Association of lexical and collocation knowledge: A comparative analysis of a learner corpus of English and a native speaker corpus

Sung-Yeon Kim
Department of English Education
Hanyang University
sungkim@hanyang.ac.kr

Dongkwang Shin
Department of English Education
Gwangju National University of Education
sdhara@gmail.com

Kyung-Sook Kim
College English Education Center
Hanyang University
cindytesol@hanyang.ac.kr

Abstract

This study compared L2 learners' (EFL) use of single words and collocations with that of native speakers (NSs). The study compiled two text corpora, one from an existing native corpus and the other with texts from Korean EFL learners at three proficiency levels. The study found a significant difference in the lexical diversity of single words between the two corpora. The difference between the advanced EFL learners and NSs was, however, non-significant, which means that advanced learners were on par with NSs in terms of using a diverse range of words. In a comparison of lexical distribution of single words, however, these advanced learners were found to use words between the 1K (the first 1,000 words) and 4K level (4,001 ~ 5,000 words) whereas NSs used words beyond this level. This suggests the need to cover a wider range of words in classroom teaching and assessment. Regarding collocational knowledge, the difference between the NSs and EFL group was apparent and statistically significant, regardless of learners' proficiency levels. Namely, EFL learners used far fewer collocations in a smaller range than their counterparts. The learners' limited collocational competence indicates that collocations should be considered an integral component of the curriculum, instruction, and assessment.

Keywords: single words, collocations, lexical competence, collocational knowledge, corpus.

Resumen

Este estudio comparó el uso de palabras sueltas y colocaciones de los estudiantes de L2 (ILE) con el de los hablantes nativos (HNs). El estudio compiló dos corpus de texto, uno de un corpus nativo existente y otro con textos de estudiantes coreanos de ILE en tres niveles de competencia. El estudio encontró una diferencia significativa en la diversidad léxica de palabras individuales entre los dos corpus. Sin embargo, la diferencia entre los estudiantes avanzados de ILE y los HNs no fue significativa, lo que significa que los estudiantes avanzados estaban a la par con los NS en términos de uso de una amplia gama de palabras. Sin embargo, en una comparación de la distribución léxica de palabras individuales, se encontró que estos estudiantes avanzados usaban palabras entre el nivel 1K (las primeras 1000 palabras) y 4K (4001 ~ 5000 palabras), mientras que los HNs usaban palabras más allá de este nivel. Esto sugiere la necesidad de cubrir una gama más amplia de palabras en la enseñanza y la evaluación en el aula. Con respecto al conocimiento de colocación, la diferencia entre el grupo HNs y ILE fue evidente y estadísticamente significativa, independientemente de los niveles de competencia de los alumnos. Es decir, los estudiantes de ILE utilizaron muchas menos colocaciones en un rango más pequeño que los otros. La competencia de colocación limitada de los alumnos indica que las colocaciones deben considerarse un componente integral del plan de estudios, la instrucción y la evaluación.

Palabras clave: palabras sueltas, colocaciones, competencia léxica, conocimiento colocacional, corpus.

1. Introduction

There are two common generalizations about lexical competence that deserve attention in the era of globalization where language acquisition is facilitated with internet technology and transnational mobility (Duff, 2015; Ma, 2017; Yu & Trainin, 2022). First, native speakers are superior to EFL learners in terms of vocabulary knowledge (Demir, 2017; Zareva 2007); and second, lexical knowledge automatically entails collocational competence (Nation, 2001; Zareva et al., 2005). The first argument has been received as a fact based on a considerable body of research. The second, however, has been questioned by a substantial number of L2 vocabulary studies. Many researchers have cautioned that lexical competence does not imply collocational knowledge (Bahns & Eldaw, 1993; Laufer & Waldman, 2011; Paquot & Granger, 2012). For instance, Siyanova-Chanturia (2015) claimed that L2 learners' general

vocabulary knowledge is not transferred to collocational competence, reporting diverse problems that second language learners often experience when using collocations. This is a particularly thorny problem for language learners that creates a sizable gap between them and their native counterparts.

Learner difficulty with collocations partly arises because they are "a type of formulaic expression made of strongly associated pairs of words characterized by restricted substitutability", (Bestgen, 2017, p. 67), as in the following examples: "make a mistake" vs. "*do a mistake", "strong tea" vs. "*powerful tea", and "heavy drinker" vs. "*extravagant drinker" (Supasiraprapa, 2018). This makes it more difficult for learners to acquire collocations as they have to differentiate the collocational restrictions and use the right combinations. However, classroom instruction and assessment generally focus on single words (SWs) rather than on collocations. In addition, learners' lexical competence is often evaluated based on their knowledge of SWs (Laufer & Nation, 2001; Nation, 2001, 2006; Paquot, 2007; Shin, 2015). This negligence of teaching and assessing collocations is, however, problematic as it can lead to unbalanced development of lexical competence, which is not complete without collocational knowledge. Collocations should not be neglected as they comprise 20 to 50 percent of the spoken and written discourse of native speakers (Erman & Warren, 2000; Foster, 2001; Laufer & Waldman, 2011). In addition, a strong relationship was found between formulaic competence and writing quality, and thus formulaic measures are known to predict writing quality better than single-word measures (Bestgen, 2017). Yet, as discussed in previous studies, L2 learners have problems with collocations, and as a result, underuse, overuse, or misuse them (González-Fernández & Schmitt, 2015; Siyanova-Chanturia, 2015). Against this backdrop, collocation deserves special attention, particularly in an EFL context like Korea, where English teaching disregards language areas such as collocation that are not measured in the college entrance exam (Kim, 2021). However, given that collocational knowledge is known as a yardstick for distinguishing L1 users from L2 users of English, it is critical to assess L2 learners' collocational competence and teach collocations as well as single words (Siyanova-Chanturia, 2015).

In recent years, numerous studies have been conducted on learners' use of collocations; however, they have focused on a certain number of lexical bundles or a limited set of combinations, such as adjective-noun (e.g., "heavy rain"), verb-noun (e.g., "tell lies"), or intensifier-adjective (e.g., "deeply rooted") collocations (Altenberg & Granger, 2001; Boers et al., 2014; Cross & Papp, 2008; Durrant & Schmitt, 2009; García Salido & Garcia, 2018; Granger, 1998a; Kashiha & Chan, 2015; Li & Schmitt, 2010; Nesselhauf, 2003; Siyanova-Chanturia, 2015; Siyanova-Chanturia & Schmitt, 2008). Also, few studies have investigated single words and collocations together (Shin et al., 2018; Vedder & Benigno, 2016), and those have been limited in scope, simply

describing the frequencies of either single- or multi-words. This is unfortunate because they offer an incomplete account of L2 learners' linguistic competence (Bahns & Eldaw, 1993; Read & Nation, 2006). Consequently, little is known about L2 learners' collocational competence.

Given that vocabulary and collocations have rarely been examined together, this study aims to analyze the two using the BNC-COCA25 Range (a vocabulary analysis program loaded with 25 wordlists constructed from the Corpus of Contemporary American English or COCA, and British National Corpus or BNC, Nation, 2012) and COCA_MWU20 ColloGram (a collocation analysis program based on 20 collocation lists constructed from COCA, Shin et al., 2018), respectively, to yield a thorough assessment of L2 learners' lexical competence and examine the association between lexical and collocational knowledge. Considering that these two areas of lexical competence deserve equal attention, they should be included as units of lexical assessment. In addition, L2 learners' knowledge of these two areas should be compared to that of their native counterparts. To ensure the validity of the research, the learner corpus used in the study was carefully compiled from three graded proficiency levels (high, mid, low), and a reference corpus of approximately the same size as the learner corpus was selected from a native corpus (LOCNESS) to facilitate a comparison of the two corpora (Granger, 1998b). Such a comparison allows us to measure what learners across proficiency levels can do relative to their native counterparts in terms of using single words and collocations in their written text production.

