; NNCs: $\chi 2$=21.359, df=1, p=***0.000, Phi(Φ)=0.000; ANCs: $\chi 2$=0.653, df=1, p=***0.000, Phi(Φ)=0.000).

Notable findings in respect of the first research question, on collocation density in textbook materials from middle to high school levels, are: 1) on average, each textbook corpus contains approximately 60 collocation tokens per 1,000 word count, which corresponds with one collocation in every 16~17 word counts; 2) the average of 25% of head nouns have a collocational relationship with adjacent verbs, and 15% with adjacent nouns or adjectives; 3) significantly more VNC and ANC tokens but fewer NNCs than the native reference corpus provides.

4.1.2. Collocation Diversity

While the collocation density discussed in the previous subsection considered the total number of co-occurrences, collocation diversity examines how many different types of collocation appear in each of the eight textbook corpora. Thus, the analysis of collocation diversity set out to estimate the total number of statistically verified collocation types in relation to: 1) text length (total word count); and 2) the total number of head nouns. The top half of Table 4.3 presents the total number of collocation types in each corpus; an average of nearly 1,012 collocation types (633 VNCs, 122 NNCs, and 257 ANCs), comprising 713 curriculum-related head nouns as node words, were found in each textbook material. The lower half of Table 4.3 summarizes the estimated collocation diversity rates within set text lengths. A particular finding is that VNC and ANC collocation subtypes in the textbook

corpus show significantly higher diversity rates than in the reference corpus.

Table 4.3
Collocation Diversity in the Textbook and Reference Corpora

	Subtype	Textbook Corpus	Reference Corpus
Collocation Type Counts (Raw Frequency)	VNC	633.63	38,676
	NNC	122.00	22,920
	ANC	257.00	21,499
	Total	1,012	83,095
Collocation Diversity Rates To Text Length [1)]	VNC	4.24	1.20
	NNC	0.82	0.71
	ANC	1.72	0.67
	Total	6.78	2.58

[1)] Collocation diversity rates to text length = Collocation types/√Word token counts

Analyzing the detail; the textbook corpus demonstrates an overall diversity rate of 6.78, subdivided by 4.24 for VNCs, 0.82 for NNCs, and 1.72 for ANCs. Although diversity rates for all three subtypes were higher than those for the reference corpus, statistical significance of difference was found only in VNCs and ANCs (VNCs: $\chi 2$=1150.055, df=1, p=***0.000, Phi(Φ)=0.019); NNCs; $\chi 2$=2.395, df=1, p=0.122, Phi(Φ)=0.001; ANCs; $\chi 2$=248.037, df=1, p=***0.000, Phi(Φ)=0.009).

Next, the estimated diversity rates of collocate types associated with each head noun are reported in Table 4.4. As can be seen, the results confirm that the diversity of VNCs and ANCs was significantly higher in the textbook corpus than in the reference corpus. Of even greater interest, the diversity of VNCs and ANCs in the textbook corpus was observed to rate more highly than that of the reference

corpus by three to four times.

Table 4.4
Collocation Diversity in the Textbook and Reference Corpora

	Subtype	Textbook Corpus	Reference Corpus
Collocation Diversity Rates to Head Nouns [1)]	VNC	11.01	3.49
	NNC	2.11	2.07
	ANC	4.46	1.94
	Total	17.58	7.5

[1)] Collocation diversity rates to head nouns = Collocation types/√Head noun token counts

To illustrate this in more detail; on average, 17.58 types of all collocational categories, subdivided by 11.01 types for VNC, 2.11 for NNC, and 4.46 for ANC were found in the set number of node words (√total node word counts). Similar to the previously reported data, statistical significance of difference was found only in VNCs and ANCs (VNCs: $\chi 2$=948.990, df=1, p=*0.000, Phi(Φ)=0.029; NNCs: $\chi 2$=0.069, df=1, p=0.793, Phi(Φ)=0.000; ANCs: $\chi 2$=190.179, df=1, p=*0.000, Phi(Φ)=0.013).

In summary, the textbook corpus was found to present significantly more diversified VNC and ANC types than the reference corpus when text length and number of node words are controlled, and possible reasons for this observation will be discussed in the next chapter. We now turn to report on the degree of collocation repetition.

4.1.3. Repetition Rate

The repetition in the current study indicates the degree to which individual collocation recurs throughout the corpus. Regarding the number of repetitions for

each collocation type in the textbook corpus, Table 4.5 presents the average frequency of individual collocation type in the textbook corpus. Since the frequency of collocation is not normally distributed, median value best summarizes the data.

Table 4.5
Average Number of Repetitions of Collocations in the Textbook Corpus

Subtype	*Mdn*	M	N	SD
VNC	1	1.28	5,069	0.775
NNC	1	1.33	976	1.013
ANC	1	1.37	2,056	1.126

What stands out in the table is that all collocation subtypes appear only once throughout the middle to high school curriculum materials. When comparing each subtype, VNCs with a mean of 1.28 frequencies tend to repeat less than ANCs (M=1.37, SD=1.126), NNCs (M=1.33, SD=1.013).

To look into the detailed distribution, the proportion of collocations according to frequency range (number of repetitions) was analyzed, and the results are summarized in Table 4.6. A notable finding is that 64% of VNC types appeared only once throughout the textbook materials, whereas 0.4% of VNCs are presented more than 10 times. For NNCs and ANCs in general, they tend to repeat slightly more than VNCs, although the proportion of items repeating over 10 times remains limited to 2.51% of NNCs and 2.50% of ANCs. Moreover, nearly 61% of NNCs and 59% of ANCs are never revisited. This result indicates that Korean learners are likely to encounter half of the collocations only once throughout the middle and high school textbook materials.

Table 4.6
Distribution of Collocations according to the Number of Repetitions in the Textbook Corpus

Subtype	1	2 to 5	6 to 10	over 10
VNC	519.88* (64.00%)	269.88 (33.20%)	19.75 (2.40%)	2.88 (0.40%)
NNC	99.38 (61.30%)	54.50 (33.60%)	4.63 (2.90%)	3.63 (2.51%)
ANC	207.25 (58.90%)	112.88 (32.10%)	23.25 (6.60%)	8.63 (2.50%)

* Collocation type counts (%)

Next, Table 4.7 compares the repetition rate of the textbook corpus with the reference baseline. With RTTR, the modified type-token ratio gives a reverse score for the number of repetitions made by each collocation type: a higher RTTR score indicates less repetition of the items within the corpus. This table is revealing, as it shows that collocations of all subtypes used in the textbooks are less recursive than their equivalents in the reference corpus.

Table 4.7
Repetition Rate by RTTR* in the Textbook and Reference Corpora

Subtypes	Textbook Corpus	Reference Corpus
VNC	22.23	7.84
NNC	9.58	7.86
ANC	13.70	6.08

RTTR* = Collocation type/√Collocation token

By looking more closely at the data presented in Table 4.7, the highest RTTR of VNCs (22.23) in the textbooks can be seen to be particularly significant,

since the data suggest that VNCs in the textbook materials occur with least repetition compared to other subtypes in the materials corpus (9.58 for NNCs, 13.70 for ANCs) or any other collocation subtypes in the reference corpus (7.84 for VNCs, 7.86 for NNCs, and 6.08 ANCs). Statistical significance of difference in RTTR scores of collocation between textbook and reference corpora was confirmed (VNCs: $\chi 2$=804.789, df=1, p=*0.000, Phi(Φ)=.040**; NNCs: $\chi 2$=5.168, df=1, p=*0.000, Phi(Φ)=.004**; ANCs: $\chi 2$=188.490, df=1, p=*0.000, Phi(Φ)=.023**).

4.1.4. Association Strength

This subsection examines the probabilistic nature of collocation; that is, association strength as the likelihood that two component words would co-select each other over others. First, Table 4.8 summarizes the overall association strength of collocations in the textbook and reference corpora, measured by the logDice score. A higher logDice score indicates a higher probability that component words would be associated strongly enough to be a collocation in general English use[4]. Data from the table shows counterintuitive results: the association strength of VNCs and ANCs were found to be significantly higher in the textbook corpus than in the reference corpus.

[4] 'Plus 1 point' in logDice scale means twice ($\approx 2^1$) as frequent collocation while plus 7 points means roughly 100 times frequent ($\approx 2^7$) collocations (Rychlý, 2008)

Table 4.8
logDice Score of Collocations (Tokens)
in the Textbook and Reference Corpora

Corpus	VNCs		NNCs		ANCs	
	Mdn	Sig	*Mdn*	Sig	*Mdn*	Sig
Textbook	6.76	*.000	7.09	.166	7.38	*.000
Reference	6.61		6.93		7.21	

To further illustrate the results of the statistical analysis seen in Table 4.8; the median logDice score in the textbook corpus is 6.76 for VNCs, 7.09 for NNCs, and 7.38 for ANCs, while the reference corpus yields a median score of 6.61 for VNCs, 6.93 for NNCs, and 7.21 for ANCs. The independent Mann-Whitney U test confirmed the statistical significance of the difference between the association measure within the two corpora in ANCs (U=16,869,367,455.500, p(two-tailed)=.000), and VNCs (U=73,843,651,609.000, p(two-tailed)=.000), but not in NNCs (U=5,388,721,989.500, p(two-tailed)=.166). Thus, the higher gains in the textbooks demonstrate that collocations in the materials tend to be the pairs associated with higher probability. The reference corpus, on the other hand, seems to favor somewhat weaker associations, indicating native speakers' tendency to use less frequent and less predictable collocations.

Next, in order to provide a detailed comparison between association scores in the two corpora, Table 4.9 illustrates the proportion of collocations in each range of logDice scores. Interestingly, collocations in the textbook corpus are largely restricted to items at the upper-mid level of association strength, falling short of lower-mid collocation strength, which explains a large proportion of collocations in the reference data.

Table 4.9
Distribution of Collocations by Association Strength
in the Textbook and Reference Corpora

Corpus	Subtype	5-7.5	7.5-10	10-12.5	Over 12
Textbook	VNC	4,429* (68.15)	1,930 (29.70)	140 (2.15)	0
	NNC	776 (59.83)	463 (35.70)	58 (4.47)	0
	ANC	1478 (52.49)	1164 (41.34)	174 (6.18)	0
Reference	VNC	17,684,508 (72.64)	6,186,010 (25.41)	474,577 (1.95)	0
	NNC	5,344,978 (62.89)	2,666,606 (31.38)	482,542 (5.68)	4,176 (0.05)
	ANC	6,972,411 (55.77)	4,716,734 (37.73)	759,042 (6.07)	54,168 (0.43)

*Collocation type counts (%)

To elaborate; Table 4.9 shows that the reference corpus presents a larger body of collocation of all subtypes (appx. VNCs 73%, NNCs 63%, ANCs 56%) at the lower-mid level of logDice scores ranging from 5 to 7.5, than the textbook corpus (appx. VNCs 68%, NNCs 60%, ANCs 52%). That is, the textbook corpus presents a larger proportion of collocations at the upper-mid to high level of logDice scores ranging from 7.5 to 12 (appx. VNCs 32%, NNCs 40%, ANCs 47%) compared to the reference corpus (appx. VNCs 27%, NNCs 37%, ANCs 45%).

To provide a detailed description of the different collocational strengths in the two corpora, Figure 4.1 and Figure 4.2 exemplify the collocates which the reference corpus (left) and textbook corpus (from the right) associate with the node

words “*money*” and “*idea*,” respectively. While collocates from the reference data are presented on the left-most side of the horizontal lexis, textbook data is shown on the right-hand side (A~H), with logDice scores being on the vertical axis.

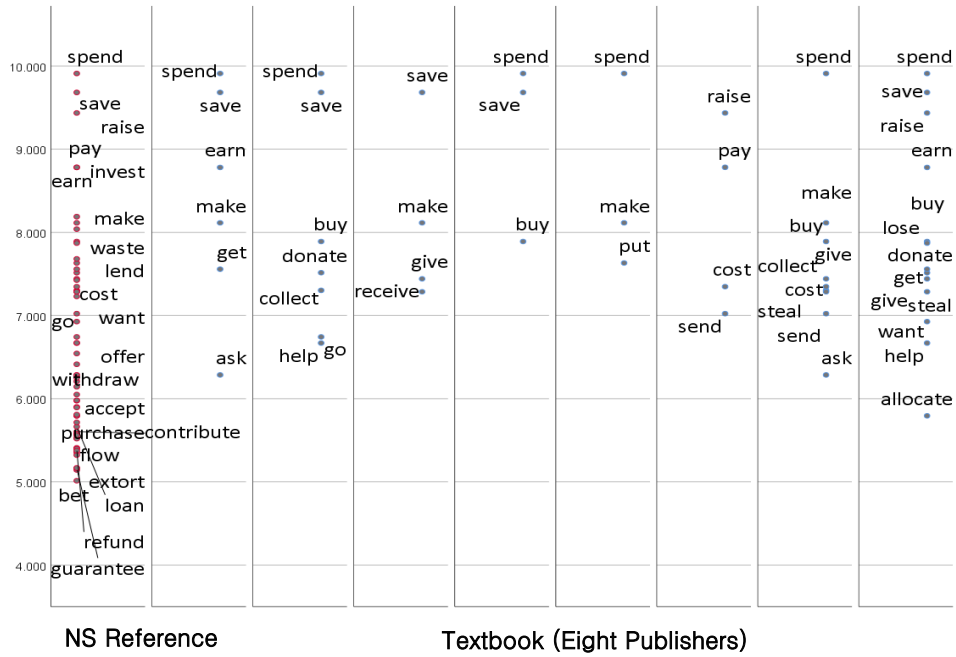

Figure 4.1 VNC collocates for “money” in the textbook corpus

The figures illustrate a narrower range of collocates in the textbook corpus as well as scarcity of items with lower-mid strength (e.g., *offer, loan, obtain* for “*money*”, *generate, challenge, pursue* for “*idea*”). The textbook corpus seems to use more “probable” collocations than the reference corpus, which may be why the textbook corpus gained higher median logDice scores.