2. Literature review

It is generally accepted that lexical knowledge plays a crucial role in shaping target language proficiency (Laufer & Nation, 1995; Lewis, 1993; Milton, 2013; Nation, 2001, 2006; Olinghouse & Leaird, 2009; Schmitt, 2008). This awareness of the role of vocabulary has led to a surge in vocabulary research, further accelerated by computer-aided language analysis (Cobb, 2010; Coxhead, 2000, 2016; Read, 2007). For example, corpus analysis is widely used to measure the size of learner vocabulary using frequency-based wordlists. Researchers have divided English vocabulary into fourteen different levels according to their frequency and the first 3,000-word families (1K to 3K) are generally considered as high-frequency vocabulary, words beyond the 9K frequency bands are treated as advanced words, and those between the first 3K and 9K as mid-frequency words (Leech et al., 2001; Nation, 2006; Schmitt & Schmitt, 2014). An analysis of lexical profiles can be also used to identify the distribution of L2 learners' vocabulary and guide them to learn low-frequency words as well as high-frequency words (Johnson et al., 2016). Ha (2021) analyzed the lexical profiles of input texts used in the IELTS listening and reading tests using the BNC-COCA25 Range program, and found that the 3K words cover 95% of the words in the listening texts, and the 5K

words have 95% coverage of the running words in the reading passages.

Vocabulary research has focused on estimating the vocabulary size required for performing language tasks, describing learner language according to frequency-based wordlists, or analyzing L2 learner language against a native corpus in terms of lexical properties, such as word frequencies, lexical diversity, and lexical density (Dang & Webb, 2014; Gregori-Signes & Clavel-Arroita, 2015; Kao & Wang, 2014; Laufer & Nation, 2001; Nation, 2006; Paquot, 2007; Shin, 2015). For instance, Nation (2006) used fourteen one-thousand-word-family lists (1K though 14K) created from the BNC to analyze the vocabulary size of nine different written and spoken corpora and suggested that L2 learners need receptive vocabulary of the most frequent 8,000 (8K) to 9,000 (9K) words to adequately comprehend written text and 6,000 (6K) to 7,000 (7K) words for spoken text. Recently, Dabbagh and Enayat (2019) investigated the association between productive vocabulary size and writing score using the Vocabulary Levels Test (VLT, Schmitt et al., 2001) to measure learner knowledge of vocabulary at four frequency bands (2000, 3000, 5000, and 10,000 words). They found that mid- (the 5,000-word or 5K) and low (the most frequent 3,000 families or 3K) level vocabulary contributed significantly to EFL students' writing scores. This finding supports an earlier suggestion that teachers and learners should pay attention to 5K words as they cover a large portion of different types of texts (González, 2013; Schmitt & Schmitt, 2014).

There are also other studies that have compared the lexical diversity and distribution of L2 learners with those of L1 writers. Doro (2007), for example, compared Hungarian EFL learners' texts with L1 users' and found that L2 learners used a narrower range of words than L1 users, mostly confined to high-frequency words from the first 2K. In a study conducted with graduate students, Nasseri and Thompson (2021) also reported that EFL graduate students incorporated significantly less word types in their dissertation abstracts compared to L1 users and ESL graduate students. Native speakers (NSs) of English also differed from non-native speakers (NNSs) in terms of lexical distribution (Coxhead & Boutorwick, 2018). While NSs mastered K2 and K3 words by Grade 6 (G6), K5 words by G8, and most of K10 words by G10, NNS acquired K2 words by G6, K3 words by G8, and K5 words by G10. Notably, these studies concentrate only on single words. The question is whether single words alone are sufficient, particularly for processing connected discourse in written language production. This is where knowing a word means "much more than knowing the form-meaning link" (Tsai, 2015, p. 724) through word connections. This connected feature is often observed in the way that words are combined, and collocation is one example of this.

Collocational knowledge plays a pivotal role in shaping lexical competence, given that collocations constitute about half of the native written English corpus (Ehrman

& Warren, 2000; Vedder & Benigno, 2016). According to Ehrman and Warren (2000), more than half (52.3% to be exact) of written discourse consists of formulaic sequences or "collocations" (Dickins, 2020, p. 34). Accordingly, it would be incomplete to estimate lexical competence without considering the ability to understand and use collocations. Unfortunately, collocation poses enormous challenges to L2 learners, who are known to make collocational errors when using language, regardless of proficiency or duration of learning (Kuo, 2009; Laufer & Waldman, 2011; Nesselhauf, 2005; Shitu, 2015; Vedder & Benigno, 2016). For example, from an analysis of 900 essays written by 300 Nigerian college students, Shitu (2015) found that even advanced ESL learners make collocational errors due to L1 interference, overgeneralization, and lack of collocational knowledge; among the six subtypes of lexical errors, a verb-noun (or prepositional phrase) pattern was the most problematic. There are also other studies that have examined native speakers and L2 learners in terms of their collocational competence. For instance, in a comparison of L1 and L2 postgraduate students' English writing, Durrant and Schmitt (2009) found that L2 writers used significantly more collocations from the high-frequency bands, but fewer collocations from the low frequency bands compared to the native speakers. More recently, García Salido and Garcia (2018) found that unlike native speakers, advanced learners of Spanish underused low frequency collocations with high mutual information scores (MIS), or strongly associated words with MIS ≥ 5, while overusing high frequency collocations with low mutual information scores (MIS < 5). Demir (2017) also compared L1 and L2 writers by analyzing the research articles published in top-tier journals in the field of ELT in terms of seven categories (verb + noun, verb + adj./adv., noun + verb, noun + noun, adjective + noun, adverb + adjective, and adverb + verb). He found that the L1 writers used three times (n = 1,548 vs. n = 499) as many collocations as non-native counterparts. Interestingly, the L1 writers used more collocations than the L2 writers in all categories except for noun + verb. The fact that L2 learners lack collocational competence may be because collocations are difficult to acquire (Siyanova-Chanturia, 2015). Collocations are known to be acquired late (Vedder & Benigno, 2016), and their development is "slow and uneven" (Laufer & Waldman, 2011, p. 664).

The significance of collocations and their learnability has ignited interest in studying them. The research in this area can be classified into several categories according to their purposes. While some studies attended to describing the distribution of collocations in L2 written or spoken texts (Kim et al., 2020; Tsai, 2015), others focused on teaching and learning collocations (Boers et al., 2014; Nurmukhamedov, 2017). Furthermore, some studies have compared native speakers to learners with the same L1 (Chen & Baker, 2010; Forsberg, 2010; Henderson & Barr, 2010), along with those conducted with learners of different native languages (Alejo-González, 2010; Cross & Papp, 2008; Waibel, 2008). Another difference is noted in the target of the analysis: some studies

focused on spoken production (Crossley & Salsbury, 2011; Gotz & Schilk, 2011) and others on written production (Bestgen, 2017; Siyanova-Chanturia, 2015).

Despite the diverse range of studies, these do not present a complete picture of L2 learners' lexical competence. This is primarily because their analyses have been restricted to specific lexical bundles and collocational combinations, such as adjective-noun or verb-noun (Altenberg & Granger, 2001; Cross & Papp, 2008; Durrant & Schmitt, 2009; Granger, 1998a; Kashiha & Chan, 2015; Laufer & Waldman, 2011; Li & Schmitt, 2010; Siyanova-Chanturia, 2015; Siyanova-Chanturia & Schmitt, 2008). This tendency may have to do with the inadequacy of existing word lists for multi-word (MW) analysis (Martinez & Schmitt, 2012) or the dearth of tools for the extensive analysis of collocation. While there have been endeavors to develop collocation checker programs (Kuo, 2009) or multi-word lists, such as the Academic Formulas List (Simpson-Vlach & Ellis, 2010), and the Phrasal Expressions List (Martinez & Schmitt, 2012), they seem to be incomplete in terms of coverage.

Another problem with vocabulary research is that only a handful of studies have examined both single words and collocations simultaneously (Bestgen, 2017; Kim et al., 2020; Shin et al., 2018; Vedder & Benigno, 2016), and even these are limited in scope. For example, Bestgen (2017) reported formulaic and single-word measures of lexical richness from an analysis of learner language data, but the study focused exclusively on bigrams (e.g., private correspondence, target audience, good example, and more than). Kim et al. (2020) also considered both single words and collocations, and analyzed Korean college students' lexical competence using the BNC-COCA25 Range, COCA_MWU20 ColloGram, and CLAWS Web Tagger. The problem is, however, that the study focused on the learners' lexical competence according to proficiency levels, and thus little is known about their lexical and collocational knowledge in relation to native speakers. Shin et al. (2018) attempted to compare an EFL learner corpus and a native corpus simply in terms of the distribution of single words and multi-words using the BNC-COCA25 Range and COCA_MWU20 ColloGram. This urges us to describe learners' lexical competence systematically and thoroughly by considering both single words and collocations as the unit of analysis and using a tool that facilitates comprehensive analysis.