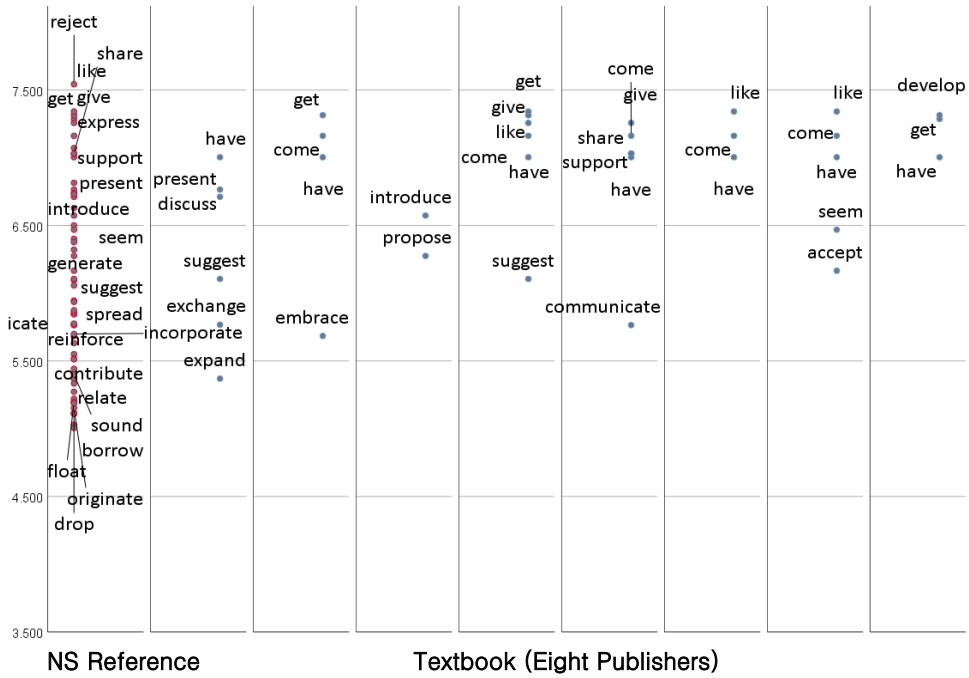

Figure 4 .2 VNC collocates for "*idea*" in the textbook corpus

As shown in the Figure 4.1 and Figure 4.2, the collocates used in the textbook corpus are highly restricted to items ranked at the top of the logDice scale, such as *spend*, *raise*, *save*, *earn* for the node word "*money,"* and *have*, *get*, *come* for the node word "*idea*." This suggests that the range of collocations presented in the textbooks are limited to items with logDice scores over 7; thus, Korean learners may not encounter as many collocations with lower or mid-level association strength, for which the pedagogical value acknowledged by other researchers will be discussed in Chapter 5.

Continuing the analysis of association strength, we further examined how far collocations represented in the textbook materials correspond to their associative strength. The test results of the correlation between the two variables, association

strength and co-occurrence frequency[5], are presented in Table 4.10. Data from the table reveals that the rank order by co-occurrence frequency and association strength of each collocation is less concordant in the textbook corpus than in the reference corpus. In other words, collocation use indicated by the frequency of the target items in the materials is less likely to match with the level of association strength, and vice versa.

Table 4.10
Correlation between Association Strength and Co-occurrence Frequency of Collocations (type) in the Textbook and Reference Corpora

Subtype	Index	Textbook Corpus			Reference Corpus		
		Mdn	Corr.	N	*Mdn*	Corr.	N
VNC	logDice	6.577	.181***	5,069	5.700	.321**	38,676
	Co-occurrence	1			277		
NNC	logDice	6.930	.073*	976	5.852	.316**	22,920
	Co-occurrence	1			153		
ANC	logDice	7.039	.210***	2,056	6.110	.338**	21,499
	Co-occurrence	1			581.53		

***(p<.001), **(p<.005), *(p<.05)

To explain the details with regard to the textbook corpus, the correlation coefficient between the logDice score and co-occurrence frequency of individual VNCs is τ=.181 (p=.000). The correlation of two variants in the textbook is lower

[5] Association strength of collocation is established by the reference corpus; that is, association score indicates what is strong in the native reference data. On the other hand, co-occurrence frequency is calculated within the target corpus and thus indicates how frequently the target collocation appears within either the textbook or Korean EFL learners' writing corpora.

than in the reference corpus, τ=.321 (p=.000), which means that the frequency of collocations in textbook corpus is less likely to reflect the association strength than in the reference corpus. In other words, the extent of collocational input in the textbook materials has little relation to the association strength. For instance, strongly associated collocates of the head-noun *"money"* (eg., *spend, raise, save, pay, make*) tend to occur only once or never appear in most of the textbook materials, of which frequencies are not distinguishable from that of weaker associations (eg., *allocate, ask, cost, steal*). In contrast, in the reference corpus, the latter group occurs almost six times more frequently (median co-occurrence frequency=10,714) than the former does (1,644). The potential problem indicated by this lower correlation between the two variables in the textbook materials will be discussed in the following chapter.

4.2. Collocation Use in the Korean EFL College Students' Writing Corpus

This section reports on the findings regarding the use of collocation represented in the writing samples of Korean EFL first-year college students, in comparison to the reference corpus data. Subsection 4.2.1. presents collocation density based on collocation token counts. Subsection 4.2.2. then addresses collocation diversity with regard to type counts. The ensuing subsections 4.2.3. and 4.2.4. each demonstrate the repetition rate and association strength of collocation used in the target and reference corpora.

4.2.1. Collocation Density

To examine how often Korean EFL learners produce statistically verified collocations in their writings, collocation density was calculated as the proportion of collocation tokens in relation to: 1) text length (total word count); and 2) total head word count.

The top half of Table 4.11 presents the total number of collocation tokens in each textbook corpus; with almost 1,025 curriculum-related head noun types. Each learner corpus contains nearly 29,433 VNC tokens, 2,030 NNCs, and 6,700 ANCs, totaling 38,163 statistically verified collocation token counts for all three subtypes.

Table 4.11
Collocation Density in the Korean EFL learner and Reference Corpora

	Subtype	Learner Corpus	Reference Corpus
Collocation Token Counts [1)]	VNC	29,432.60	24,345,386
	NNC	2,030.40	8,498,302
	ANC	6,699.80	12,502,402
	Total	38,162.80	45,346,090
Collocation Token Counts per 1,000 Words [2)]	VNC	32.07	23.38
	NNC	4.14	8.16
	ANC	18.61	12.01
	Total	54.82	43.55

[1)] Average collocation tokens per each textbook corpus by publishers
[2)] Collocation tokens/ Word counts *1,000

Closer investigation of the data presented in Table 4.11 reveals that estimated collocation density was subdivided by VNCs appearing 32.07 times per 1,000 words, which is the most frequent type among all three sub-categories,

followed by ANCs at 18.61, and NNCs at 4.14 times per 1,000 words. In comparison to the reference corpus learners' use of ANCs, 18.61 is particularly noteworthy, since its frequency is higher than that in the reference data (12.01) and textbook corpus (15.79). By contrast, the NNC token count in the learner corpus (4.14 per 1,000 word count) was lower than in the other two corpora (Textbook corpus=7.27, Reference corpus=8.16), suggesting learners' tendency to underuse NNCs. For all three subtypes, statistical significance of difference between learner corpus and reference corpus was found (VNC: $\chi 2$=648.424, df=1, p=*.000, Phi(Φ)=0.001; NNC: $\chi 2$=29830946.770, df=1, p=*.000, Phi(Φ)=0.169; ANC: $\chi 2$=11356644.481, df=1, p=*.000, Phi(Φ)=0.104).

Next, in order to avoid the error produced by the biased coverage of the curriculum wordlist within the corpus, we alternatively measured density by calculating the number of collocation tokens per head nouns. Table 4.12 summarizes the result, which endorses the previous findings in this subchapter that learners tend to produce significantly more VNCs (23.23%) and ANCs (13.48%), but significantly fewer NNCs (3.00%) with head nouns used in their writings.

Table 4.12
Collocation Density in the Korean EFL Learner and Reference Corpora

	Subtype	Learner Corpus	Reference Corpus
Collocation Density by Head Noun Counts *	VNC	23.23%	19.85%
	NNC	3.00%	6.93%
	ANC	13.48%	10.19%

*Collocation tokens / Head noun token counts *100 (%)

This result is similar to that of Table 4.11, which compared collocation token

counts normalized by text length; where ANCs and VNCs serve higher proportions in the learner corpus than reference corpus, by nearly 3% for each. On the contrary, the underrepresentation of NNCs in the learner corpus was confirmed by the fact that learners presented only 3% of head nouns as a base of collocation, which is 4% lower degree than NNCs in the reference corpus. Statistical significance of difference was found in all three subtypes (VNC: $\chi 2$=852808677.786, df=1, p=*.000, Phi(Φ)=2.637; NNC: $\chi 2$=814442069.495, df=1, p=*.000, Phi(Φ)=2.577; ANC: $\chi 2$=820297574.969, df=1, p=*.000, Phi(Φ)=2.586).

Collocation density in the learner corpus can be summarized as follows: 1) generally, learners use collocations with a higher degree of density than native speakers; 2) collocation frequencies may differ in sub-groups classified by syntactic relations. Learners seem to overuse VNCs and ANCs, but underuse NNCs compared to the reference corpus.

4.2.2. Collocation Diversity

Collocation diversity in Korean EFL learner writings was first analyzed with regard to how many collocation types were used in a set text length. The top half of Table 4.13 summarizes raw collocation type counts of collocations in the learner and reference corpora, while the lower half shows comparable type counts in relation to text length. According to the data, all collocation subtypes except NNCs show higher diversity rates in the learner corpus than in the reference corpus. When normalized by text length, a total of 5.66 collocation types were found in the learner corpus (3.82 for VNCs, 0.49 for NNCs, and 1.36 for ANCs), which is higher than the estimated type counts in the native data. All three subtypes of the two corpora

show statistical significance of difference (VNC: $\chi 2$=1034813239.935, df=1, p=*.000, Phi(Φ)=17.787; NNC: $\chi 2$=1034813154.079, df=1, p=*.000, Phi(Φ)=17.787; ANC: $\chi 2$=1034813068.677, df=1, p=*.000, Phi(Φ)=17.787)

Table 4.13
Collocation Diversity in the Korean EFL Learner and Reference Corpora

	Subtype	Learner Corpus	Reference Corpus
Collocation Type Counts (Raw Frequency)	VNC	7,463	38,676
	NNC	867	22,920
	ANC	2,230	21,499
	Total	10,560	83,095
Collocation Diversity Rates to Text Length [1)]	VNC	3.82	1.20
	NNC	0.49	0.71
	ANC	1.36	0.67
	Total	5.66	2.58

[1)] Collocation diversity rates to text length = Collocation types/√Word token counts

Next, diversity rates in relation to head noun counts are presented in Table 4.14, indicating how many collocate types are associated with a set number of node words. The analysis results reaffirm that diversity rates of VNCs and ANCs in the learner corpus are far higher, and NNCs lower, than their equivalents in the reference corpus when the number of head nouns is controlled.

Analyzing the detail: Table 4.14 shows 15.24 types in total, subdivided by 10.27 types for VNCs, 1.32 for NNCs, and 3.65 for ANCs, found in a set number of head nouns. Statistical significance of difference was found in all three subtypes (VNC: $\chi 2$=1039090337.044, df=1, p=*.000, Phi(Φ)=30.406; NNC: $\chi 2$=1039089760.082, df=1, p=*.000, Phi(Φ)=30.406; ANC: $\chi 2$=1039089780.023,

df=1, p=*.000, Phi(Φ)=30.406).

Table 4.14
Collocation Diversity in the Korean EFL Learner and Reference Corpora

	Subtype	Learner Corpus	Reference Corpus
Collocation Diversity Rates to Head Nouns [1)]	VNC	10.27	3.49
	NNC	1.32	2.07
	ANC	3.65	1.94
	Total	15.24	7.5

[1)] Collocation diversity rates to head nouns = Collocation types/√Head noun token counts

Overall, results show that learner writings present more collocation types than the reference corpus, except for NNCs, within normalized text lengths or node word counts. Discussion on the reasons for these counterintuitive results will be covered in Chapter 5. The following section reports on the degree of repetition in learners' collocation use.

4.2.3. Repetition Rate

Collocation repetition rate in the learner writing corpus operationalized by RTTR is summarized in Table 4.15. Given that lower RTTR scores indicate a higher degree of repetition by individual collocation type, the higher RTTR of VNCs and ANCs estimated in the learner corpus scores compared to the reference corpus show that learners tend to use each VNC and ANC type with less repetition than the native baseline.

Table 4.15
Repetition rate by RTTR*
in the Korean EFL Learner and Reference Corpora

Subtype	Learner Corpus	Reference Corpus
VNC	21.31	7.84
NNC	7.59	7.86
ANC	9.95	6.08

RTTR* = Collocation type/√Collocation token

Notably, NNCs show lower RTTR in the learner corpus (7.59) than in the native corpus (7.86), as seen in Table 4.15, suggesting that each NNC type in learner writings appeared more repetitively than native norms. The repetition rates of all three subtypes in the two corpora show statistically significant difference (VNC: $\chi 2$=1040333619.372, df=1, p=*.000, Phi(Φ)=45.553; NNC: $\chi 2$=1040746427.345, df=1, p=*.000, Phi(Φ)=59.460; ANC: $\chi 2$=104061637.2197, df=1, p=*.000, Phi(Φ)=53.792).

4.2.4. Association Strength

First, to demonstrate the overall association strength of collocations in the learner corpus and reference corpus, the estimated median logDice scores in each corpus are summarized in Table 4.16. The analysis uncovered a revealing fact, that learners tend to prefer more strongly associated and highly probable collocations of all three subtypes than native speakers.

Table 4.16
logDice Score of Collocations (Tokens)
in the Korean EFL Learner and Reference Corpora

Corpus	VNC		NNC		ANC	
	Mdn	Sig	*Mdn*	Sig	*Mdn*	Sig
Learner	7.11	*.000	8.05	*.000	7.53	*.000
Reference	6.61		6.93		7.21	

Closer inspection of the result (Table 4.16) shows that the median logDice in the learner corpus is 7.11 for VNCs, 8.05 for NNCs, and 7.53 for ANCs. The association scores for all three types of collocation are higher than those in the reference corpus (VNCs 6.61, NNCs 6.93, ANCs 7.21). Statistical difference was confirmed between logDice scores of all three collocation types in the learner and reference corpora using independent Mann-Whitney U test (ANC: U=101,265,219,131.5, p(two-tailed)=.000; VNC: U=331,239,147,103, p(two-tailed)=.000; NNC: U=11,687,073,088, p(two-tailed)=.000). Learners' choice of highly probable collocations and their pedagogic implications will be discussed further in the next section.

To provide in-depth analysis of Korean EFL learners' use of stronger, and thus more probable collocation, and deficiency in relatively weaker, less predictable collocation, Table 4.17 summarizes the distribution of collocations in each corpus according to a band of logDice scores.

Table 4.17
Distribution of Collocations by Association Strength
in the Korean EFL Learner and Reference Corpora

Corpus	Subtypes	5-7.5	7.5-10	10-12.5	Over 12.5
Learner	VNC	18,422* (58.49)	12,521 (39.76)	552 (1.75)	0
	NNC	1,377 (33.85)	1,801 (44.27)	890 (21.88)	0
	ANC	9,091 (49.73)	7,575 (41.44)	1,614 (8.83)	0
Reference	VNC	17,684,508 (72.64)	6,186,010 (25.41)	474,577 (1.95)	0
	NNC	5,344,978 (62.89)	2,666,606 (31.38)	482,542 (5.68)	4,176 (0.05)
	ANC	6,972,411 (55.77)	4,716,734 (37.73)	759,042 (6.07)	54,168 (0.43)

*Collocation type counts (%)

It can be clearly seen from the data that collocations at lower-mid level strength in the range of 5 to 7.5 logDice scores were less productive in learner writings (appx. VNCs 58%, NNCs 34%, ANCs 50%) than in the native baseline (appx. VNCs 73%, NNCs 63%, ANCs 56%). Conversely, a substantial number of collocations at the upper-mid to high level of logDice scores, ranging from 7.5 to 12.5, appeared in the learner corpus (appx. VNCs 42%, NNCs 66%, ANCs 50%). Another notable trend found in the learner corpus, similar to collocation use in the textbook materials, is that of the highest level of logDice score, while learners used a larger body of NNCs at logDice score 10-12.5.

Learners' tendency to rely on highly probable collocations is portrayed in Figure 4.3 and Figure 4.4. These plots present collocates of the node word *idea* and

money and its association strength expressed by logDice scores. The left-most column observes natives' choice of collocates, densely populated at the lower range of score bands (logDice 5-6), whereas the five columns from the right (A~E) show collocates in learners' writings piled up above logDice 5.5~6.

For example, as Figure 4.3 shows, Korean learners seem to prefer the collocates with logDice score over 5.5 (eg., *get*, *share*, *support*, *propose*) for the headword "*idea",* while collocates below 5.5 logDice score (eg., *reinforce*, *combine*, *reflect*, *borrow*, *drop*, *entertain*) rarely appeared in learner writings.

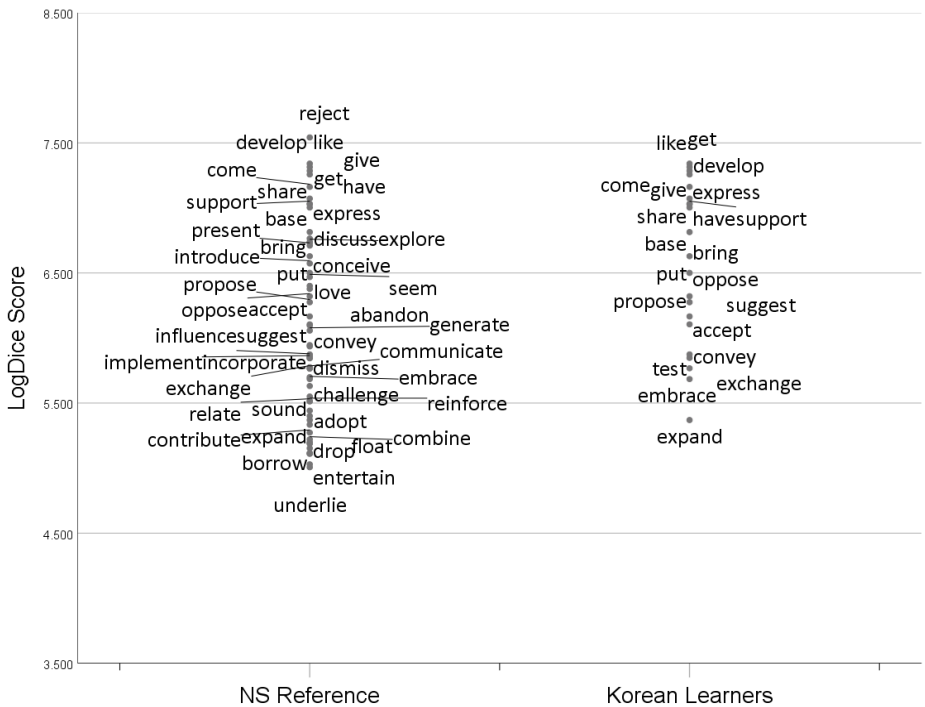

Figure 4.3 VNC Collocates for "*idea*" in the Korean Learener Corpus

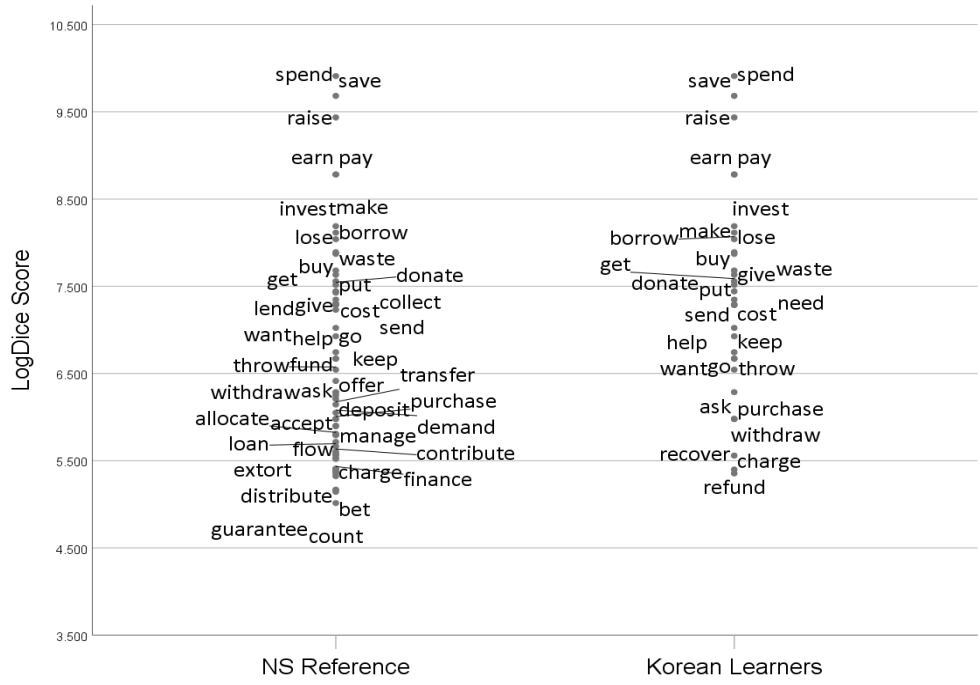

Figure 4.4 VNC Collocates for "*money*" in the Korean Learener Corpus

Moreover, Figure 4.4 shows that collocates with higher logDice scores tend to have a high degree of commonality between the learners and native speakers; for example, collocates for "*money"* such as "*spend, save, pay,* and *earn"* which topped the logDice score, appeared in both corpora. By contrast, items with logDice scores below 6 (eg., *offer, demand, accept, refuse, count*) were scarce in the learner text even if they are included in the curriculum wordlist.

Lastly, to demonstrate the extent to which learners' use of collocation represented by frequency corresponds with association strength established by native English, the results of Kendall's correlation test between logDice scores and co-occurrence frequency is presented in Table 4.18.

Table 4.18

Correlation between Association Strength and Co-occurrence Frequency in the Korean EFL Learner and Reference Corpora

Subtype	Index	Learner Corpus			Reference Corpus		
		Mdn	Corr.	N	*Mdn*	Corr.	N
VNC	logDice	6.371	.200***	8,457	5.700	.321***	38,676
	Co-occurrence	1			277		
NNC	logDice	6.921	.173***	1,083	5.852	.316***	22,920
	Co-occurrence	1			153		
ANC	logDice	6.773	.217***	3,007	6.110	.338***	21,499
	Co-occurrence	1			581.53		

***(p<.001), **(p<.005), *(p<.05)

When compared to the reference corpus, the correlation between the two variables in learners' writing was weaker, because learners' frequency of use corresponds less with the association strength of the collocation compared to the frequency of use in native texts.

To analyze the detail; the correlation between association strength and co-occurrence frequency was highest in ANCs (τ=.217, p=.000), followed by VNCs (τ= 0.2, p=.000), and NNCs (τ=.173, p=.000). According to Cohen's standard, correlation coefficients ranging from .173 to .217 are interpreted as weak. Thus, all three collocation subtypes have shown a positive but weak correlation with the level of association strength.

The data suggests that the amount of collocational input tends to moderately correspond with the association strength in native English, but this is not the case for the Korean EFL learners' production. A relatively weaker correlation between the co-occurrence frequency and the association strength in the learner corpus indicates

that there might be some mismatch between learners' preference and the actual collocability of the item.

To summarize this chapter, notable findings in respect of the first research question, on collocation use in the textbook materials from middle to high school levels, are the use of significantly more VNC and ANC tokens but fewer NNCs than the native reference corpus. Readers of textbook materials would encounter approximately 60 collocation tokens per 1,000 word count, which corresponds with one collocation in every 16~17 word counts. Moreover, the textbook corpus is found to present significantly more diversified VNC and ANC types than the reference baseline. The materials, however, presents a lesser degree of repetition for individual collocation of all three subtypes. Lastly, when overall association strength is estimated using logDice score as a measure, the textbook corpus shows surprisingly higher association scores, indicating that collocations used in the textbook corpus tend to be more 'likable' association than those used in the reference corpus. The correlation test result shows, however, that input frequency in the textbook corpus does not match as closely with its association strength as that in the reference corpus. The result indicates that the frequency of each collocation in the materials may not provide reliable cues for its association strength. Additionally, the distributions of collocation in the textbook materials and the native English seem to differ in that collocations in the materials are less likely to be repeated with due frequency corresponding to its own association strength.

Collocation density in the learner corpus can be summarized as overuse VNCs and ANCs, but underuse of NNCs compared to the reference corpus. The result also shows that learner writings tend to present more diversified collocation

types than the reference corpus, while the opposite trend is found in NNCs. On the other hand, the average association level in the learner corpus is higher than that in the reference corpus. This suggests that learners tend to use more typical and highly probable collocations, as opposed to the reference corpus which presents a much broader range of association strength and contains a higher proportion of collocations at the mid-low level of association strength. The correlation between association strength and co-occurrence frequency was weaker in the learner writings than the reference corpus, indicating that learners' use of collocation is less sensitive to its association strength.

Chapter 5. Discussion

The current study explored the use of collocation in the textbook materials and Korean EFL learners' writing. Section 5.1. addresses the extensive use of collocations but lack of its intensity in the textbook materials. Focusing on the core element of collocation in the frequency-based tradition, Section 5.2. discusses the representation of association strength in the textbook corpus, with regard to its correlation with co-occurrence frequency. Turning to Korean EFL learners' data, Section 5.3. explores possible reasons for the learners' extensive collocation use. Lastly, Section 5.4. looks into the learners' restricted collocational repertoire indicated by association strength as well as the underused collocation subtypes.

5.1. The Extensive Use of Collocations and Compromised Formulaicity in the Textbook Materials

Collocation density is one of the most widely investigated properties of collocation use as it indicates the total amount of co-occurrences in the text. Many researchers have suggested that a wide exposure to collocational input may foster the implicit learning of collocation (Durrant & Schmitt, 2010; Parkinson, 2015). In the current study, the textbook corpus shows a higher density of VNCs and ANCs than the reference corpus. A similar pattern was reported by Koya (2004) who found that VNC token and type counts did not differ between English textbooks used in Japan and history textbooks used in the UK. The author attributed the unexpectedly lower degree of collocation density in the history textbook for native learners to the use of language appropriate to the subject area, as opposed to the English textbooks which were purposefully designed to provide L2 input. In a similar vein, Tsai (2015)

reported that the English textbook series used in Taiwan also exhibited collocational profiles comparable to that of NS productions (appx. 21~27 VNCs per 1,000 words in the textbook corpus, 19 VNCs per 1,000 words in the native corpus), concluding that collocation density may be a less prominent issue compared to other more problematic features such as limited collocation repertoire or repetition rates.