This study is an attempt to accurately estimate EFL learners' lexical competence in comparison with their native counterparts in terms of SWs and collocations used in writing. This investigation is based on the two areas of lexical competence with carefully graded written texts from beginning to advanced, following Granger and Bestgen's (2014) suggestion to include texts at different levels when sampling. The study used the BNC-COCA25 Range program (Nation, 2012) and a multi-word analysis program, COCA_MWU20 ColloGram (Shin et al., 2018). These programs have explanatory

power because they can systematically analyze the frequency and types of SWs and collocations across the graded lists constructed from BNC-COCA and COCA, respectively. The present study is significant in that it examined the role of collocation in explaining lexical competence in an EFL context, where excessive attention is paid to single words in terms of curriculum, instruction, and materials. Examining the relation between the two constituents of lexis (single words and collocations) will enable us to capture a holistic picture of EFL learners' lexical competence, aside from filling the gap in prior studies. In addition, the comparison of learner language with native data offers solid comparative ground for mapping out the degree and direction of vocabulary instruction in a more balanced way. The two specific research questions to pursue are as follows:

(1) Are there differences between native English speakers and Korean EFL learners in their knowledge of SWs? How different are native speakers from EFL learners at different proficiency levels (high, mid, and low)?

(2) Are there differences between L1 users of English and L2 learners in their collocational knowledge? How different are native speakers from EFL learners with different levels of proficiency (high, mid, and low) in terms of their collocational competence?

3. Methods

3.1. Data

The data used in this study comprised two text corpora: an EFL learner (EFL-KR) corpus and a native speaker (Native-US) corpus. The learner (EFL-KR) corpus consisted of 150 argumentative essays written by Korean college freshmen for an English placement test. The essays were randomly selected from three (high, mid, low) proficiency groups (50 from each).

Table 1 shows the range of scores for each proficiency level. For instance, advanced learners' scores ranged from 15 to 16, which was the maximum score obtainable.

Table 1: Range of scores for the three proficiency levels in the EFL-KR corpus

Level	Score range	CEFR Level	N
High	15~16	C1	50
Mid	10~11	B2	50
Low	6~7	B1	50

The Native-US corpus was compiled from the LOCNESS (Louvain corpus of native English essays), a corpus constructed by the Université Catholique de Louvain (Granger, 1998b). LOCNESS was chosen because it contains academic English texts written by native speakers between the year 1900 and 2000.

Table 2 summarizes the number of tokens for the learner corpus (n = 46,190) and the native corpus (n = 46,791), respectively.

Table 2: Composition of the two corpora

EFL-KR corpus			Native-US corpus		
Level	**N**	**Token**	**Source**	**N**	**Token**
High	50	20,102	Argumentative essays from University of South Carolina	30	34,474
Mid	50	15,995			
Low	50	10,093	Argumentative essays from Presbyterian College, South Carolina	6	12,317
Total		46,190	Total		46,791

3.2 Data collection

For a comparison of vocabulary and collocation used in writing by EFL learners and native speakers, data were compiled from two different sources: a learner corpus and a native corpus. First, for the learner corpus, 150 texts were randomly selected from a collection of essays (n = 1,141) written by college freshmen for an in-house English placement test at a university in Seoul. Before the test, the students provided consent for their texts to be used for research purposes. For the test, the students were asked to take a position regarding the following statement and write an argumentative essay for 30 minutes: "Television, newspapers, magazines, and other media pay too much attention to the personal lives of famous people such as public figures and celebrities. Use specific reasons and details to explain your opinion".

The students' essays were then scored by English-speaking professors, who used a rubric developed by the director and coordinator of the English program at the university. The rubric was designed to measure four dimensions of writing: content, organization, accuracy, and vocabulary. The score for each area ranged from 1 to 4, and thus the maximum score the student could earn was 16 points. To ensure suitability for a comparative analysis of the three proficiency levels, essays that received a minimum of six points were selected because essays with scores lower than that were

roughly 50 words long. The students were generally more proficient than average EFL learners, as they had been admitted to one of the most prestigious colleges in Korea.

For a native speaker corpus, argumentative essays written by American university students were chosen, given that Korean schools and universities adopt American English as the norm. To make this comparable with the EFL-KR corpus (46,190 words), 36 essays (30 from the University of South Carolina and 6 from Presbyterian College, South Carolina) were chosen from the LOCNESS corpus so that its size (46,791 words) was almost the same. All the texts were argumentative essays, and thus identical in genre to the EFL learner texts. The topics, however, varied, as they were written on a variety of topics, such as gender discrimination, drug control, media and self-esteem, role of college, conflict and divorce, adolescent suicide, and water pollution. These texts were submitted by native speakers of English as college writing assignments, and thus their written output may have been different from the EFL learners' due to the nature of the tasks: writing assignments versus placement tests. Given that assignments are scored for course grades, and placement tests generally have lower stakes, hopefully, the psychological pressure associated with these tasks may not have been so different from the students' perspectives.

3.3 Research instruments

BNC-COCA25 Range Program

For lexical analysis, we used a vocabulary analysis program, BNC-COCA25 Range (Nation, 2012), loaded with 25 wordlists extracted from the Corpus of Contemporary American English (COCA) and the British National Corpus (BNC). The program can analyze up to 25 grades (25,000 words) by 1,000 words per grade. The analysis of the present study is confined to the first 20 grades, as the number of words that belonged to the higher levels (21 through 25) was found to be minimal in the preliminary analysis.

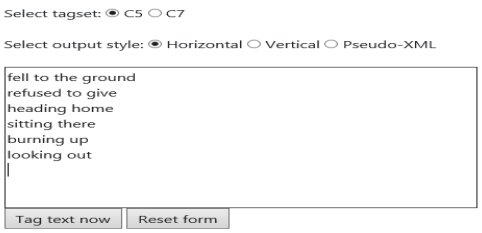
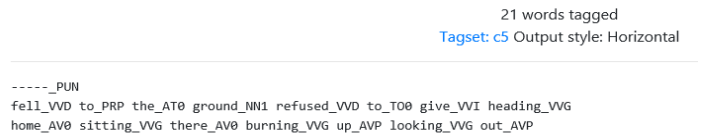
COCA_MWU20ColloGram

To compare the collocational diversity and distribution of the two text corpora, COCA_MWU20 ColloGram was used (Shin et al., 2018). The program is similar to the BNC-COCA Range program, in that it is loaded with 10,000 collocations (20 graded lists, 500 on each) extracted from the COCA and can analyze the collocational distribution by grade.

CLAWS Web Tagger

For an analysis of the part-of-speech (PoS) combination pattern of collocates, a program called CLAWS Web Tagger (University of Lancaster, n.d.) was used.

Figure 1: Part-of-speech tagging using CLAWS Web Tagger

Submitting Collocations into CLAWS Web Tagger	
Output of CLAWS Web Tagger	
Definition of 13 POS Tags	AJ: Adjective AV: Adverb AVP: Adverb particle (e.g., up, off, out) AVQ: WH-adverb (e.g., when, why) CJC: Conjunction DT: Determiner (e.g., a/an, the, this, these) NN: Noun PNQ: WH-pronoun (e.g., who, whoever) POS: The possessive (or genitive morpheme) 's or ' PRP: Preposition VB: Verb TO: Infinitive marker (e.g., to –infinitive) DTQ: WH-determiner (e.g., whose, which)

CLAWS Web Tagger is a free PoS tagging program with two types of tagsets: c5 and c7. For this study, the c5 tagset, a basic tagset (C5) with 62 tags, was adopted. It is sufficiently comprehensive as it was used for tagging the 100-million-word BNC (Garside, 1996). The tagset grouped different varieties of verbs, such as VVI (infinitive of lexical verb), VVD (past tense form of lexical verb), and VVG (-ing form of lexical verb) and merged them into one. Consequently, the 62 PoS were integrated into 13 parts of speech as shown in Figure 1 (Kim et al., 2020).