Similar to the findings that show a higher level of density of VNCs and ANCs in the textbook materials, the analysis of the collocation diversity rate shows that in comparison to the reference baseline, within a given length of text, textbook materials present a larger number of collocation types of all subcategories. In other words, when text length is taken into account, learners are likely to encounter a considerable number of collocation types throughout textbook materials. In accordance with the present findings, previous studies have also demonstrated higher collocation diversity in ELT materials (eg., Koya, 2004; Tsai, 2015). Koya (2004) found that the English textbooks used in Japan contained slightly more collocation types than native history textbooks. Likewise, Tsai (2015) reported a lesser degree of collocation diversity in the native corpus data, where NS writing produced 4.2 type counts compared to 5.4-6.2 found in the textbook corpus (normalized by the formula used in the current study).

Although textbook materials may have some advantages over those in natural English with its superior coverage of the overall token and type counts of VNCs and ANCs, it still remains inconclusive whether textbooks outdo the natural setting as input. The higher collocation density and diversity, in fact, may compromise other aspects which are also central to the definition and learning of collocation. While collocation density and diversity is related to the extent of use of

collocation as a whole set of category, the repetition expresses a formulaic nature of collocation as "habitual" co-occurrence of words. Such formulaicity of collocation and the idiom principle at work in native language is represented by the current findings of the higher level of repetition of VNCs and ANCs in the reference data. Together with a lower density and diversity, the highly repetitive co-occurrence of collocational category as a whole in the reference corpus seems to demonstrate its idiomatic tendency that a highly selective set of formulaic language recurs to a greater extent than other random combinations.

In the pedagogic context, many researchers acknowledge the significant role repetition plays in language learning, especially in the usage-based tradition, as the co-occurrence frequency predicts the level of entrenchment (Bybee, 2006; Ellis, 2002; Ellis, 2001; Ferrand & New, 2003; Wolter & Gyllstad, 2013). According to Ellis (2001), efficient language learning is promoted through the repeated exposure to input, which fosters a stronger entrenchment of typical language use. The frequent co-occurrence of two words in linguistic input will lead to a stronger association in long-term memory, which is consolidated into 'chunks.' Similarly, emphasizing the significance of repeated exposure in learning collocations, Webb, Newton, and Chang (2013) suggest that 10 to 15 encounters with the target items through reading are necessary for learning collocations. In Durrant and Schmitt's study (2010), at least 8 to 10 exposures were required to gain the initial receptive knowledge of collocations.

Compared to these findings, the amount of repetition in the textbooks examined in the current study seems insufficient and offsets any benefits gained from the higher diversity and density of collocations. The present data show that textbook

materials underrepresent the recurrent nature of collocations by introducing VNCs and ANCs at a lower level of repetition than the reference corpus. Furthermore, the medium frequency of individual collocations of all three subtypes was 1, while the proportion of collocations with a frequency of 1 reached up to 64% of VNCs, 61% of NNCs, and 59% of ANCs, with nearly 90~95% of collocations being repeated less than five times. Despite extensive encounters with a large number of collocation tokens and types, it is unlikely that learners will notice the formulaicity of the items or consolidate a memory trace of the word association after only a single exposure to the textbook materials.

Meanwhile, the infrequent repetition in the textbook corpus is consistent with the findings of Koya (2004) and Tsai (2015) who observe that over half of the collocations that appear in the textbook series never recur and call for the provision of a sufficient amount of repetition for learners. Even though at least six or seven instances of repetition are necessary for vocabulary learning (Peters, 2014; Webb et al., 2013), only a few collocations recur at such rates in the textbook corpus.

In summary, the current study holds that textbook materials offer "extensive" though "less intensive" collocational input. In other words, while the materials seem to introduce a wide variety of collocation, this may, in fact, compromise the repetition of individual collocation types and the formulaic nature of authentic collocation use. In this regard, the current study upholds the tenet proposed by researchers such as Groom (2009) and Koprowski (2005) that "more collocations do not mean the better", and suggests that teachers using collocations need to be aware of this lack of repetition in the textbooks in order to prepare supplemental activities or materials.

5.2. A Mismatch between the Collocation Use and Association Strength in the Textbook Materials

The importance of the frequency of co-occurrence of lexical items and its distributional patterns in language input has been recognized in the literature (e.g. Durrant, 2014; Durrant & Schmitt, 2010; Ellis, 2003; Lorenz, 1999; Northbrook & Conklin, 2018). It has been shown that the co-occurrence frequency in the input is a major facilitative force in the acquisition of formulas (Ellis, 2003) and that learners tend to pick up "attestedly viable, recurrent collocations" (Lorenz, 1999, p. 181) more effectively than less frequent ones (Durrant & Schmitt, 2009). In particular, Northbrook and Conklin (2018) demonstrate the effect on learners' processing of target lexical units of co-occurrence frequency in textbook materials. It is also suggested that language input that is insufficiently representative of the distributional patterns of authentic native language can consequently lead to the entrenchment of non-native-like language in memory.

Supporting these views, the current study shows that collocation use in the textbook materials may not represent the typical collocational distribution in the native language. If we take frequency of co-occurrence as a measure of language experience and association strength as a probabilistic indicator of "true collocations" (Evert, 2009, p. 5), a higher correlation between the two variables found in the reference corpus may indicate that the amount of exposure tends to match the level of collocability to a greater extent. Collocational input in natural English, which reliably predicts association strength, seems to enable native speakers to develop their intuition of the associative relationship. By contrast, a distinctively weaker correlation between the frequency data and association strength in the textbook

corpus suggests that Korean EFL learners are less likely to have such language experience, and collocational distributions presented in the materials are less indicative of collocability than that of their native language equivalent. The current data, therefore, raise a pedagogic issue regarding the insufficient representation of association strength in textbooks, corroborating the findings of Northbrook and Conklin (2018) who found that the lexical units in textbooks, although abundant, are qualitatively different to and followed very different distributional patterns to the reference corpus.

Furthermore, the lower correlation in the textbook corpus suggests that Korean learners would benefit less from input frequency in order to consolidate their knowledge of stronger collocations. As higher input frequency is known to create a deeper entrenchment in the memory of learners, it would be advisable to align the frequency of target collocations with their association strengths for more efficient teaching and learning of collocations. Given that textbook materials constitute the primary learning experience of EFL students, that learners are sensitive to frequency information (Northbrook & Conklin, 2018), and that any shortfall in non-natives' knowledge of collocational associations between words is due to inadequate input (Durrant & Schmitt, 2010), it follows that insufficient or mismatching frequency cues of collocation strength level need to be addressed in order to foster the development of native-like sensitivity to the collocational relationship.

5.3. The Extensive Use of Collocations in the Korean EFL Learner's Writings

Groom's (2009) intermediate and advanced learner data show more lexical bundles than their native counterparts, while a higher degree of lexical variations in the native data seemingly lead to less coverage of the target constructions. This suggests that collocational development may be indicated by "a downward adjustment to native-like use" (p. 33) rather than by the increasing number of collocation use. A similar trend was found in the present study regarding the extensive use of collocations in Korean EFL learners' writing.

First, the learner data show a high density of VNCs and ANCs. The overuse of ANCs by learners is particularly noteworthy; the ratio of collocation to combination reached nearly 53%, indicating that more than half of A-N combinations attested in both corpora (18,280 statistically verified collocation tokens among 33,491 combinations) fall into the category of statistically verified collocations. This shows that learners rely more heavily on collocational pairs than on free associations, exceeding the native baseline where collocations account for 48% of all candidates (12,502,402 statistically verified collocation tokens among 26,548,752 combinations). This result possibly contradicts the earlier finding that L2 learners tend to rely more on individual words rather than to process lexical units as a single choice (Erman & Warren, 2000; Sinclair, 1991; Wray, 2005). Maintaining that learners rely more on the idiom principle than the open principle, Durrant and Schmitt (2010) report that learners exhibited collocational capacity and were able to recall collocations through implicit learning without instructions. Likewise, learners' extensive use of ANCs corroborates the results presented by Siyanova and Schmitt

(2008) who, with reference to the BNC corpus, found that almost half (45%) of adjective-noun combinations produced by learners met the criteria, which matched that of native production (48%).

As with the higher density of ANC use, learners have also been shown to overuse VNCs, which similarly corresponds with previous findings. Tsai (2015), for example, reports that learners used 60% more VNCs than their native counterparts. The author suggests that learners may be capable of using, or have a need to use, more collocations in writing than native speakers or textbooks. Furthermore, Chang (2018) found that VN and AN-types account for the largest sum of all collocations used by Korean learners and native speakers, with no significant difference between the two groups in the overall collocation frequency, meaning Korean learners produce as much formulaic language as their native counterparts. Overall, these findings confirm the current data, which seemingly evidence both the collocational capacity of L2 learners and the idiom principle-based L2 system.

It may be even more surprising to find a higher diversity rate in learner production than in the reference native corpus as it is commonly assumed that the ability to use diverse collocations characterizes the final stages of L2 development (eg., Durrant & Schmitt, 2009; Men, 2018). The current data rebuke this assumption by demonstrating that a higher diversity of collocation use does not necessarily indicate a native-like distribution. One may find the lower diversity in the reference corpus rather counterintuitive, but there are several possible explanations for this result. First, the observation could relate to a restrictive nature of collocational categories and its role in the language acquisition; that is, compared to the infinite potential L1 lexical combinations, only a small proportion (a total of 83,095

collocation types identified from the 1 billion-word reference corpus) can achieve sufficiently high idiomaticity to constitute a formulaic category. In other words, in the native lexicon collocational category seems to form a highly restricted and exclusive membership. This supports the idea that language is in fact close-class (Pawley & Syder, 1983), and that native speakers tend to rely on a smaller range of idiomatic expressions to achieve "economy of effort" (Ellis, 2001, p. 45). According to the studies of first language acquisition, at the earlier stage native children tend to rely on a small set of formulas, and then gradually develop the analytic ability to break up the formulas into individual words and to produce creative syntactic constructions. Consequently, the child's reliance on formulaic language declines with the emergence of productive grammar (Clark, 2009). While fully mature adult grammar also exhibits formulaicity, it is much more variant and independent of particular lexical items than the highly stereotypical child language. In this regard, the extensive use of collocation by Korean EFL learners seems to demonstrate the characteristic formulaicity characteristic of the earlier stage of language development as a "consequence of a lack of analytical ability" (Perkins, 1999, p. 62), while the lower coverage of collocation in the reference data reflects the greater flexibility of native language use, consistent with Groom's account of the decrease of lexical units in the native corpus.

Another plausible interpretation could be that collocation diversity is a less reliable indicator of phraseological complexity or linguistic proficiency. This hypothesis seems to be in line with Paquot (2018) who examined whether phraseological complexity measures (e.g., RTTR, the MI-score) can delimit learners' proficiency and found that there was no systemic increase in the level of

collocational diversity from one proficiency level to the next. He proposed that collocational diversity is possibly of less discriminatory power because the variable is quantified by somehow crude metrics and that association scores might be a better discriminator of L2 performance at upper intermediate to very advanced level.

Data from other studies also support the idea that learners with lower proficiency can produce a relatively larger number of collocation types (e.g., Siyanova & Schmitt, 2008; Parkinson, 2015; Chen, 2017; Paquot, 2019). When we calculate the normalized type counts (collocation type counts per $\sqrt{}$ corpus token counts) based on the reported data from Chen (2017), the estimated diversity rate also disputes the simplistic view that would attribute lower diversity to learners' lower proficiency. In the writings of Chinese tertiary students in their 1st to 3rd academic years, the diversity rate did not show a gradual increase by learner proficiency; RTTR of VNCs was found to be 4.77(1st) < 5.26(3rd) < 5.71(2nd), ANCs 5.08(1st) <5.09(2nd) <6.14(3rd), NNCs 1.95(3rd) < 2.37(2nd) <2.55(1st). If we assume that third-year learners would be exposed to more collocations and thus have higher proficiency than the other groups, the lower diversity rate found in the third-grade students still does not comply with the general assumption. The re-analyzed data from Siyanova and Schmitt (2008) also reveal that learner writing (normalized type counts = 2.31) uses as diverse collocation as native speakers (2.45), without a statistically significant difference (x^2=0.46, df=1, p=0.498, Cramer's V=0.0032), which supports the claim that collocation diversity and the native-like property of collocation are independent to each other.

5.4. The Restricted Collocational Repertoire in the Korean EFL Learner's Writings

While Korean learners tend to use VNCs and ANCs extensively, the range of association strength of the collocation they produce seems to be restricted somehow. This result corroborates Paquot (2019) who also demonstrates that association measure tends to provide the most useful summary of learners' ability to select word combinations.

The analysis of mean logDice scores reveals that collocations that appear in the learner corpus are most strongly associated compared to the textbook corpus and the reference corpus, which has the least strongly associated. That is, Korean EFL learners show a stronger preference for highly probable collocations, while they rarely present weaker collocations with lower logDice scores. The reference corpus, on the other hand, shows the lowest median logDice scores, with larger bodies of collocation being of low-to-mid level association strength.

The higher association strength found in the learner corpus suggests that non-native texts tend to prefer "more likely" collocations and lack "less predictable" combinations with a medium-degree of association strength. For instance, for the head noun *idea,* Korean learners preferred collocates with a logDice score over 5.5 – such as *get, share, support, propose* – while collocates below a 5.5 logDice score – such as *reinforce, combine, reflect, borrow, drop, entertain* – never appeared in their writings. For the head noun *money,* the collocates with a logDice score over 5.5 – such as *spend, need, keep, borrow* – were commonly used by learners, while collocates with lower association scores – such as *loan, distribute, purchase, manage, loan* – rarely occurred in their writings. This may suggest that learners' collocational

repertoire tends to be restricted to highly likely, strongly associated pairs which are more noticeable or prominent in language input, while native speakers' knowledge of collocation covers a wider range of association strength, including combinations with loose connectivity. This result seems to support Howarth (1998) who maintained that the greatest challenge of collocation learning is in "differentiating between combinations that are free and those somehow limited in substitutability" (p. 42). Hence, L2 learners tend to have difficulty distinguishing the level of association strength, especially when identifying the middle ground of restriction.

Korean learners' difficulty in finding a "happy medium" between the two extremes is also supported by Nesselhauf (2003) who investigated the verb-noun collocation use of German learners and found that "collocations with a low degree of restriction [collocations which occur with a number of other nouns but with some degree of restriction in choice] are the most difficult kind of combinations for the learners" (p. 234). This is supported by Schmid (2003) who held that the concept of collocation is hard to pin down, suggesting that because of the ambiguity of having "a medium degree of observable combined recurrence, mutual expectancy, and idiomaticity" (p. 255), collocations with a moderate level of association strength, ranging from a logDice score of 5 to 6, could be less pronounced in language input and thus underrepresented in learner writings.

The learners' shortage of knowledge in this central area between the two ends of free association and idioms may be partly explained by the fact that these items are often underrepresented in the textbook materials. As Howarth (1998) points out, the true significance of collocations is often disregarded as the "unrelated residue of arbitrary co-occurrences and familiar phrases" (p. 42), and so most

scholarly and pedagogical attention has been directed towards the extreme ends of the spectrum; from the application of generative rules to lexis in free combinations on the one hand, to complete frozen idioms on the other. Durrant and Schmitt (2010) also maintain that mid-strength collocations are less likely to be consciously taught than those more idiomatic, highly salient units.

Corroborating these points, the current study shows that the overall association strength of collocations in the textbook materials is higher than that in the reference corpus. While those collocations introduced in current textbooks are restricted to strong, and thus somehow too predictable associations, collocations ranging from low- to mid-level association strength have been somehow underrated despite their saliency in native English. In other words, the selected items in the current textbooks are limited to overly typical and likely combinations, which might be of less pedagogical value since students might learn such items without explicit instruction. This seems to be consistent with other research which questioned the usefulness of the lexical items in the teaching materials (Koprowski, 2005; Kim, 2004). Closer investigation might be needed, therefore, to determine if collocations in the textbook materials are to be worth the class time assigned for explicit learning (Durrant & Schmitt, 2010).

Another important finding related to the restricted collocational repertoire in learners' production is their underuse of NNCs, which is the inverse to their productivity with VNCs and ANCs. The exceptionally restricted use of NNCs by Korean EFL writers seemingly fits with the developmental progression index of noun phrase complexity hypothesized by Biber et al. (2011), who propose that noun modification, a feature of L2 development, progresses from the simple modification

of the noun with an adjective towards complex modification that uses other nouns and postmodifying structures, such as prepositional phrases, relative clauses, and complement clauses. This may explain why Korean learners tend to favour ANCs over NNCs, in that it suggests they follow the theorized developmental sequence of acquiring a simpler structure earlier than they acquire noun premodifiers. Other studies have similarly observed such underdevelopment of NNCs in learner writings and their different developmental patterns, depending on subtype. According to Chen (2017), who compared the use of three different collocation types in the writings of Chinese tertiary learners, no progress was found in NNCs through the academic years, while learners came to use more VNCs and ANCs at the advanced stage. Likewise, Parkinson and Musgrave (2014) report that learners' choice of modifiers differed according to their proficiency levels. While less proficient learners preferred adjectives as noun modifiers, higher proficiency learners frequently used other modifiers, including nouns. Taken together, the current study provides support for the hypothesis that NNCs develop at a later stage of language learning, causing greater difficulties for EFL learners than other types of collocation.

The reason for the belated development of NNCs in learner production is not clear, but it could be attributed to the lack of sufficient input. As shown in the current data, NNCs are relatively underrepresented in the textbook corpus, for which we cannot rule out the possibility that there might be a link between the insufficient input provided by textbook materials and learners' underuse of NNCs in their writing. This relationship was explored by Northbrook and Conklin (2018) who found that Japanese secondary school students were sensitive to the frequency of lexical bundles occurring in their textbooks, exhibiting the clear advantage of input

frequency on the deeper entrenchment in memory and faster processing of higher-frequency items. In a similar vein, Parkinson (2015) compared the use of NNCs in learner groups from different English language learning environments and pointed out that sufficient exposure to natural English in the SLA environment might be one of the conditions that positively affects the amount of collocation use. It can be inferred from this result that Korean EFL learners may have difficulty learning NNCs or avoid using unfamiliar subtypes due to the deficient language input from their textbook materials or EFL environment.

Chapter 6. Conclusions

The final chapter begins by summarizing the major findings of the present study (Section 5.1). Then, theoretical and pedagogical implications for collocation teaching and learning are discussed (Section 5.2). The thesis concludes by stating its limitations and proposing suggestions for future research (Section 5.3).

6.1. Major Findings

This study has attempted to investigate the four distributional patterns in the use of the three subtype collocations of VNCs, ANCs and NNCs, in the middle and high school English textbook and Korean EFL learner writing corpus. The first research question is concerned with the use of collocations in the textbook materials, with four sub-questions on distributional variables; that is, collocation density, diversity, repetition rate, and association strength. The first variable, collocation density, is related to the overall proportion of collocations appearing in the textbook corpus and whether collocations are sufficiently provided in the language input in the materials. In Chapter 4, we showed that textbook corpus presents a higher proportion of VNCs and ANCs than the reference corpus. The analysis of collocation density indicates that Korean EFL learners may be exposed to a large number of collocations through language input in their textbooks. The density of NNCs is lower in the textbook corpus than the reference corpus, but with no significant difference.

To address the second subquestion regarding collocation diversity in the textbook materials, we examined to what extent collocation repertoire is restricted in the target corpus. Our data has shown a higher diversity of VNCs and ANCs in

the materials than in the reference corpus. Supporting Koproski's statement, "more is less" (2005, p. 329), we suggested that higher collocation density and diversity can be disadvantageous**;** the extensive use of collocation with higher diversity and density in textbook materials may compromise the intensity of collocational input.

The third variable, repetition rate, is an important element of formulaic language learning. According to Ellis (2001), repeated exposure to target items benefits learners with frequency effect, permitting a deeper entrenchment of word association into learners' memory trace. The current data reveals, however, that the textbook materials are somehow inauthentic in the way individual collocation types are repeated to a lesser degree, and consequently the formulaic nature of collocation as a habitual recurrence is underrepresented in the language input. This suggests the need for sufficient repetition to be provided in the materials so as to enhance the opportunities for learners' to store collocations in their long-term memory.

The last research sub-question investigates the level of association strength. To address this question, we first compared the average (median) logDice score and then examined the correlation between the logDice score and co-occurrence frequency. The estimated high median logDice score in the textbook corpus indicates that the teaching materials present a larger body of somewhat likable, strongly associated collocations. On the other hand, the proportion of collocations with the mid-lower level of association strength is smaller in the textbook corpus, suggesting a need for incorporating less-than-typical collocations into the materials. When we looked into the correlation between the association measure and the frequency level, the weakest level of correlation in the ranks between the two variables was found in the textbook corpus. This indicates that the input frequency of collocation and

association strength do not reliably predict each other; that is, the frequency distribution of co-occurrence in the materials is not a strong indicator of an associative relationship. In general, the association strength of collocation used in the textbooks seems to be limited in range and somehow underrepresented, without being correlated with frequency signals. This may give rise to concerns about learners' development of collocational sensitivity and exposure to a wide range of association strengths.

The second main research question addresses the use of collocation in the Korean EFL learner writing corpus. As for the collocation density, learners seem to rely heavily on prefabricated units of the two subtypes, presenting a higher proportion of VNCs and ANCs than the reference corpus. The distribution differs on the subtypes, however, such that the significantly lower density of NNCs was found in the learner corpus than the reference corpus.

The analysis of the second variable, collocation diversity, revealed that learners tend to use VNCs and ANCs with higher diversity rate, while a significantly lower diversity rate being found in NNCs. This finding counters the assumed link between higher collocation diversity and the native-like command of collocation. It was reasoned that mature native speakers find a balance between formulaic and creative language use, whereas learners' language development remains at the formulaic stage. In addition, supporting Paquot (2018) who maintains that the collocation diversity rate may not reliably demarcate language proficiency, we suggest that learners' limited repertoire of collocation can better be explained by variables other than the diversity rate, such as association strength.

The analysis of association strength in the learner corpus shows that learner

production is highly restricted to collocations with high logDice scores, indicating their reliance on typical and likely associations. Conversely, the observed scarcity of lower-mid level association in the learner corpus seems to support the idea that learners may have difficulties with middle ground restricted collocations due to the "challenge in differentiating between combinations that are free and those that are somehow limited in substitutability," as Howarth (1998, p. 42) proposed. Hence, special attention is needed to address the observed scarcity of lower-mid level associations in the learner corpus.

The last finding worth noting is the difference between the subtypes of collocation. We found the opposite trend in the use of NNCs, which occur in the target corpus with lower density and diversity but higher repetition. In particular, Korean EFL learners' underuse of NNCs, in contrast to their preference for ANCs, seems to support the developmental trajectory of complex noun phrases hypothesized by Biber et al., (2011), suggesting that, for most Korean EFL writers, the learning of NNCs could be more difficult than other subtypes and that this subtype may need to be given more coverage in textbook materials.

Taken together, the use of VNCs and ANCs in both textbook and learner corpora can be summarized as having higher density, diversity, and association strength, but less repetition in comparison to the reference corpus. Discussing corroborating findings reported in the previous research in Chapter 5, we first suggested that the extensive coverage of collocations in the materials may compromise the amount of repetitive exposure through language input, jeopardizing learning efficiency. Besides, a mismatch of the co-occurrence frequency with the association score in the materials was highlighted in relation to the data regarding

learners' restricted collocational repertoire and lack of native-like sensitivity to a wider range of associative relationship. Finally, the paucity of collocational knowledge was also observed in Korean EFL learners' use of NNCs, which occur in their writing with significantly lower density and diversity, but higher repetition.

6.2. Theoretical and Pedagogical Implications

6.2.1. Theoretical Implications

Although the collocation use in the reference corpus was not part of the major research questions, the analysis was nevertheless carried out to provide a native baseline for data interpretation, and the result offers several theoretical implications on the nature of collocation in native English. Overall, the higher repetition rates exhibited in the native reference corpus strengthens the Firthian notion of collocation, which is theorized as habitually co-occurring word combinations. The results also lend support for the usage-based theory which emphasizes the role of repeated exposure to lexical units as a driving force of L1 development and the emergence of language structure. If we assume that the implicit tallying of collocational probabilities is central to developing sensitivity to collocations, as Ellis (2002) suggested, then learners in the natural setting are more likely to notice the formulaicity of word associations through repeated exposure and develop native-like control of collocations. The facilitative role of recurrent patterns is acknowledged by Durrant and Schmitt (2010), who pointed out that repetition is essential in forging and strengthening the association links between the constituents. They suggest that substantial exposure to a language is necessary in order to gradually build up knowledge of a large number of collocational pairs.

In addition, the analysis undertaken here provides empirical data to understand the nature of associative relations in English. Chapter 3 introduced the concept of mutual expectancy as another criterion to define the association between constituent words; the probability that a pair of words co-occur more often than individual frequencies. In the current data, the reference corpus contains a larger body of collocations at the lower-mid ranges of association strength, presenting lower median logDice score than the textbook and the learner corpus. This indicates that relatively weaker associations may represent the prototypical idiomaticity of collocation in natural English, which reflects the theoretical notion of collocation – the lexical units which take the "middle" road between idioms and free associations (Cowie, 1998; Howarth, 1998; Nesselhauf, 2003; Schmid, 2003). Supporting Howarth (1998) who acknowledged the significance of the large and complex middle ground of restricted collocations, our data demonstrated that there is a nativity in the collocational relationship which is somehow loosely associated and relatively less likable collocations, rather than strongly associated and thus highly probable ones.

6.2.2. Pedagogical Implications

The analysis of the textbook corpus raises pedagogical issues concerning the expected benefits and shortcomings of collocation learning using the current textbook materials. First, the data suggest that Korean learners' knowledge of collocations may be better achieved by increasing the number of repetitions in textbook materials. Previous studies have established that repeated exposure to collocation benefits learners in noticing its formulaicity and developing sensitivity to associative strength. Tsai (2015), for example, proposed that collocations with

high pedagogical value should be revisited in a principled manner throughout the curriculum. Similarly, another practical application might involve the provision of verbatim repetition in post-reading activities, which Durrant and Schmitt (2010) found to be the most effective way to consolidate collocational knowledge.