4. Results

4.1. Korean EFL learners and native English speakers: Knowledge of single words

To compare the lexical knowledge of the two groups, this study examined the distribution of single words used in the two written corpora: an EFL learner corpus (EFL-KR, 46,190 words) and a native speaker corpus (Native-US, 46,791 words). Table 3 summarizes the total number of words (tokens), types, and type-token ratios (TTR) for the native and the learner corpus with three levels: high, mid, and low.

Table 3: Lexical diversity of EFL-KR and Native-US corpora

Group		Token	Type	TTR
EFL-KR	High	20,102	2,468	0.12
	Mid	15,995	2,188	0.14
	Low	10,093	1,625	0.16
EFL-KR Total		46,190	3,936*	0.09
Native-US Total		46,791	5,501	0.12

Note: The number marked with an asterisk (*) is smaller than the sum of the high-, mid-, and low-level data because the same word, repeated across levels, is counted as one.

Given that TTR is a measure of lexical diversity, notably the words used in the native corpus (TTR = 0.12) were more varied than those of the learner corpus (TTR = 0.09). To check whether the two type-to-token proportions differed significantly between the native and EFL learners, a *z*-test was run to compare the two proportions with a web-based *z*-score calculator (Social Science Statistics, 2022). The type-token proportion of the native corpus (12%) was found to be significantly higher than that of the learner corpus (9%) in the test [z (df =1) = 16.34, $p < .001$]. That is, the TTR, a measure of lexical diversity, was significantly higher for native speakers than for EFL learners.

The differences were also noted in the comparisons of native speakers and EFL learners at mid and low levels. Specifically, the two type-to-token proportions were found to be different between the NSs and mid-level [z (df =1) = 6.40, $p < .001$], and between the NSs and low-level [z (df =1) = 11.96, $p < .001$] learners. However, in a comparison of advanced EFL learners with native speakers, it was found that advanced learners used diverse types (n = 2,468) of words for the size of their corpus (n =20,102). Considering that the advanced learner corpus size was 43% of the native speaker corpus, it is surprising that advanced learners did not differ much from their counterparts,

i.e., native speakers, in terms of lexical diversity, as shown by the identical TTR. The seemingly non-significant difference was confirmed by a z-test of the proportions from the two groups [z (df =1) = 1.91, p =.056]. The type-to-token proportion of the advanced learners' vocabulary was identical to that of the native speakers. This means that the advanced learners were as competent as the native speakers in terms of lexical diversity. This finding merits attention, as it deviates from a general assumption that EFL learners are inferior to native speakers with regard to their linguistic repertoire, including lexical knowledge. It is particularly notable that the TTR measures of the low- and mid-level students were higher than that of the native speakers, which was due to the size of the corpora. The small size of the two text corpora seems to have contributed to boosting the TTR measures.

For a more detailed analysis of the profiles of words from the corpora, a BNC-COCA 25 Range program was used, because it includes 25 wordlists (BNC-COCA 25) graded from 1K (the most frequent 1,000 words) through 25K according to frequency. Then, the lexical profiles of the two groups (native speakers and EFL learners) were compared across the 20 graded bands. Table 4 shows the distribution of words across the groups and word levels.

Table 4: Lexical profiles of EFL-KR and Native-US corpora in token (%)

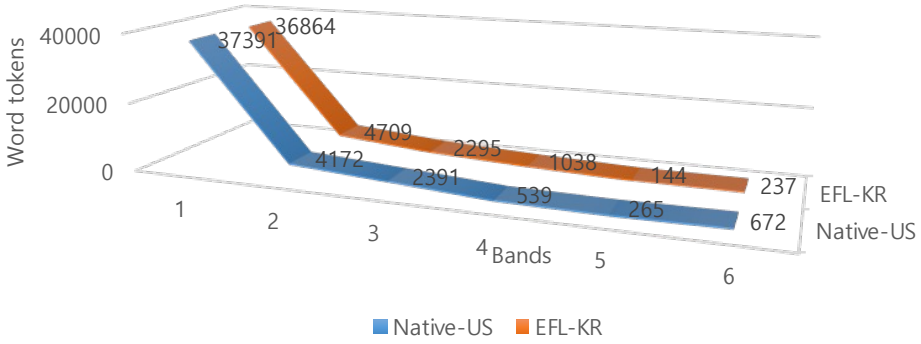
Level	Band	EFL-low	EFL-mid	EFL-high	EFL-KR	Native-US
One	1K	8,023 (79.49)	12,795 (79.99)	16,046 (79.82)	36,864 (79.81)	37,391 (79.91)
Two	2K	1,063 (10.53)	1,641 (10.26)	2,005 (9.97)	4,709 (10.19)	4,172 (8.92)
Three	3K	501 (4.96)	730 (4.56)	1,064 (5.29)	2,295 (4.97)	2,391 (5.11)
Four	4K	198 (1.96)	362 (2.26)	478 (2.38)	1,038 (2.25)	539 (1.15)
Five	5K	32 (0.32)	55 (0.34)	57 (0.28)	144 (0.31)	265 (0.57)
Six-twenty	6K and Above	37 (0.37)	81 (0.51)	119 (0.59)	237 (0.51)	672 (1.44)

Note: The shaded lists (6K and above) were combined for analysis; off-list words were excluded; thus, the total percentage for each column is not 100.

As seen in the table, level 1 through level 5 contain one thousand words frequently used at the respective level, and they are expressed as 1K band through 5K band, respectively. As there are few words beyond level six, level six through level 20 were combined for analysis, which was then categorized as the '6K and above' band. For this reason, the 6K and above band seems to contain more words than the 5K band, but in general, as the band goes up (e.g., from 1 to 2), the frequency of words declines.

The table also shows that the native corpus contained more words from the 5K and the 6K and above levels, which was almost double and triple the number of words in the learner corpus, respectively. However, interestingly, the lexical distribution of the two corpora displayed similar patterns (see Figure 2).

Figure 2: Word tokens: EFL-KR and Native-US corpora

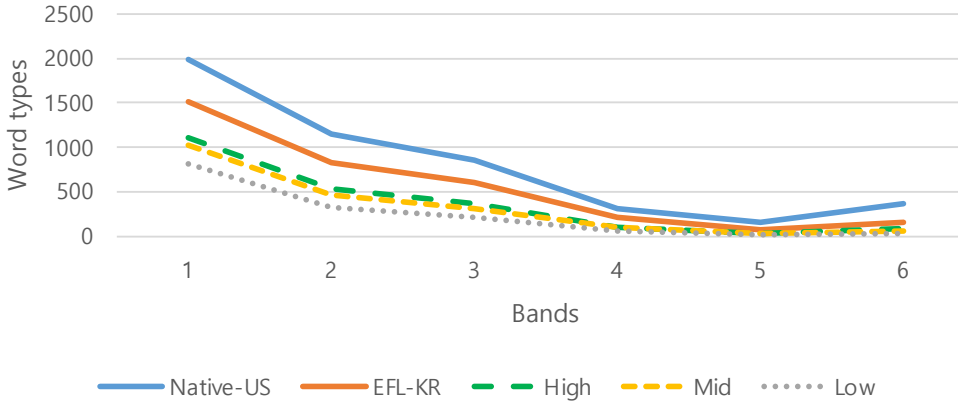


Considering that the two corpora were almost the same size and thus were directly comparable, the similar pattern that was observed deserves attention. First, regardless of the groups, about 80% of the words used were from the 1K band. In fact, more than 90% of the words were from the first two thousand words, regardless of group. Second, while the EFL learner corpus showed a steady decrease between the 2K and 5K bands and a plateau in the levels above, the native corpus displayed a steep decrease in the 4K band, and then a gradual increase in the subsequent levels. Third, native speakers used the 5K words about twice as much and the 6K words almost three times as much as the EFL learners. In other words, the differences are more pronounced in the 5K and 6K and above bands. With this exception, the lexical profiles or distributions were similar across the groups. This may have been because the students who contributed to the learner corpus were considerably competent in using productive vocabulary compared to average EFL learners, particularly the high-level learners.

Figure 3 shows the comparison of word types across the proficiency groups in relation to the native speakers. Types are used instead of tokens because the three

proficiency groups comprised of the EFL learner corpus, and thus were much smaller than the native corpus in terms of tokens.

Figure 3: Word types: High-, mid-, and low-level learners and native speakers



As seen in Figure 3, EFL learners used varied types of words up to 4K, and showed a gradual decrease thereafter in the learner corpus, although the native corpus showed slight increase from the 5K to 6K and above bands. The difference in learners' vocabulary use differed according to their proficiency, although the difference between the high and the mid-level learners was marginal. Interestingly, the differences across proficiency levels were larger in the 1K through 4K bands than in the sequential levels, such as the 5K and 6K and above bands, where the differences were minimal.

4.2. Korean EFL learners and native English speakers: Collocational knowledge

4.2.1 Collocational diversity and distribution

In addition to the learners' knowledge of single words, their knowledge of collocations was compared with that of native speakers for a holistic measurement of lexical competence. To analyze collocational diversity and distribution, the study used COCA_MWU20, loaded with 20 graded-multiword-unit (MWU) lists from the COCA, with each list containing 500 MWU families. Table 5 presents the tokens, types, and TTR of collocations used by native speakers and EFL learners in three levels: high, mid, and low.

Table 5: Collocational diversity of EFL-KR and Native-US corpora

Group		Token	Type	TTR
EFL-KR	High	798	350	0.44
	Mid	590	304	0.52
	Low	392	208	0.53
EFL-KR Total		1,780	670*	0.38
Native-US Total		1,749	1,091	0.62

Note: The number marked with an asterisk (*) is smaller than the sum of the high-, mid-, and low-level data because the same word, repeated across levels, is counted as one occurrence.

As seen in the table, the tokens and types of collocations in the learner corpus increased in proportion to learner proficiency. In other words, the more proficient the learners were, the more tokens and types of collocations they used. Compared to the learner corpus, the native speaker corpus contained a more diverse range of collocations. Although the collocation size of the native speaker corpus (n = 1,749) was approximately the same as that of the learner corpus (n = 1,780), there were far fewer types of collocations used by EFL learners (n = 670) than by native speakers (n = 1,091), and the difference was significant in the statistical comparison of the two proportions. The type-to-token proportion of the native corpus (62%) was significantly higher than that of the learner corpus (38%) in the z-test [z (df =1) = 14.70, p <.001]. This indicates that the native speaker corpus differed significantly from the learner corpus in terms of collocational variety, and that native speakers employed (about 1.6 times) more diverse types of collocations than EFL learners.

The study also examined whether there are statistical differences between the native speakers and EFL learners at different proficiency levels, and found significant differences between NSs and learners at all levels. The type-to-token proportions for the two groups were found to be statistically different [z (df =1) = 8.75, p <.001] with the proportion of the native speaker corpus (62%) being significantly higher than that of the advanced learner corpus (44%). The type-to-token proportions were also different between the NSs and EFL learners at the intermediate level [z (df =1) = 4.65, p <.001] or at the low-level [z (df =1) = 3.41, p <.001]. This shows that native speakers had a greater range of collocations than EFL learners of all proficiency levels.

From the two comparisons, we can see that EFL learners, regardless of proficiency level, used a more limited set of collocations than native speakers. The difference in the type-to-token proportions (62% vs. 38%) for the two groups manifests the disparity in their collocational competence. Although the difference was slightly smaller for

advanced learners versus native speakers (62% vs. 44%), the two corpora were still significantly different in terms of the type-token proportions, with the native speakers employing more diverse types of collocations than their counterparts. Table 6 shows how the collocations in the two corpora were distributed across the 20 graded lists.

Table 6: Collocational profiles of EFL-KR and Native-US corpora in token (%)

Level	Band	EFL-low	EFL-mid	EFL-high	EFL-KR	Native-US
One-two	1K	194 (49.49)	280 (47.46)	365 (45.74)	839 (47.13)	775 (44.31)
Three-four	2K	49 (12.50)	74 (12.54)	103 (12.91)	226 (12.70)	280 (16.01)
Five-six	3K	31 (7.91)	49 (8.31)	80 (10.03)	160 (8.99)	200 (11.44)
Seven-eight	4K	47 (11.99)	82 (13.9)	122 (15.29)	251 (14.10)	111 (6.35)
Nine-ten	5K	39 (9.95)	51 (8.64)	48 (6.02)	138 (7.75)	102 (5.83)
Eleven-twenty	6K and Above	32 (8.16)	54 (9.15)	80 (10.03)	166 (9.33)	281 (16.07)

Note: As each collocational level has 500 collocation families, two levels were combined to yield 1,000; Levels 1 and 2 were combined to make the 1K collocation family list. Likewise, Levels 3 and 4 (2K), 5 and 6 (3K), 7 and 8 (4K), 9 and 10 (5K), 11 to 20 (6K and above) were merged to make the collocation groups comparable to the wordlist groups; the shaded lists (6K and above) were combined for analysis.

In general, the frequency of collocations decreased as the collocation level increased, although there were occasional exceptions. In the learner corpus, the lower the learners' level, the higher the frequency of the 1st band collocations (High: 45.74%; Mid: 47.46%; Low: 49.49%). In addition, the frequency of the 2nd band collocation was consistently low in the EFL learner corpus regardless of proficiency (High: 12.91%; Mid: 12.54%; Low: 12.50%), as opposed to the native speaker corpus, which contained 16.01% of the 2nd band collocations. Notably, 42% (n = 740) of the collocations that the EFL learners used were from the first 500 collocation families. For instance, the collocate "too much" was used 208 times, taking up 28% of the 740 occurrences in the 1st level. Furthermore, the collocational distribution showed an unusual rise in level 7 or band 4 (level 7 and 8 combined) due to the writing topic. Namely, the students' use of topic-related collocates (such as "public figure") inflated the frequency of the 4th band or the 7th level collocation.

Figure 4 delineates the differences in collocational distribution between the native and learner corpora. The collocations of the native corpus were spread across all the levels and displayed a steady decrease from the 2K to 5K band, followed by a rise in the 6K and above band. By contrast, the learner corpus showed a drop from the 5K to 6K band and some fluctuations between the 3K and 5K band, as the number of collocations in percent drastically increased due to the frequent use of topic-related collocation, "public figure".

Figure 4: Collocation tokens: EFL-KR and Native-US corpora

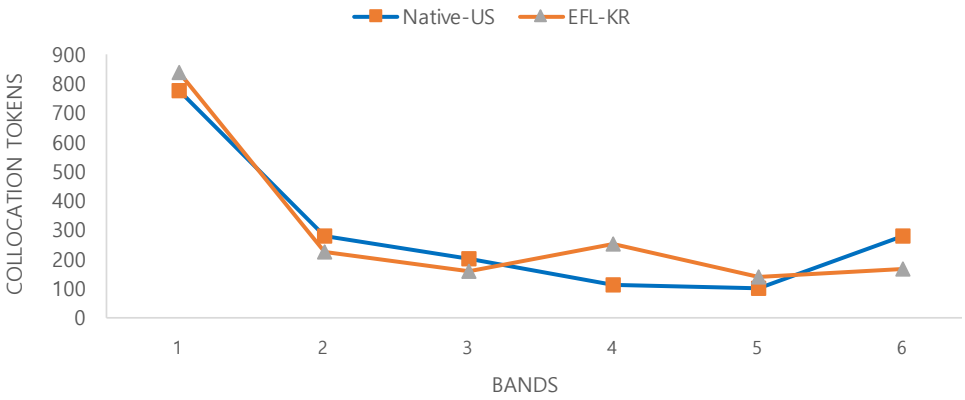
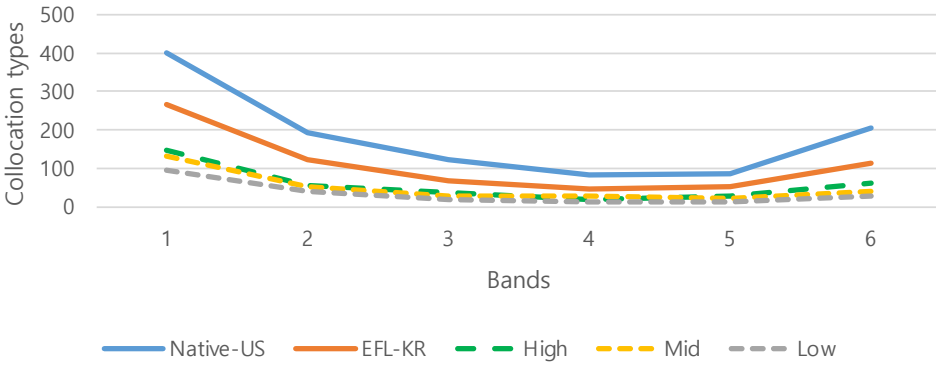


Figure 5 presents the distribution of collocations across the proficiency groups in comparison with the native corpus. It is particularly noteworthy that the EFL learners used far fewer collocations than single words and in a limited range only. The number of collocations used by the three proficiency groups ranged from 13 to 147; for the entire learner groups, it ranged from 53 to 266. This indicates that EFL learners used substantially less varied forms of collocations than native speakers, whose collocation types ranged from 84 to 400.