Next, the study may contribute to existing knowledge of collocation by providing empirical data to determine the optimal level of association strength which would worth to be the focus of the instruction for EFL learners. The current data highlights the significance of collocations with lower mid-level association strength, by demonstrating a relative prominence of lower logDice median scores in the reference corpus. The current textbook materials, in contrast, seem to fall short of items associated at such a moderate level (logDice score of 5-7.5). According to Boers and Lindstromberg (2009), in order to extend learners' collocational repertoire the chunks in the middle frequency should be prioritized over those in the highest and lowest bands. They contend that teacher intervention is most fruitful when directed toward not-so-frequent chunks as students are less likely to learn these incidentally. On the same grounds, the current study proposes that efforts are needed to supplement collocations at the lower-mid level of association strength, which are seemingly underused in the current textbook materials.

Furthermore, textbook data has shown a weaker correlation between association strength and co-occurrence frequency. This may lead to textbook users' difficulties with predicting the collocation strength based on the degree to which they encounter the items. In contrast, the higher correlation between the two variables in the reference corpus shows that collocational strength may be more predictable in natural English, since the language users are given more frequency cues on the level

of associative strength. Increasing the level of correlation input between association strength and frequency of individual collocation, therefore, could be suggested as a way to foster learners' sensitivity to word association.

Turning to the collocational knowledge exhibited in the writings of Korean EFL learners, the data reveals inconsistencies in the level of learners' collocational knowledge which needs to be addressed in pedagogical practice. The first issue raised for Korean learners' use of collocations is the underuse of NNCs. The lower density and diversity of NNCs in learner writings may indicate learners' avoidance of the use of NNCs due to their unfamiliarity to this subtype. Interestingly, the use of NNCs was also distinguished from VNCs and ANCs in the textbook materials, for which we may find some link between the insufficient input and learners' underuse of this subtype. It was reported that in the textbook corpus, VNCs and ANCs are significantly overrepresented, exhibiting a higher density and diversity. It is then possible that these two structures are more familiar to learners and thus used with higher density and diversity in learner writings. By contrast, NNCs are somehow underrepresented than the other subtypes in the textbook corpus, and this insufficiency of collocation input might possibly lead to learners' shortage of collocational knowledge. As Durrant and Schmitt (2009) suggest, the level of exposure to collocations has a significant impact on L2 collocation acquisition, and insufficient coverage of a particular subtype may lead to structurally impoverished collocation use. If structures emerge from the repetitive encounter, as explained by Ellis (2002), EFL learners' use of collocation subtypes may be affected by its distribution in the textbook materials. Chang's (2018) suggestion that the difference in distributional patterns between subtypes also needs to be considered in teaching

collocations may also be applicable to the present study. The study therefore suggests a need to identify specific subtypes which have not been given sufficient input from the textbook materials, and underused by learners, and provide more input for these underused types of collocation.

Another pedagogical implication is concerned with the association strength of collocation in the learner corpus. It was observed that learners tend to overuse strongly associated and thus more predictable collocations, seemingly lacking the knowledge of "less predictable" combinations. As the difficulty of learning collocations with mid-level restriction and learners' typical insensitivity to different levels of association strength were suggested in the previous Chapter 4 and 5, it is recommended to provide examples of collocations with a wider range of association strength, especially focusing on the loosely connected word pairs. In addition, a weak correlation between co-occurrence frequency and association strength suggests the frequency of collocations used in learner corpus has little accordance with the association strength attested in native English. Given that language develops through the implicit tallying of co-occurrence frequency (Ellis, 2002), continued efforts are needed to introduce a broader range of association strengths with due frequency to ensure the authentic representation of collocation in language input and to develop learners' sensitivity to association strength.

6．3． Limitations and Suggestions for Future Research

There were several limitations of the data collection which may have affected the findings. First, it was noted in Section 3 that the different corpus size brought challenges in comparing the frequency data. Comparison with a large-sized

reference data has both benefits and shortcomings. While the reference corpus used in the present study is currently available samples of maximum representativeness, and thus its analysis provides a look into the native "norm," the effects from the different sample size were an unavoidable issue. Although we attempted to use various mathematic formulae to reduce sample size effect, they may still be insufficient to accurately standardize the corpus size. Furthermore, the distribution of lexical items in the large-sized corpus are different from those in the sampled writing by nature; thus, careful interpretation is required when comparing corpus of different size.

Another limitation is the comparison between corpora with different timelines. It can be questioned whether textbook materials based on the 2015 National Curriculum have any relevance to the learner writing corpus compiled in 2009. In this study, we used textbook materials to represent the expected input given to learners, and the YELC corpus as a sample for Korean EFL learner writings. These target corpus, however, are compiled in different time periods, and any inference driven from the two should remain only hypothetical. In order to draw on more direct comparison, it would be necessary to use the up-to-date corpus of learner productions.

Next, the mathematical operationalization of collocation density, diversity, and repetition may be rather unreliable. It further needs to be verified whether the metrics show suitability for counting the co-occurrence of word pairs as for a single word level. Although the current study attempts to validate the estimated result by applying two formulae for each of the measures, the applicability of such measures as TTR and modified RTTR to collocational level may need deeper investigation.

Lastly, since the present study is primarily based on frequency measures,

strictly following statistical criteria to automatically identify the maximum number of associative links, various relevant variables such as opacity or transparency of meaning, degree of L1-L2 congruency, task conditions (e.g., listening vs. reading) relations have not been examined.

There are a number of additional ways to enrich the present research topic. First, future research can exploit the compatible corpora in their timeline and size. The analysis of more recent learner data, produced by those who use the revised textbooks based on the 2015 National Curriculum, will allow a more plausible comparison between the input provided in the textbook materials and the subsequent learner productions. Furthermore, native writing corpus of similar size would permit ease of comparison with the target corpus.

Next, the correlation between learner proficiency and other four measures can be investigated. The current study analyzed each of the measures estimated in Korean learners' writing in general, but did not examine the validity of each measure for demarcating collocational competence. A comparison of the measures between different proficiency groups will aid understanding of which aspect of collocation use is more relevant than any other.

Further research may include other variables based on the phraseological notion of collocations. Analysis of the relationship between association measures such as logDice score and semantic features may provide a more in-depth explanation of the difference in association strength.

References

Ackermann, K., & Chen, Y. H. (2013). Developing the Academic Collocation List (ACL) - A corpus-driven and expert-judged approach. *Journal of English for Academic Purposes*, *12*(4), 235–247.

Altenberg, B. (1991). Amplifier collocations in spoken English. *English Computer Corpora: Selected Papers and Research Guide*, *128*.

Arnon, I., & Snider, N. (2010). More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language*, *62*(1), 67–82.

Baisa, V., & Suchomel, V. (2014). SkELL: Web interface for English language learning. *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN*, 63–70.

Baroni, M., Kilgarriff, A., Pomikálek, J., & Rychlý, P. (2006). WebBootCaT: a web tool for instant corpora. *12th EURALEX International Congress*, (2003), 123–131.

Barrs, K. (2016). *Using the Sketch Engine Corpus Query Tool for Language Teaching*. (April).

Bartsch, S. (2004). *Structural and functional properties of collocations in English: A corpus study of lexical and pragmatic constraints on lexical co-occurrence*. Gunter Narr Verlag.

Bartsch, S., & Evert, S. (2014). Towards a Firthian notion of collocation. *Online Publication Arbeiten Zui Linguistik*, (1985), 48–61.

Bestgen, Y., & Granger, S. (2018). Tracking L2 writers' phraseological development using collgrams: Evidence from a longitudinal EFL corpus. In *Language and Computers* (Vol. 81, pp. 277–301).

Biber, D., & Clark, V. (2002). Historical shifts in modification patterns with complex noun phrase structures. *Teresa Fanego, Maria Lépez—Couso and Javier Perez—Guerra (Eds.). English Historical Morphology. Selected Papers From*, *11*, 43–66.

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Retrieved from https://books.google.co.kr/books?hl=ko&lr=&id=0HUhombmOJUC&oi=fnd&pg=PR9&dq=Biber+%26+Conrad+2009&ots=tMUsm9ANDu&sig=R5ARewt5ImXYc_GFGzl_Uln5Ids#v=onepage&q=Biber %26 Conrad 2009&f=false

Biber, D., & Gray, B. (2011). Grammatical change in the noun phrase: the influence of written language use. *English Language and Linguistics*, *15*(02), 223–250.

Biber, D., Gray, B., & Poonpon, K. (2011). Should We Use Characteristics of Conversation to Measure Grammatical Complexity in L2 Writing Development? *TESOL Quarterly*, *45*(1), 5–35.

Boers, F., Demecheleer, M., Coxhead, A., & Webb, S. (2014). Gauging the effects of exercises on verb-noun collocations. *Language Teaching Research*, *18*(1), 54–74.

Boers, F., & Lindstromberg, S. (2009). *Optimizing a lexical approach to instructed second language acquisition*. Springer.

Bybee, J. (1998). The Emergent Lexicon. *Chicago Linguistic Society*, *34*(2), 421–435.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language*, *82*(4), 711–733.

Chang, Y. (2018). Features of Lexical Collocations in L2 Writing. *English Teaching*, *73*(2), 3–36.

Chen, W. (2017). Profiling Collocations in EFL Writing of Chinese Tertiary Learners. *RELC Journal*, 003368821771650.

Chen, W. (2019). Profiling Collocations in EFL Writing of Chinese Tertiary Learners. *RELC Journal*, *50*(1), 53–70.

Choi, H. Y., & Chon, Y. V. (2012). A corpus-based analysis of collocations in tenth-grade high school English textbooks. *Multimedia Assisted Language Learning*, *15*(2), 41–73.

Choi, Y., Chon, Y. V, & Han, M.-S. (2015). L2 learners' knowledge of verb-noun collocations: Congruency, L2 proficiency and learning strategies. *Korean Journal of Applied Linguistics*, *31*(3), 31–63.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, *16*(1), 22–29.

Clark, E. V. (2009). *First language acquisition*. Cambridge University Press.

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics*, *32*, 45–61.

Cowie, Anthony P. (1992). Multiword lexical units and communicative language teaching. In *Vocabulary and applied linguistics* (pp. 1–12). Springer.

Cowie, Anthony Paul. (1998). *Phraseology: Theory, analysis, and applications*. OUP Oxford.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing*, *26*, 66–79.

Durrant, P. (2014a). Corpus frequency and second language learners' knowledge of collocations: A meta-analysis. *International Journal of Corpus Linguistics*, *19*(4), 443–477.

Durrant, P., & Doherty, A. (2010). Are high-frequency collocations psychologically real? Investigating the thesis of collocational priming. *Corpus Linguistics and Linguistic Theory*, *6*(2), 125–155.

Durrant, P. L. (2008). High frequency collocations and second language learning (Vol. 35). Retrieved from http://etheses.nottingham.ac.uk/622/

Durrant, P., & Schmitt, N. (2009). To what extent do native and non-native writers make use of collocations? *IRAL - International Review of Applied Linguistics in Language Teaching*, *47*(2), 157–177.

Durrant, P., & Schmitt, N. (2010). Adult learners' retention of collocations from exposure. *Second Language Research*, *26*(2), 163–188.

Ellis, N. (2002a). Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, *24*(2), 143–188.

Ellis, N. C. (1996). Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition*, *18*(1), 91–126.

Ellis, N. C. (2001). Memory for language. *Cognition and Second Language Instruction*, pp. 33–68.

Ellis, N. C. (2002). Reflections on Frequency Effects in Language Processing. *Studies in Second Language Acquisition*, *24*(2), 297–339.

Ellis, N. C. (2003). Constructions, Chunking, and Connectionism: The Emergence of Second Language Structure. *The Handbook of Second Language*

Acquisition, *14*, 63–103.

Ellis, N. C. (2006). Language acquisition as rational contingency learning. *Applied Linguistics*, *27*(1), 1–24.

Ellis, N. C. (2012). Formulaic language and second language acquisition: Zipf and the phrasal teddy bear. *Annual Review of Applied Linguistics*, *32*(2012), 17–44.

Ellis, N. C., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly*, *42*(3), 375–396.

English corpus for SkELL | Sketch Engine. (n.d.). Retrieved January 12, 2020, from https://www.sketchengine.eu/english-skell-corpus/?highlight=skell

Erickson, L. C., & Thiessen, E. D. (2015). Statistical learning of language: Theory, validity, and predictions of a statistical learning account of language acquisition. *Developmental Review*, *37*, 66–108.

Erman, B., & Warren, B. (2000). The idiom principle and the open choice principle. *Text*, *20*(1), 29–62.

Evert, S. (2009). Corpora and collocations. *Corpus Linguistics: An International Handbook*, *2*, 1212–1248.

Fan, M. (2009). An exploratory study of collocational use by ESL students - A task based approach. *System*, *37*(1), 110–123.

Ferrand, L., & New, B. (2003). Semantic and associative priming in the mental lexicon. *Mental Lexicon: Some Words to Talk about Words*, 25–43.

Firth, J. R. (1957). A synopsis of linguistic theory, 1930-1955. *Studies in Linguistic Analysis*.

Flower, J., Berman, M., & Powell, M. (1989). *Build your vocabulary*. Language Teaching Publications.

Foster, P. (2013). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In *Researching pedagogic tasks* (pp. 85–104). Routledge.

Frankenberg-Garcia, A. (2018a). Combining user needs, lexicographic data and digital writing environments. *Language Teaching*, (November 2017), 10–11.

Frankenberg-Garcia, A. (2018b). Investigating the collocations available to EAP writers. *Journal of English for Academic Purposes*, *35*, 93–104.