Figure 5: Collocation types: High-, mid-, and low-level learners and native speakers



Although the collocational distributions were seemingly similar up to the 5K band, the native speakers used far more diverse forms of collocations, whereas the three proficiency groups used collocations in a narrow range. In other words, regardless of proficiency, the three groups of learners used a limited range of collocations (Low: 13~94, Mid: 22~131, High: 18~147) across the graded bands.

4.2.2 Collocational patterns

In addition to the collocational diversity and distribution, collocational patterns were analyzed by tagging Part-of-Speech (PoS) combinations in each text corpus. As seen in Table 7, the native speaker corpus contained slightly more diverse combinations of collocations.

Table 7: Total number of POS combinations of collocations

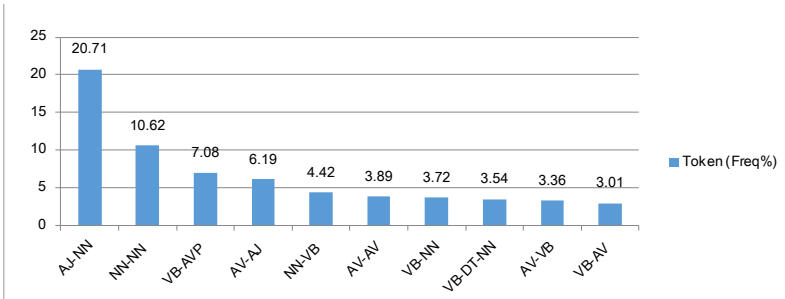
Native-US	EFL-KR	High	Mid	Low
108	95	61	65	55

Note: There were 86 overlaps in the POS combinations of collocations across the three levels.

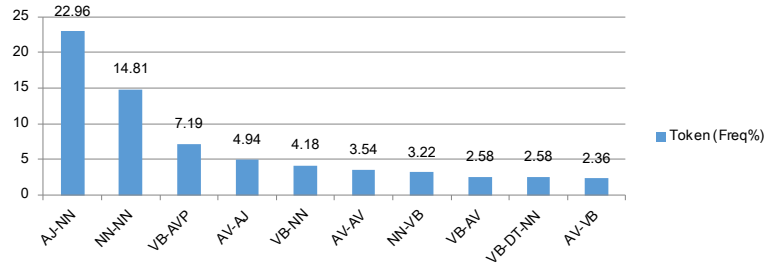
The number of PoS combinations extracted was 95 for the learner corpus, whereas it was slightly greater (n = 108) for the native speaker corpus. Figure 6 presents the distribution of different PoS combinations of collocations. Despite the differences between the learner and native speaker corpora in terms of collocational diversity and distribution, the collocational combinations displayed similar patterns. For example, the most frequent PoS combination was AJ-NN (e.g., "romantic comedy") both in the learner corpus (20.71%) and native corpus (22.96%), and the top four PoS combinations (AJ-NN, NN-NN, VB-AVP, and AV-AJ) were identical. Although there were slight differences in the frequency rank, the top 10 patterns were the same across the two corpora.

Figure 6: Comparison of top 10 collocational patterns: PoS combinations of the two corpora

EFL-KR



Native-US



The collocational patterns observed in the two corpora were in line with five of the six PoS combinations noted by Lewis (2000). Likewise, the five patterns proposed by McCarthy and O'Dell (2005) were also included in the top 10 PoS patterns. Consequently, the following seven patterns from Lewis (2000) and McCarthy and O'Dell (2005) matched the top 10 POS patterns in this study: AJ-NN (e.g., "bright color"), NN-NN (e.g., "radio station"), AV-AJ (e.g., "extremely inconvenient"), NN-VB (e.g., "economy boomed"), VB-AV (e.g., "smiled proudly"), VB-DT-NN (e.g., "submit a report"), and AV-VB (e.g., "happily married").

To summarize, as shown in the distribution of the top 10 collocation patterns, the PoS combinations of collocations were the same between the EFL learner and the native corpus. There were, however, some differences between the groups in the way they used combinations of words. For instance, the native speakers used the NN-NN combination more frequently, whereas the high-, mid-, and low-level EFL learners used the VB-, ADV-, and ADJ-associated collocational patterns more frequently than their counterparts, respectively. This finding may be because L2 learners tend to prefer certain combinations of high frequency collocations and overuse them (Durrant & Schmitt, 2009; García Salido & Garcia, 2018; Granger, 1998a).

5. Discussion

The study focused on comparing native speakers and L2 learners in terms of their knowledge of single words (research question 1) and collocations (research question 2). As expected, native speakers excelled against EFL learners in terms of lexical diversity. The EFL learners relied greatly on the first 2K high-frequency words, whereas the native speaker group used many low-frequency words from the 5K, and the 6K and above bands. In other words, the EFL group underused low-frequency words, unlike their native-speaking counterparts who utilized advanced vocabulary in addition to high-frequency words from the first 2,000 words.

In a comparison of the NSs and the learners at different proficiency levels, the native speaking counterparts outperformed the mid- and the low-level EFL students in

their lexical and collocational knowledge. This is not surprising as it is in line with the findings from earlier studies (Ädel & Erman, 2012; Dabbagh & Enayat, 2019; Doro, 2007; Lessard-Clouston, 2006; Siyanova-Chanturia & Schmitt, 2007). It is, however, interesting to note that advanced EFL learners were as competent as native speakers in terms of lexical variety. This finding departs from the general expectation that L2 learners, regardless of proficiency levels, are less skilled than NSs. In that regard, the non-significant difference between the NSs and advanced learners is promising, in that it signals the potential development of L2 learners' interlanguage in terms of lexis. In other words, L2 learners' knowledge of SWs can grow to the level of native speakers, and this growth can be facilitated with the advancement of internet technology (e.g., synchronous computer-mediated communication) and transnational mobility in the era of globalization (Duff, 2015; Kim & Kim, 2022; Ma, 2017). L2 learners are now able to readily access target language input on the Internet and easily travel to English-speaking countries. Consequently, they can gain sufficient exposure to L2 input and increased opportunities for interaction and communication. As there is room for development of L2 learners, we should be wary of regarding them as inferior to native speakers, let alone promoting native speakerism.

Despite their excellence in using a diverse range of words, the advanced learners were not as competent as native speakers in their collocational knowledge, as indicated in the significant difference between the two groups, and also corroborated by earlier studies (Ädel & Erman, 2012; Demir, 2017; Shitu, 2015). The difference was also found between the EFL learners as a whole and the native speakers. Also notable was that the learners, regardless of proficiency levels, used far fewer collocations than native speakers and in a narrower range. This result confirms the findings from Durrant and Schmitt (2009) who demonstrated that L2 postgraduate writers used substantially more collocations from the high-frequency bands in contrast to native speakers. This finding also supports Vedder and Benigno's (2016) claim that collocations are acquired late. In addition, the finding that the EFL learners lack collocational competence seems to be in line with Kim et al. (2020) who examined the collocational competence of Korean college students at high, mid, and low levels, and found that the distribution of collocations used by advanced learners was similar to the other groups' collocational distribution. This means that L2 learners, regardless of proficiency, used a limited repertoire of collocations, mostly confined to the 2K collocation families.

Tsai (2015) also reported that Taiwanese EFL learners used far fewer and less varied collocations than native speakers. As Granger (1998a) noted, learners tend to adopt a safe approach in their use of collocations, that is, using a small set of formulaic sequences that they are certain about. Likewise, Siyanova-Chanturia (2015) suggested that L2 learners have trouble using collocations and are inferior to native speakers in their collocational knowledge. This has been also addressed in other previous studies

(Kim et al., 2020; Kuo, 2009; Laufer & Waldman, 2011; Nesselhauf, 2005; Shitu, 2015; Vedder & Benigno, 2016). This difficulty may be because collocational development occurs gradually at a slow pace (Laufer & Waldman, 2011).

Considering that collocations are acquired later than single words (Vedder & Benigno, 2016), it is crucial to design a curriculum and teaching method that go beyond SWs. As Laufer and Waldman (2011) put, collocations are "a necessary component of second-language (L2) lexical competence in addition to the knowledge of single words" (p. 648). For this reason, Bestgen (2017) suggested that formulaic competence or "the native-like use of ready-made sequences of words" (p. 65) should be considered in L2 writing assessment, as it best predicted the quality of learner texts. Bestgen argued that formulaic competence has been neglected in automated L2 writing assessment programs, such as ETS' e-Raters. This negligence can have a harmful influence on teachers and learners alike. As Henriksen (2013) cautioned, teachers are likely to place less focus on collocations and fail to use materials effective for learner awareness raising for collocations.

Similarly, curriculum wise, little attention has been paid to collocation in the primary through secondary school context in Korea (Kim et al., 2020). Given that the non-native participants of this study were newly admitted college freshmen, it is lamentable that they learned English under the secondary school curriculum that trivialized the role of collocation. The Korean national curriculum seems to place more emphasis on SWs than collocations, in that it has specified wordlists but not collocation lists for different grades. This curricular negligence is manifest in the learning materials that the students use to study for the college SAT. Considering that the test-preparation materials often contain low-frequency idioms, collocations are not likely to be adequately level-based and evenly covered. In addition, classroom teaching focuses on preparing for the Korean SAT (KSAT), the college entrance exam (Kim, 2021). As the KSAT has a harmful washback on classroom instruction, particularly in the secondary school context, teachers only attend to the skill areas that are measured in the test. As knowledge of individual words suffices for students to excel on the test, they do not acknowledge the need for studying multi-words or collocations. However, as Bahns and Eldaw (1993) cautioned, tests that do not assess collocational knowledge do not yield comprehensive measures of learners' lexical competence. In particular, given that written discourse is composed of formulaic sequences (Ehrman & Warren, 2000), it is critical to teach and assess collocations along with lexis, and maintain a balance between word knowledge and collocational knowledge (Kim et al., 2020). Once collocational knowledge is measured in a language test, it will receive adequate attention from teachers and students, who will then change how they teach and learn for lexical competence development (Bestgen, 2017). This will eventually help to make a positive contribution to L2 writing proficiency development.

6. Conclusion

The present study is significant in that it analyzed L2 learners' use of single words and collocation in their writing, in comparison to that of native speakers. The study found significant differences between the native speaker and the whole learner group in both lexical and collocational knowledge. The difference between the advanced learners and the NSs was, however, significant only in their collocational competence. These findings are meaningful, as they offer holistic information about learners' knowledge of SWs and collocations. They also have pedagogical implications for curriculum design and vocabulary instruction and assessment. First, it is of paramount importance that single words across the graded lists be used in teaching and assessment. It was found from the lexical distribution that L2 learners used a limited range of words compared to native speakers. This indicates the need for sampling a diverse range of words across the graded lists and covering them in classroom instruction. Learners can then acquire words from the advanced bands, such as 5K and 6K and above, and increase their lexical repertoire. Moreover, curriculum designers should consider collocation an integral component of the curriculum and develop collocation lists in addition to wordlists. Once collocation is included in the curriculum as a core unit of learning, material designers should sample collocational expressions from different bands, so that a diverse range of collocations can be used in textbooks. These expressions should then be covered in classroom teaching to ensure that students be exposed to a vast array of collocational combinations. Once they become familiar with collocational expressions through repeated exposure, they are likely to reduce the number of collocational errors and acquire collocational knowledge. In addition, classroom tests should be constructed to measure collocational knowledge and provide complete information about learner's lexical competence (Bahns & Eldaw, 1993). In doing so, collocational expressions should be sampled across the graded lists, and various combinations of collocations should be used to ensure beneficial washback effect of the tests.

Notwithstanding the pedagogical merits of these findings, they are not without limitations. First, despite the attempt to equalize the size and text genre of the two corpora, they were not completely comparable, as the texts were written on different topics, and the nature of the tasks was different. The learner corpus was composed of texts from a timed-writing task in a test-taking situation whereas the native speaker corpus comprised college writing assignments. Thus, it is plausible that the difference in the nature of the tasks may have affected the students' written output. Another limitation is that the native corpus was available in its entirety without individual writers' demographic information, and thus it was not possible to run a statistical comparison of the two corpora in terms of the lexical and collocational distribution.

Thus, a follow-up study is recommended with two groups of text corpora that are tightly controlled in terms of topics, task type, and writers. Finally, as the non-native corpus was constructed with written texts by Korean college students who were generally more proficient than average college students, the findings of the study may not be directly applicable and generalizable to groups of learners in Korea or other EFL contexts. It should also be noted that the learner corpus was compiled from a context where the KSAT has a negative washback on the learning of collocations. Thus, the findings of the study should be interpreted with caution, and a further study should be deigned to resolve the limitations by controlling for the effects of the writer, tasks, topics, and learning contexts.

References

Ädel, A., & Erman, B. (2012). Recurrent word combinations in academic writing by native and non-native speakers of English: A lexical bundles approach. *English for Specific Purposes, 31*(2), 81-92.

Alejo-González, R. (2010). L2 Spanish acquisition of English phrasal verbs: A cognitive linguistic analysis of L1 influence. In M. C. Campoy, B. Belles-Fortuno, & M. L. Gea-Valor (Eds.), *Corpus-based approaches to English language teaching* (pp. 149-166). London, UK: Continuum.

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics, 22*(2), 173–195.

Bahns, J., & Eldaw, M. (1993). Should we teach EFL students collocations? *System, 21*(1), 101–114.

Bestgen, Y. (2017). Beyond single-word measures: L2 writing assessment, lexical richness and formulaic competence. *System, 69*, 65-78.

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.

Chen, Y.-H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*, 30-49.

Cobb, T. (2010). Learning about language and learners from computer programs. *Reading in a Foreign Language, 22*(1), 181-200.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Coxhead, A. (2016). Reflecting on Coxhead (2000), "A new academic word list." *TESOL Quarterly, 50*(1), 181-185.

Coxhead, A., & Boutorwick, T. J. (2018). Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths, and science.

TESOL Quarterly, 52(3), 588-610.

Cross, J., & Papp, S. (2008). Creativity in the use of verb + noun combinations by Chinese learners of English. In G. Gilquin, S. Papp, & M. B. Diez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 57–81). Amsterdam, Netherlands: Rodopi.

Crossley, S. A., & Salsbury, T. (2011). The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Teaching, 49*, 1-26.

Dabbagh, A., & Enayat, M. J. (2019). The role of vocabulary breadth and depth in predicting second language descriptive writing performance. *The Language Learning Journal, 47*(5), 575-590.

Dang, T., & Webb, S. (2014). The lexical profile of academic spoken English. *English for Specific Purposes, 33*, 66-76.

Demir, C. (2017). Lexical collocations in English: A comparative study of native and non-native scholars of English. *Journal of Language and Linguistic Studies, 13*(1), 75-87.

Doró, K. (2007). The use of high-and low-frequency verbs in English native and non-native student writing. In Z. Lengyel & J. Navracsics (Eds.), *Second language lexical processes: Applied linguistic and psycholinguistic perspectives* (pp. 117-129). Clevedon, UK: Multilingual Matters.

Duff, P. A. (2015). Transnationalism, multilingualism, and identity. *Annual Review of Applied Linguistics, 35*, 57-80.

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

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

Forsberg, F. (2010). Using conventional sequences in L2 French. *International Review of Applied Linguistics, 48*(1), 25-50.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 75-93). Harlow, UK: Longman.

García Salido, M., & Garcia, M. (2018). Comparing learners' and native speakers' use of collocations in written Spanish. *International Review of Applied Linguistics in Language Teaching, 56*(4), 401-426.

Garside, R. (1996). The robust tagging of unrestricted text: The BNC experience. In J. Thomas & M. H. Short (Eds.), *Using corpora for language research: Studies in honour of*

Geoffrey Leech (pp. 167–180). London. Longman.