Frankenberg-Garcia, A., Lew, R., Roberts, J. C., Rees, G. P., & Sharma, N. (2019). Developing a writing assistant to help EAP writers with collocations in real time. *ReCALL*, *31*(1), 23–39.

Frost, R., Armstrong, B. C., Siegelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences*, *19*(3), 117–125.

Gablasova, D., Brezina, V., & McEnery, T. (2017). Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence. *Language Learning*, *67*(June), 155–179.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure*. University of Chicago Press.

Granger, S. (1998). Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae. *World Englishes*, *23*(2), 258.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *IRAL - International*

Review of Applied Linguistics in Language Teaching, *52*(3), 229–252.

Groom, N. (2009). Effects of second language immersion on second language collocational development. In H. Barfield, A., & Gyllstad (Ed.), *Researching collocations in another language* (pp. 21–33). Springer.

Guiraud, P. (1954). *Les characteres statistiques du vocabulaire francaise*. P.

Harmer, J., & Rossner, R. (1997). *More than words: vocabulary for upper intermediate to advanced students: book 2*. Addison Wesley Longman.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development–a progress report. *C. Bardel, C. Lindqvist, & B. Laufer (Eds.) L*, *2*, 29–56.

Hirata, Y., & Hirata, Y. (2018). Students' Evaluation of SkELL: The 'Sketch Engine for Language Learning.' *International Conference on Blended Learning*, 368–377. Springer.

Hoey, M. (1991). *Patterns of lexis in text* (Vol. 299). Oxford University Press Oxford.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Taylor & Francis.

Howarth, P. (1998). Phraseology and second language proficiency. *Applied Linguistics*, *19*(1), 24–44.

Howarth, P. A. (1996). Phraseology in English academic writing: Lexicographica Series Maior, 75. *Tubingen: Max Niemeyer*.

Hunston, S. (2002). *Corpora in applied linguistics / Susan Hunston.* Cambridge ; New York: Cambridge University Press.

Jui-Hsin, Wang, T., & Good, R. L. (2007). *The Repetition of Collocations in EFL*

Textbooks: A Corpus Study. Retrieved from https://files.eric.ed.gov/fulltext/ED502758.pdf

Kim, N. . (2004). Collocational Analysis of Korean High School English Textbooks and Suggestions for Collocation Instruction. *English Language & Literature Teaching*, *10*(3), 41–66.

Kjellmer, G. (1991). A mint of phrases IN English Corpus Linguistics: Studies. In et al. K. Aijmer (Ed.), *in Honour of Jan Svartvik* (pp. 111–127). London: Longman.

Koprowski, M. (2005). Investigating the usefulness of lexical phrases in contemporary coursebooks. *ELT Journal*, *59*(4), 322–332. https://doi.org/10.1093/elt/cci061

Koya, T. (2004). Collocation Research Based on Corpora Collected from Secondary School Textbooks in Japan and in the UK. *Dialogue*, *3*(3), 7–18.

Lafferty, J. D., Lebanon, G., & Lafferty, J. D. (2002). Cranking: Combining Rankings Using Conditional Probability Models on Permutations. *Proceedings of the Nineteenth International Conference on Machine Learning*, 363–370. Retrieved from http://dl.acm.org/citation.cfm?id=645531.655830

Lapata, M. (2003). Probabilistic text structuring: experiments with sentence ordering. *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, 545–552.

Laufer, B., & Waldman, T. (2011). Verb-Noun Collocations in Second Language Writing: A Corpus Analysis of Learners' English. *Language Learning*, *61*(2), 647–672.

Lee, Jinkyong. (2015). The Repetition of Chunks in Korean Middle School English

Textbooks. *English Language Teaching*, *8*(10), 60–75.

Lee, M.-B., & Shin, D.-K. (2015). Development of the Korean basic English word list of the 2015 revised national curriculum of English. *Journal of the Korea English Education Society*, *14*(4), 115–134.

Lewis, M., Gough, C., Martínez, R., Powell, M., Marks, J., Woolard, G. C., & Ribisch, K. H. (1997). *Implementing the lexical approach: Putting theory into practice* (Vol. 3). Language Teaching Publications Hove.

Lorenz, G. R. (1999). *Adjective intensification: learners versus native speakers: a corpus study of argumentative writing* (Vol. 27). Rodopi.

Lu, X. (2011). A Corpus-Based Evaluation of Syntactic Complexity Measures as Indices of College-Level ESL Writers' Language Development. *TESOL Quarterly*, *45*(1), 36–62.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT press.

McCarthy, M., & O'Dell, F. (1994). *English vocabulary in use: 100 units of vocabulary reference and practice*. Cambridge University Press.

Men, H. (2015). *Vocabulary Increase and Collocation Learning : A Corpus-Based Cross-Sectioal Study of Chinese EFL Learners A Thesis for the Degree of Doctor of Philosophy March*. (March).

Men, H. (2018). *Vocabulary Increase and Collocation Learning*.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

Nesselhauf, N. (2003). The Use of Collocations by Advanced Learners of English and Some Implications for Teaching. *Applied Linguistics*, *24*(2), 223-

242+268.

Northbrook, J., & Conklin, K. (2018). Is What You Put in What You Get Out? — Textbook-derived Lexical Bundle Processing in Beginner English Learners. *Applied Linguistics*, 1–19.

O'keeffe, A., McCarthy, M., & Carter, R. (2007). *From corpus to classroom: Language use and language teaching*. Cambridge University Press.

Paquot, M. (2018). Phraseological Competence: A Missing Component in University Entrance Language Tests? Insights From a Study of EFL Learners' Use of Statistical Collocations. *Language Assessment Quarterly*, *15*(1), 29–43.

Paquot, M. (2019). The phraseological dimension in interlanguage complexity research. *Second Language Research*, *35*(1), 121–145.

Paquot, M., & Granger, S. (2012). Formulaic language in learner corpora. *Annual Review of Applied Linguistics*, *32*, 130–149.

Paquot, M., & Naets, H. (2015). Adopting a relational model of co-occurrences to trace phraseological development. *Learner Corpus Research 2015*.

Parkinson, J. (2015). Noun-noun collocations in learner writing. *Journal of English for Academic Purposes*, *20*, 103–113.

Parkinson, J., & Musgrave, J. (2014). Development of noun phrase complexity in the writing of English for Academic Purposes students. *Journal of English for Academic Purposes*, *14*, 48–59.

Pawley, A., & Syder, F. H. (1983). Natural selection in syntax: Notes on adaptive variation and change in vernacular and literary grammar. *Journal of Pragmatics*, *7*(5), 551–579.

Pawley, A., Syder, F. H., Richards, J. C., & Schmidt, R. W. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. R. & R. W. Schmidt (Ed.), *Language and communication* (pp.191-225). London: Longman.

Perkins, M. R. (1999). Productivity and formulaicity in language development. *Issues in Normal and Disordered Child Language: From Phonology to Narrative*, (January 1999), 51–67.

Peters, E. (2014). The effects of repetition and time of post-test administration on EFL learners' form recall of single words and collocations. *Language Teaching Research*, *18*(1), 75–94.

Peters, E. (2016). The learning burden of collocations: The role of interlexical and intralexical factors. *Language Teaching Research*, *20*(1), 113–138.

Redman, S., Ellis, R., & Viney, M. (1989). *A way with Words*. Cambridge University Press Cambridge.

Rhee, S., & Jung, C. (2012). Yonsei English learner corpus (YELC). *Proceedings of the First Yonsei English Corpus Symposium*, 26–36.

Rudzka, B., Channell, J., Putseys, Y., & Ostyn, P. (1981). *The words you need*. macmillan London.

Rychlý, P. (2008). A lexicographer-friendly association score. *Raslan*, 6–9.

Schmid, H.-J. (2016). *Entrenchment and the psychology of language learning: how we reorganize and adapt linguistic knowledge*. Walter de Gruyter GmbH & Co KG.

Schmid, H. J. (2003). Collocation: Hard to pin down, but bloody useful. *Zeitschrift Fur Anglistik Und Amerikanistik*, *51*(3), 235–258.

Schmitt, N, Candlin, C. N., & Hall, D. R. (2010). Researching Vocabulary: A Vocabulary Research Guide. *Language Teaching*.

Schmitt, Norbert. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Seretan, V. (2005). Induction of syntactic collocation patterns from generic syntactic relations. *IJCAI International Joint Conference on Artificial Intelligence*, 1698–1699.

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

Siyanova-chanturia, A. (2015). Collocation in beginner learner writing : A longitudinal study. *System*, *53*, 148–160.

Siyanova-Chanturia, A., & Janssen, N. (2018). Production of familiar phrases: Frequency effects in native speakers and second language learners. *Journal of Experimental Psychology: Learning Memory and Cognition*, *44*(12), 2009–2018.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *Canadian Modern Language Review*, *64*(3), 429–458.

Sonbul, S., & Schmitt, N. (2013). Explicit and Implicit Lexical Knowledge: Acquisition of Collocations Under Different Input Conditions. *Language Learning*, *63*(1), 121–159.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. *Functions of Language*, *2*(1), 23–55.

Stubbs, M. (2001). *Words and phrases: Corpus studies of lexical semantics*.

Blackwell Publishers Oxford.

Sung, M.-C. (2017). Verb-noun collocations in EFL writings: Crosslinguistic influence and input frequency. *Korean Journal of Applied Linguistics*, *33*(3), 91–115.

Tomasello, M. (2009). *Constructing a language*. Harvard university press.

Toomer, M., & Elgort, I. (2019). The Development of Implicit and Explicit Knowledge of Collocations: A Conceptual Replication and Extension of Sonbul and Schmitt (2013). *Language Learning*, *69*(2), 405–439.

Tsai, K.-J. (2015). Profiling the Collocation Use in ELT Textbooks and Learner Writing. *Language Teaching Research*, *19*(6), 723–740.

Webb, S., Newton, J., & Chang, A. (2013). Incidental Learning of Collocation. *Language Learning*, *63*(1), 91–120.

Williams, J. (2019). Talking about the weather: exploring collocation with SkELL. In *New Ways in Teaching with Corpora*. TESOL Press.

Wolter, B., & Gyllstad, H. (2013). Frequency of input and L2 collocational processing. *Studies in Second Language Acquisition*, *35*(3), 451–482.

Wray, A. (2005). *Formulaic language and the lexicon*. Cambridge University Press.

Yi, W. (2018). Statistical Sensitivity, Cognitive Aptitudes, and Processig of Collocations. *Studies in Second Language Acquisition*, *40*(4), 831–856.

Appendix

List of English Textbooks for Research

Lee, Jae-young, Ahn, Byeong-gyu, Oh, Jun-il, Bae, Tae-il, Kim, Suncheon, Park, Seong-geun & Shin, Sujin (2017-8) *Middle School English 1, 2*. Seoul: Chunjae Education Inc.

Min, Chankyu, Kim, Yunkyu, Jung, Hyun-sung, Lee, Sang-gi, Choi, Jin-hee, Park, Se-ran, Yeom, Ji-sun & Foreman, Walter (2017-8) *Middle School English 1, 2*. Seoul: Jihaksa Publishing Co.

Kim, Jinwan, Hwang, Jong-bae, Yin, Judy, Lee, Yun-hui, Shin, Migyeong, Jo, Seong-ok & Jo, Hyunjung (2017-8) *Middle School English 1, 2*. Seoul: Visang Education.

Kang, Yongsun, Kim, Haedong, Kwon, Hye-yeon, Whitehead, George Elliott, Gu, Na-hyeon, Han, Kyung & Hong, Kiman (2017-8) *Middle School English 1, 2*. Seoul: Darakwon Inc.

Kim, Seong-gon, Seo, Seong-gi, Lee, Seok-yeong, Choi, Dongseok, Kang, Yong-gu, Kim, Seong-ae, Choi, Incheol, Yang, Bit-na & Jo, Yu-ram (2017-8) *Middle School English 1, 2*. Seoul: Neungyule Inc.

Choi, Incheol, Park, Taeja, Suh, Wonhwa, Hong, Woojung, Kang, Yuna, Song, Haeri, Kim, Ji-yun & Lee, Jeong-ha (2017-8) *Middle School English 1, 2*. Seoul: Kumsung.

Park, Jun-eon, Kim, Myeong-hui, Park, Byeong-ryun, Yang, So-yeong & Choi, Heejin (2017-8) *Middle School English 1, 2.* Seoul: YBM.

Song, Mi-jeong, Kwon, Jin-a, Mo, Yoonsook, Shin, Jeong-a, Lee, Sooha, Yoo, Hyeonju & Jung, Ji-yun (2017-8) *Middle School English 1, 2*. Seoul: YBM.

Lee, Jaeyoung, Ahn, Byeong-gyu, Oh, Jun-il, Moon, Anna, Kim, Choonsoo, Kim, Hyunjin & Manning, Shaun Justin (2017) *High School English*. Seoul: Chunjae Education Inc.

Min, Chan-gyu, Jung, Hyunsung, Lee, Sang-gi, Kim, Yun-gyu, Kwak, Noh-jin, Won, Jangho, Woo, Eo-jin & Klinkner, Robin Eric (2017) *High School English*. Seoul: Jihaksa Publishing Co.

Han, Sang-ho, Jung, Eun-gwi, Park, Seon-ha, Lee, Bo-hui, Lee, Hye-eun & Jang, Eun-gil (2017) *High School English*. Seoul: YBM.

Hong, Min-pyo, Ahn, Hyunki, Park, Yeon-mi, Kim, Jungtae, Jang, Hyun-ok, Shin, Jeongseop, Jo, Keumhee & Pak, Richard (2017) *High School English*. Seoul: Visang Education.