González, M. C. (2013). *The intricate relationship between measures of vocabulary size and lexical diversity as evidenced in non-native and native speaker academic compositions.* [Unpublished doctoral dissertation]. University of Central Florida.

González-Fernández, B., & Schmitt, N. (2015). How much collocation knowledge do L2 learners have? The effects of frequency and amount of exposure. *International Journal of Applied Linguistics, 166*(1), 94–126.

Gotz, S., & Schilk, M. (2011). Formulaic sequences in spoken ENL, ESL, and EFL. In J. Mukherjee & M. Hundt (Eds.), *Exploring second language varieties of English and learner Englishes: Bridging a paradigm gap* (pp. 79-100). Amsterdam: John Benjamins.

Granger, S. (1998a). Prefabricated patterns in advanced EFL writing: Collocations and formulae. In A. Cowie (Ed.), *Phraseology: Theory, analysis, and applications* (pp. 145-160). Oxford: Oxford University Press.

Granger, S. (1998b). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). London: Addison-Wesley Longman.

Granger, S., & Bestgen, Y. (2014). The use of collocations by intermediate vs. advanced non-native writers: A bigram-based study. *International Review of Applied Linguistics, 52*(3), 229-252.

Gregori-Signes, C., & Clavel-Arroita, B. (2015). Analysing lexical density and lexical diversity in university students' written discourse. *Procedia-Social and Behavioral Sciences, 198*, 546-556.

Ha, H. T. (2021). Exploring the relationships between various dimensions of receptive vocabulary knowledge and L2 listening and reading comprehension. *Language Testing in Asia, 11*(1), 1-20.

Henderson, A., & Barr, R. (2010). Comparing indicators of authorial stance in psychology students' writing and published research articles. *Journal of Writing Research, 2*, 245-264.

Henriksen, B. (2013). Research on L2 learners' collocational competence and development: A progress report. In C. Bardel, C. Lindqvist, & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 29-56). Amsterdam: Eurosla.

Johnson, M. D., Acevedo, A., & Mercado, L. (2016). Vocabulary knowledge and vocabulary use in second language writing. *TESOL Journal, 7*(3), 700-715.

Kao, S. M., & Wang, W. C. (2014). Lexical and organizational features in novice and experienced ELF presentations. *Journal of English as a Lingua Franca, 3*(1), 49-79.

Kashiha, H., & Chan, S. H. (2015). A little bit about: Differences in native and non-native speakers' use of formulaic language. *Australian Journal of Linguistics, 35*(4), 297-310.

Kim, S.-Y. (2021). Developing disciplinary writing proficiency of pre-service English teachers; Writing-to-learn through online PBL. *Multimedia-Assisted Language Learning, 24*(1), 37-63.

Kim, S.-Y., Shin, D., & Kim, K.-S. (2020). Korean college students' use of vocabulary and collocation according to writing proficiency. *Multimedia-Assisted Language Learning, 23*(2), 215-235.

Kim, S.-Y., & Kim, K.-S. (2022). Vocabulary transfer from reading to writing: A comparison of essay writing and synchronous CMC. *TESL-EJ, 26*(1), 1-21.

Kuo, C. (2009). An analysis of the use of collocation by intermediate EFL college students in Taiwan. *ARECLS, 6*, 141-155.

Laufer, B., & Nation, I. S. P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics, 16*, 307-322.

Laufer, B., & Nation, I. S. P. (2001). Passive vocabulary size and speed of meaning recognition: Are they related? *EUROSLA Yearbook, 1*, 7-28. https://doi.org/10.1075/eurosla.1.05lau

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647-672.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English: based on the British national corpus.* London: Longman.

Lessard-Clouston, M. (2006). Breadth and depth specialized vocabulary learning in theology among native and non-native English speakers. *Canadian Modern Language Review, 63*(2), 175-198.

Lewis, M. (1993). *The lexical approach.* Hove, UK: Language Teaching Publications.

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

Li, J., & Schmitt, N. (2010). The development of collocation use in academic texts by advanced L2 learners: A multiple case study approach. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 2-46). London: Continuum.

Ma, Q. (2017). Technologies for teaching and learning L2 vocabulary. In C. A. Chapelle & S. Sauro (Eds.), *The handbook of technology and second language teaching and learning* (pp. 45-61). Hoboken, NJ: Wiley.

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics, 33*(3), 299-320.

McCarthy, M., & O'Dell, F. (2005). *English collocations in use* (5th ed.). Cambridge: Cambridge University Press.

Milton, J. (2013). Measuring the contribution of vocabulary knowledge to proficiency in the four skills. In C. Bardel, C. Lindqvist & B. Laufer (Eds.), *L2 vocabulary acquisition, knowledge and use: New perspectives on assessment and corpus analysis* (pp. 57-78). EuroSLA Monograph 2. The European Second Language Association.

Nasseri, M., & Thompson, P. (2021). Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing, 47*. https://doi.org/10.1016/j.asw.2020.100511

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

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review, 63*(1), 59-82.

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved January, 17 2020, from http://www.victoria.ac.nz/lals/about/staff/publications/BNC_COCA_25000.zip

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223-242.

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

Nurmukhamedov, U. (2017). The contribution of collocation tools to collocation correction in second language writing. *International Journal of Lexicography, 30*(4), 454–482.

Olinghouse, N. G., & Leaird, J. T. (2009). The relationship between measures of vocabulary and narrative writing quality in second- and fourth-grade students. *Reading and Writing, 22*, 545–565. https://doi.org/10.1007/s11145-008-9124-z

Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens & S. Gozdz-Roszkowski (Eds.), *Practical applications in language and computers 2005* (pp. 127–140). Frankfurt & Main: Peter Lang.

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

Read, J. (2007). Second language vocabulary assessment: Current practices and new directions. *International Journal of English Studies, 7*(2), 105-125.

Read, J., & Nation, P. (2006). An investigation of the lexical dimension of the IELTS speaking test. *IELTS Research Reports, 6*, 207-231.

Schmitt, N. (2008). Instructed second language vocabulary learning. *Language Teaching Research, 12*, 329–363.

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the vocabulary levels test. *Language Testing, 18*(1), 55-88.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching, 47*(4), 484-503.

Shin, D. (2015). The influence of vocabulary regulations in the 2009 national curriculum of English on English textbooks. *The Journal of Foreign Studies, 34*, 47-76.

Shin, D., Chon, Y., Lee, S., & Park, M. (2018). A comparison of single word and multi-word unit profiles in spoken and written corpora of Korean learners and English native speakers. *Journal of the Korea English Education Society, 17*(2), 93-112.

Shitu, F. M. (2015). Collocation errors in English as a second language (ESL) essay writing. *International Journal of Cognitive and Language Sciences, 9*(9), 3270-3277.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*, 487-512.

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

Siyanova-Chanturia, A., & Schmitt, N. (2007). Native and nonnative use of multi-word vs. one-word verbs. *IRAL, 45*, 119-139.

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

Social Science Statistics. (2022, July 1). *Z score calculator for 2 population proportions.* https://www.socscistatistics.com/tests/ztest/

Supasiraprapa, S. (2018). Second language collocation acquisition: Challenges for learners and pedagogical insights from empirical research. *The Journal of Asia TEFL, 15*(3), 797-804.

Tsai, K.-J. (2015). Profiling the collocation use in ELT textbooks and learner writing. *Language Teaching Research, 19*(6), 723-740.

University of Lancaster. (n.d.). *Free CLAWS web tagger.* http://ucrelapi.lancaster.ac.uk/claws/free.html

Vedder, I., & Benigno, V. (2016). Lexical richness and collocational competence in second language writing. *International Review of Applied Linguistics in Language Teaching, 54*(1), 23-42.

Waibel, B. (2008). *Phrasal verbs: German and Italian learners of English compared.* Saarbrücken, Germany: VDM.

Yu, A., & Trainin, G. (2022). A meta-analysis examining technology-assisted L2 vocabulary learning. *ReCALL, 34*(2), 232-252.

Zareva, A. (2007). Structure of the second language mental lexicon: How does it compare to native speakers' lexical organization? *Second Language Research, 23*(2), 123–153.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition, 27*(4), 567–595.