Kim, Kiljoong, Putlack, Michael Aaron, Im, Jeongwon, Jang, Jinhwa, Kim, Gunwoo, Kim, Na-hyeon & Ahn, Migyeong (2017) *High School English*. Seoul: Darakwon Inc.

Kim, Seong-gon, Yun, Jin-ho, Koo, Eun-young, Jeon, Hyeong-ju, Seo, Jeong-hwan, Lee, Hooko, Kim, Yun-ja, Kang, Yong-gu, Kim, Seong-ae, Choi, In-cheol, Kim, Ji-yeon, Shin, Yooseung (2017) *High School English*. Seoul: Neungyule Inc.

Choi, Incheol, Park, Riri, Jang, Min-gyeong, Chae, Ji-seon, Kim, Geun-yeong, Choi, Sooha, Kim, Ju-hye, Son, Ji-hye, Jeon, Ye-ji & Ksan, Rubadeau (2017) *High School English*. Seoul: Kumsung.

Park, Jun-eon, Yoon, Byungwoo, Kim, Seon-hyeong, Choi, Sunyoung, Choi, Jiyoung, Choi, Song-yi, Kim, Jin-su & Kim, Ha-yeong (2017) *High School English*. Seoul: YBM.

Lee, Jaeyoung, Moon, Anna, Oh, Young-il, Jo, Sugyeong, Lee, Yunjeong, Kim, Junghyun & Manning, Shaun Justin (2017-8) *High School English*, *English 1*, *English 2*. Seoul: Chunjae Education Inc.

Min, Chan-gyu, Jung, Hyunsung, Lee, Sang-gi, Kim, Yun-gyu, Na, Woochul, Ahn, Hyoseon, Woo, Eo-jin & Klinkner, Robin Eric (2017) *High School English 1*. Seoul: Jihaksa Publishing Co.

Han, Sang-ho, Jung, Eun-gwi, Kim, Yeri, Kim, Jaeran, Lee, Bo-hui & Lee, Hye-eun (2017-8) *High School English1*, *High School English 2*. Seoul: YBM.

Hong, Min-pyo, Ahn, Hyunki, Park, Yeon-mi, Kim, Jungtae, Jang, Hyun-ok, Shin, Jeongseop, Jo, Keumhee & Pak, Richard (2017) *High School English 1*. Seoul: YBM.

Kim, Kiljoong, Putlack, Michael Aaron, Im, Jeongwon, Jang, Jinhwa, Kim, Gunwoo, Kim, Na-hyeon & Ahn, Migyeong (2017-8) *High School English*, *English 1*, *English2*. Seoul: Darakwon Inc.

Kim, Seong-gon, Yun, Jin-ho, Koo, Eun-young, Jeon, Hyeong-ju, Seo, Jeong-hwan, Lee, Hooko, Kim, Yun-ja, Kim, Ji-yeon & Baek, Su-ja (2017) *High School English 1*. Seoul: Neungyule Inc.

Choi, Incheol, Park, Riri, Kim, Geunyeong, Choi, Sooha, Kim, Ju-hye, Son, Ji-hye, Jeon, Ye-ji & Ksan, Rubadeau (2017) *High School English 1*. Seoul: Kumsung.

Park, Jun-eon, Yoon, Byungwoo, Kim, Seon-hyeong, Choi, Sunyoung, Choi, Jiyoung, Kim, Jin-su & Kim, Ha-yeong (2017-8) *High School English*, *English 1*, *English 2*. Seoul: YBM.

Min, Chan-gyu, Jung, Hyunsung, Lee, Sang-gi, Kim, Yun-gyu, Na, Woochul, Ahn, Hyoseon, Won, Jangho & Klinkner, Robin Eric (2018) *High School English 2*. Seoul: Jihaksa Publishing Co.

Hong, Min-pyo, Ahn, Hyunki, Pak, Richard, O'Flaherty, David Desmond, & Jo, Keumhee (2018) *High School English 2*. Seoul: Visang Education.

Kim, Seong-gon, Yun, Jin-ho, Jeon, Hyeong-ju, Seo, Jeong-hwan, Lee, Hooko, Kim, Yun-ja, Kim, Ji-yeon & Jeon, Sungho (2018) *High School English 2*. Seoul: Neungyule Inc.

Choi, Incheol, Seo, Wonhwa, Lee, Yoonkyung, Kim, Ju-hye, Jeon, Ye-ji, La, Moonsun, Moon, Youngsun & Ksan, Rubadeau (2018) *High School English 2*. Seoul: Kumsung.

국 문 초 록

둘, 혹은 그 이상의 단어가 습관적으로 연합되는 단위를 의미하는 연어는, 효율적인 언어 산출과 처리를 가능하게 하며 원어민의 유창성을 구성하는 중요한 언어적 요소인 정형화된 어구의 하위 유형에 속한다. 정형화된 어구의 반복적인 사용을 언어 생성의 기본 원리라 설명하는 Sinclair의 '숙어적 원리' 에 따르면, 원어민 화자는 영어 문법이 허용하는 무한한 조합을 새로 생성하는 대신 미리 짜여진 제한적인 수의 정형화된 표현을 반복적으로 사용한다. 숙어적 원리에 의해 생성된 정형화된 표현은 원어민의 영어에서 큰 비율을 차지하는데, 모국어를 습득하는 원어민 화자는 이들 항목에 빈번하게 노출됨으로써, 이들을 장기기억/어휘장에 깊이 각인시키고 자동화할 수 있다. 나아가 이러한 정형화된 단어 간 결합은 제한적인 어휘만을 사용할 수 있는 초기 언어 발달 단계에서 언어 사용을 촉진하는 요인으로 알려져 있다.

한편, 연어는 완전히 고착화된 숙어적 표현과, 반대로 임의적 선택에 의해 생성된 자유 결합의 중간 정도에 해당하는 정형성을 지닌 것으로 알려져 있다. 이렇듯 정형화 표현인 동시에 유연한 선택의 폭을 허용하는 연어의 '애매한' 결합적 속성으로 인해, 적절한 연어 사용은 원어민과 비원어민의 어휘 사용을 구분 짓는 특성이자, 제2외국어 학습자들이 많은 어려움을 겪는 영역으로 알려져 있다. 이러한 연어의 중요성과 학습상의 어려움에도 불구하고, 현재 한국 영어 어휘 교육과정은 개별 단어 수준의 목록만을 제시하고 있으며, 검정 교과서에 대한 교육과정 지침은 어휘의 적절한 사용보다는, 어휘 목록의 일정한 수준의 포함률(coverage)를 유지하는데 초점이 맞춰져 있는 것으로 보인다. 또한 기존의 어휘 관련 연구는 개별 어휘 항목에 대해서 어휘적 다양성과 난이도를 측정할 수 있는 다양한 도구들을 개발하고 활용해 왔으나, 연어 항목에 대해

서는 이들의 결합적 특성에 맞춘 측정법과 도구를 사용한 사례가 많지 않은 상황이다.

이에 본 연구에서는 한국 영어 교실상황에서의 주 언어 입력에 해당하는 중·고등학교 영어 교과서와, 정규교육과정을 마친 한국인 학습자의 언어 산출을 대표하는 대학 신입생 영작문을 분석하여, 한국의 EFL 상황에서의 언어 입력 및 학습자의 언어 산출 내 연어의 분포적 특징을 밝히고자 한다. 보다 객관적인 분석을 위해, 대규모 원어민 코퍼스에서 각 단어의 확률적 결합강도를 추출하고, 일정 수준 이상의 결합 강도를 지닌 조합을 연어로 정의하는 양적 접근을 취하였다. 분석에는 코퍼스적 접근 방식이 사용될 것이며 분석 대상 텍스트로는, 2015 개정 중·고등학교 8종 영어 교과서 내 읽기 지문과 연세 영어 학습자 코퍼스(YELC)가 사용되었다. 또한, 대규모 원어민 언어 자료를 사용하여 연어성을 판단하는 참조 기준으로 삼는 동시에, 원어민의 모국어 언어 습득 상황에서 주어지는 언어 입력이자 일반 영어의 연어 사용을 대표하도록 하여, 두 대상 코퍼스의 분석 결과를 해석하는 기준으로 삼았다. 연어의 사용 양상과 관련된 변인으로는, 연어의 광범위한 사용 정도와 관련된 연어 밀도, 다양성 비율과 함께, 각 연어 항목의 사용 강도와 관련된 반복률, 연합강도를 측정하여 해당 텍스트에서 연어의 분포적 특징을 분석하였으며, 언어적 변인으로는, 명사를 핵어로, 동사, 형용사, 명사 연어와 결합하는 동사 결합형 명사 연어(VNC), 형용사 결합형 명사 연어(ANC), 명사 결합형 명사 연어(NNC)의 세 가지 유형을 분석하였다.

먼저, 교과서에서 텍스트 내 연어의 양을 나타내는 상대적 비율인 연어 밀도를 측정한 결과, 교과서에서 원어민보다 VNC, ANC가 유의미하게 높은 밀도를 보이며 내 비교적 많은 수의 연어가 제시되고 있었다. 다음으로, 서로 다른 연어 항목의 유형 수를 나타내는 다양성 비율을 분석한 결과, 교과서 지문이

텍스트 크기에 비해 높은 수준의 VNC와 ANC 다양성 비율을 보이며, 비교적 많은 종류의 연어를 제시하였다. 이는 연어 유형의 개수 대비 사용된 전체 연어 개수의 비율인 반복률을 낮추는 결과로 이어졌다. 즉, 높은 강도의 반복을 통해 정형화된 언어로서의 특징을 보이는 원어민 자료에 비해, 교과서 텍스트 내에서 VNC와 ANC가 반복되는 정도가 낮았다. 마지막으로 결합 강도를 분석한 결과, 교과서에서 사용된 연어들이 보다 강한 결합력을 지니며 확률적으로 공기할 가능성이 높은 조합에 집중되어 있음을 보였다. 즉, 오히려 다소 낮은 결합력을 지닌 연어들까지 다수 분포되어 있는 원어민 참조 코퍼스에 비해, 교과서 텍스트의 연어 사용은 높은 결합력을 지닌 조합에 편향되어 다양한 수준의 연어적 결합 관계를 제시하는 데 한계를 보였다. 또한, 각 연어의 결합 강도와 사용 빈도 간 상관분석에서는, 교과서에서 연어의 사용빈도는 그 결합력에 비례하는 정도가 원어민 데이터보다 낮았다. 즉, 교과서에서 각 연어가 제시되는 빈도는 그 결합 수준에 대해 상당히 낮은 예측력을 지니며, 이는 학습자들이 목표 연어에 노출된 횟수나 친숙도로 결합력을 예측하기 어려울 수 있음을 예상할 수 있다. 또한 결합력과 빈도 수준 간의 불일치로 인해 높은 결합력을 지닌 연어들이 이에 상응하는 빈도로 충분히 반복 제시되지 못해 학습 효율이 낮을 것으로 보였다.

다음으로 한국인 대학생 학습자 코퍼스에서 사용된 연어의 분포를 분석한 결과, VNC와 ANC는 텍스트 길이에 비해 높은 빈도로, NNC는 낮은 빈도로 사용되었다. 즉, 학습자도 전반적으로 연어를 광범위하게 사용하는 경향이 있으나, 특정한 하위 유형에 대해서는 낮은 사용 빈도가 관찰되었다. 연어 다양성의 경우, VNC와 ANC는 텍스트 길이에 비해 사용된 다양한 연어 유형을 사용하고 있으나, NNC의 다양성은 낮은 것으로 나타났다. 연어 결합의 평균적인 강도는 학습자 텍스트에서 원어민 참조 코퍼스보다 높게 나타났다. 즉, 강한 결

합력을 지닌 연어들은 원어민에 상응하는 수준으로 사용하는 반면에, 결합 강도가 다소 낮은 연어들은 사용하지 않는 경향을 보였다. 이는 강한 결합력을 지닌 조합에 한정된 학습자의 제한된 연어 지식을 나태내는 동시에, 중간 정도의 결합 강도를 지닌 연어는 쉽게 예측이 어렵고 평소 언어 입력을 통해 노출될 가능성이 적어 학습자가 어려워한다는 연구 결과를 뒷받침한다. 또한 사용 빈도와 결합력 간 낮은 상관은 학습자 또한 원어민보다 연합강도에 대한 인식이 부족할 수 있음을 시사하였다.

본 연구는 교과서와 학습자영어 글쓰기 코퍼스에 나타난 연어 양상에 의거하여 다음과 같은 교육적 시사점을 제공한다. 먼저, 양적으로 광범위한 연어 항목의 제시가 반드시 효과적이지 않을 수 있다. 원어민의 코퍼스 데이터에서는 연어의 반복률이 높은 것으로 드러났으며, 이는 정형화된 표현으로서의 연어의 본질적, 분포적 특성을 영어 교과서나 교수 학습자료를 통해 구현할 필요가 있음을 시사한다. 두 번째로, 학습자나 교과서 또한 강한 결합 수준을 지닌 연어에 집중된 반면, 원어민 화자는 오히려 중-저 수준의 다소 약한 결합력을 지닌 연어를 많이 사용한다는 것을 고려하여, 보다 넓은 범위의 결합강도를 지닌 연어를 골고루 지도해야 함을 알 수 있다. 마지막으로, 교과서에서 각 연어를 제시하는 빈도와 결합력의 상관을 높임으로써, 연어 결합력에 대한 감각을 높일 수 있는 방안을 제안하고자 한다.

주요어: 연어, 연어 밀도, 연어 다양성, 반복률, 연어 결합력, 정형성, 교과서, 학습자 영어 글쓰기, 코퍼스

학번 : 2018-22214